Review of: "ScienceON Knowledge Graph System: Exploring New Frontiers in Science and Technology Information Integration System"

Silvio Peroni[1]

1 University of Bologna

Potential competing interests: No potential competing interests to declare.

In this article, the authors introduce ScienceON, a knowledge graph dedicated to publishing scholarly and research information about publications, projects, people involved in the process, etc., in RDF and making such data available to all via the Web. The data contained there come from several sources containing different types of information that focus mainly on South Korean publications and research endeavours. The paper is also enriched by some experiments that show the use of these data for addressing graph-based tasks and a list of applications built upon such data to make them available to non-expert users.

These kinds of works, i.e., creating scholarly knowledge graphs that are open, accessible, and reusable, are crucial for the community since they are the steps needed to aim at building a distributed system that, in principle, contains information about all the research information produced around the globe, following the vision that was introduced, probably for the very first time, by Robert Cameron in his work published in First Monday (https://doi.org/10.5210/fm.v2i4.522). In addition, having a knowledge graph on publications of countries and/or disciplines that usually are excluded or not fully represented in well-known proprietary indexes has even more value since it offers more equitable access to science.

While I praise these particular activities and, thus, the topic of this work, after reading the article with tremendous interest, I think it has several issues in its present form. The first one is about its narrative. Analysing its current organisation, it seems to merge two different contributions, i.e., the knowledge graph developed and all the experiments done on that knowledge graph (i.e., section 5). The authors justify section 5 by claiming that it supports the claim of the quality of ScienceON - the rationale here is that since we can run experiments on graph-related tasks on the knowledge graph, such a knowledge graph must be qualitatively sound. However, the point is that these experiments could also be run on other knowledge graphs, obtaining similar results. Thus, they do not justify the claim on the quality of ScienceON and appear to be just an exercise with no added value to the existence of ScienceON. Indeed, it would have been better to have two separate papers here: one about ScienceON, the workflow for constructing it, etc., and the other about the experiments on graph-based tasks, where ScienceON is used as one possible knowledge graph for application. I would suggest removing all the parts related to the topic in section 5 - which also affects part of the related works (i.e., section 3.2) that appears suddenly without a clear justification - and focusing entirely on the knowledge graph construction, providing even more details when needed.

As follows, other points that should be considered as well.

- In the abstract, the authors say that ScienceON Knowledge Graph should be considered a gold standard. This definition is not appropriate for several reasons. First, to claim something is a gold standard, one has to demonstrate its full quality and that it is (in principle) free of issues. However, according to my experience, it is kind of impossible to claim that for a KG of scholarly data due to the heterogeneity, possible mistakes, and coverage that any KG of this kind has. The point here is that no scholarly KG is able to have perfect coverage - actually, all of them are usually incomplete by definition, either because they explicitly put a threshold about what to include and what not to include or because the source material does not allow you to cover everything and/or may contain mistakes. Second, we usually talk about a golden standard in the context of a precise experiment to measure the results of addressing a particular task. In the tasks shown in the paper, how can the authors demonstrate that ScienceON is, for instance, better than OpenAIRE, OpenCitations, and OpenAlex?
- Table 1 in the introduction shows the authors' perceived superiority (in terms of relations and entity types) that ScienceON introduces compared with other KGs. First, I would have expected this table in the related works section. Second, and most importantly, such comparison lacks several used and well-known KGs for scholarly data developed and used systematically in the community, for instance, OpenCitations, OpenAIRE Graph, OpenAlex, CORE, and the Ukrainian Open Citation Index, to mention a few. Adding these additional KGs would make the comparison fairer and show how things are handled in the KGs (publishers, abstracts, several different types of publication entities, etc.) that are missing in ScienceON KG.
- Another aspect that needs to be clearly stated is the coverage of the data compared to other KGs. In the literature, there are KGs that are either multi-disciplinary or mono-disciplinary, based primarily on English scholarly literature, containing information about a specific country, etc. According to my understanding reading the text, it seems that ScienceON is primarily dedicated to South Korean scholarly literature, which is good since no other database provides such coverage of the research information for that country. However, saying that the authors "designed to construct a comprehensive and systematic KG to address the challenges of data-driven analysis within the science and technology domains" suggests that the resource is better than all the others available online. Is that the case? I think the claim should be softened a bit here.
- The license associated with the data we can see and download via the APIs needs to be clarified. The authors claim they have "free copyright usage," but it is rather unclear what I can do with them. Can I use them and mix them up with other data for research purposes? Do they allow me to engage in commercial activities with them? Thus, it is essential to specify a license to clarify that formally.
- Related to the previous point, I would strongly suggest that the authors and the ScienceON infrastructure provide evidence of following shared and international standards for supporting their "openness" and correct (and expected) availability of the data they provide. To this end, my suggestion would be to evaluate ScienceON against, at least, the Principle for Open Scholarly Infrastructure (POSI), the FAIR principles for data management and stewardship (FAIR), and the TRUST Principles for digital repositories (TRUST). Another well-known and complete assessment framework would be the FOREST Framework for Values-Driven Scholarly Communication (FOREST).

- In section 4, the authors introduce the main component of ScienceON. They say that the architecture provided "facilitates the progression from diverse data sources to the extraction of actionable knowledge". However, the various sources used seem to contain different kinds of data – KISTI Data Center refers to papers and authors, AccessON refers to open access information, KIPRIS contains patents, etc. Thus, even if there are several data sources, they seem complementary, i.e., the same information does not come from diverse sources. This approach essentially simplifies the ingestion of new data since, if you use multiple sources containing similar information, one has to handle the possibility that entities (e.g., papers) are present in different sources and, to ingest such data correctly, one has to develop deduplication approaches that increase the level of complexity of the ingestion process. For instance, this is the case for both OpenCitations and OpenAIRE. Thus, the question is: have the authors developed approaches to handle these situations?
- According to what is described in the text, it seems that only DOIs are considered for identifying papers included in ScienceON. Thus, what happens to all the papers that do not necessarily have a DOI? Are they excluded? Are they handled in some way? If so, how? In addition, is there a plan to also consider other relevant identifiers? Also, are only papers and journals considered, as mentioned in Table 2? What about books (which are still the primary publication in the humanities domain, for instance)? Moreover, how are the books identified if only DOIs are used, considering that books usually come with one or more ISBNs associated?
- When you talk about authors and affiliations associated with papers, it is not clear how it is handled in the data. Technically speaking, affiliations are a tripartite relation that connects an article with an author having a particular affiliation specified for such an article. How this is handled in the RDF? According to the diagram in Figure 4, there is only a relation between an author and an organisation. If so, how can I answer a simple question like: what is John Doe's affiliation within the context of paper A? For more insight into this problem, I would suggest seeing https://doi.org/10.1145/2362499.2362502.
- According to the documentation and links provided, I needed to see explicit documentation about the data model used to organise and expose the data. Is there an ontology? Is it documented? The only documentation seen is that in Figure 4, which is just introductory and does not provide any example of data definition nor a precise definition of ontological terms. Even following the ontology URL (https://scienceon.kisti.re.kr/ontologies/skg#), I could not find any ontology to look at. Where has it been defined? I want to stress that having an open data model (implemented as an ontology) is crucial for claiming that the provided data are FAIR compliant - a crucial endeavour today when infrastructures make available research information. In addition, did the authors use a particular methodology for developing the ontology? How can they claim that the ontology developed is of sufficient quality?
- To retrieve the journal categories, the authors said they scraped them from Google Scholar. Supposing these categories are available in the data they publish, it is unclear if they have the legal right to republish this information in their dataset as open material. The authors should carefully check it, and if they have the right to do so, they should also provide information to describe if and when these data can be reused.
- In Section 4.2, the authors say that the "aspect of the ETL process is the assignment of unique identifiers to each document within ScienceON, which is essential for eliminating duplicate". However, it needs to be clarified how this assignment works, what the shape of these identifiers is, and how the authors are sure that two entities (e.g., two

papers) mentioned in different sources refer to the same object. The process adopted here should be carefully explained since it is an essential passage in the production of any scholarly KG.

- Why is there the need to pass through a conversion from a relational database to an RDF triplestore? In particular, what was the rationale for handling all the data in a relational database? Wouldn't it be better to store data directly in the triplestore?
- I think the claim "ScienceON Knowledge Graph stands as a testament to the potential of semantic web technologies in facilitating advanced knowledge discovery" is a bit overstated. What about all the other efforts developed by several infrastructures and projects in the past years?
- From the diagram in Figure 4, journals may have an ISBN. However, the ISBN is an identifier for books, not journals. If this is confirmed, I honestly think the data model developed is kind of inconsistent with reality.

Minor and typo:

- a SciGraph -> SciGraph

Thus, even if I consider the ScienceON KG a valuable contribution, as explained at the beginning, I think that the authors should go through an extensive rewriting of the text to make it more focused on the details related to the KG construction, the issues, and lessons learned related to it. They should provide more and better comparisons with existing scholarly KGs to clarify what the added value of ScienceON compared to the existing literature.