Context-Compatible Information Fusion for Scientific Knowledge Graphs

Hermann Kroll(✉), Jan-Christoph Kalo, Denis Nagel, Stephan Mennicke, and Wolf-Tilo Balke

Institute for Information Systems, TU Braunschweig, Braunschweig, Germany
{kroll,kalo,mennicke,nagel,balke}@ifis.cs.tu-bs.de

Abstract. Currently, a trend to augment document collections with entity-centric knowledge provided by knowledge graphs is clearly visible, especially in scientific digital libraries. Entity facts are either manually curated, or for higher scalability automatically harvested from large volumes of text documents. The often claimed benefit is that a collection-wide fact extraction combines information from huge numbers of documents into one single database. However, even if the extraction process would be 100% correct, the promise of pervasive information fusion within retrieval tasks poses serious threats with respect to the results' validity. This is because important contextual information provided by each document is often lost in the process and cannot be readily restored at retrieval time. In this paper, we quantify the consequences of uncontrolled knowledge graph evolution in real-world scientific libraries using NLM's PubMed corpus vs. the SemMedDB knowledge base. Moreover, we operationalise the notion of *implicit context* as a viable solution to gain a sense of *context compatibility* for all extracted facts based on the pair-wise coherence of all documents used for extraction: Our derived measures for context compatibility determine which facts are relatively safe to combine. Moreover, they allow to balance between precision and recall. Our practical experiments extensively evaluate context compatibility based on implicit contexts for typical digital library tasks. The results show that our implicit notion of context compatibility is superior to existing methods in terms of both, simplicity and retrieval quality.

Keywords: Implicit context · Knowledge graph · Digital libraries

1 Introduction

Knowledge graphs have revolutionised the access to entity-centric information on the Web, with *Google's knowledge graph*[1] and the *Wikidata knowledge base* [19] being prime examples. One reason is that the old 'Web of Documents' is more and more turning into a 'Web of Linked Data', which needs new access methods

[1] https://developers.google.com/knowledge-graph/.

M. Hall et al. (Eds.): TPDL 2020, LNCS 12246, pp. 33–47, 2020.
https://doi.org/10.1007/978-3-030-54956-5_3

beyond IR-style keyword search: entity-centric information needs to be structured, disambiguated, and semantically enriched by information from various sources. Thus, also in the well-curated domains of digital libraries, a trend to augment document collections to semantically enriched content bases is clearly visible. Especially in scientific libraries *Big Scholarly Data* in heterogeneous form (see [21] for a good overview) is exploited for value-adding services, such as related work recommendation, expert search, or information enhancement using specialised entity-centric databases, like *DrugBank*[2] or *UniProt*[3]. The ultimate vision currently is to extract facts from complete digital collections into one comprehensive knowledge graph for science, supporting complex information needs and offering a variety of additional services, see e. g. [1,7,18].

Yet, the question whether a document collection may still offer more than a collection of extracted facts was already raised at an early stage. An obvious problem concerns the *trustworthiness* of sources: there is a long-standing discussion about the actual truth or plausibility of extracted facts and how well they match with facts extracted from other sources [14]. Thus, keeping lineage or provenance information and respective reputation scores as metadata for each fact is vital [2]. A second class of problems is created by errors in the *algorithmic processes* necessary for fact extraction from natural language texts, covering entity recognition, disambiguation and linking, as well as reliable relation extraction, see e. g. [15]. In fact, all tasks in this process are still error-prone, and even small errors may quickly spoil the overall quality in knowledge graphs [10].

However, even if all these problems were solved, there would be still a major, yet rarely discussed issue: the general *validity* of facts. With respect to general fact validity, current knowledge graphs on the Web vastly differ from those used in scientific digital libraries. Whereas entity-centric data in typical Linked Open Data sources on the Web may or may not be correct, it still tends to be *generally* valid, as e. g. the *birthdate of a person* or *which actors played in some movie*. In contrast, entity-centric data reported in scientific digital collections is often more problematic. Consider for instance different medical treatment options with some active ingredient. They depend on many caveats: general concerns, unresolved discourses in the community, the specific disposition of an actual patient, etc. Another prime examples are clinical trials: even if they are methodically sound, their results can only be considered valid *within the limited context* investigated by each trial. Thus, given the problems to properly control studies currently the generalisability of facts extracted from clinical trials is difficult to assess.

Assume we extract the fact (`simvastatin`, `causes`, `rhabdomyolysis`) from some document reporting on a simultaneous treatment of patients with simvastatin and amiodarone. As the resulting interaction indeed may lead to rhabdomyolysis as a side effect, the information is correct. In the same fashion, we may correctly extract the fact (`simvastatin`, `treats`, `arteriosclerosis`) from some other document on treatment options for arteriosclerosis. But if we now use the combined knowledge graph to query the side effects of *simvastatin in*

[2] https://www.drugbank.ca.
[3] https://www.uniprot.org.

treating arteriosclerosis, we run into trouble: the fact that *simvastatin causes rhabdomyolysis* is not valid *in general*. It is only valid *within the context of simultaneous treatment with simvastatin and amiodarone*. Thus, without having facts restricted by their exact context, a free combination with other facts from the knowledge graph may at least be questionable, if not plain false. Yet, current extraction procedures do exactly this: after long years of standardisation, knowledge graphs typically store facts as simple RDF-triples [3]. This way, tearing facts out of documents and putting them into a knowledge graph means losing all contextual information. If such knowledge graphs are later used for tasks like knowledge discovery, question answering and querying, serious errors can be foreseen. The central question in designing knowledge graphs for digital libraries is thus: *How can knowledge graphs maintain a sense of context for their individual collection of facts?* And concerning later applications: *How can we combine individual facts or even completely merge fact collections while still maintaining their contexts?*

When working with RDF-triples, the *technical* solution for adding context information mostly relies on reification of triples. But how is the correct context for each fact determined? To overcome this problem, two approaches are common: 1. In the community project Wikidata, uploaders are also responsible for supplying all necessary contextual information as additional triples, called qualifiers [19]. 2. In cases where clear-cut contexts can a-priori be determined for some field, the direct modelling and extraction of n-ary relations from document collection are possible [6].

Yet, in both cases, the context needs to be modelled *explicitly*. In this paper, we harness valuable work in the digital library community on standardising provenance and bibliographic metadata (such as authors or keywords) to derive a novel *implicit*, i. e. document-based context model for knowledge graphs. Documents like scientific papers interweave facts in complex contexts and can be assumed to be intrinsically coherent, e. g. by describing all relevant assumptions, methods, observations and conclusions. Thus, for all facts our model takes advantage of the respective extraction documents' characteristics and uses them as an implicit context for facts. Such implicit contexts ensure that given a retrieval problem, only facts from a coherent group of documents can be combined to produce a valid result. Indeed, our experiments show that restricting the information fusion process of knowledge graphs to (restricted) document contexts has a high impact on the number and quality of possible candidates. In addition to structural requirements (graph matching), we consider the context approximated by documents sharing different characteristics to produce valid answers to a query. To improve the result quality for any given query, we operationalise and analyse metrics to find documents having **compatible** contexts. A context compatible set of documents can then be used to obtain better results in terms of validity for tasks like knowledge discovery and querying. We analyse our document-based implicit context model in Sect. 3 and provide a detailed experimental analysis in Sect. 4. Our contributions are:

1. We design and discuss a novel implicit context model suitable for digital libraries. We demonstrate the superiority of implicitly capturing contexts for a real-world knowledge graph in the medical domain.
2. Further, we introduce the concept of context compatibility, i. e. we extend strict document contexts to compatible contexts, increasing the recall for practical applications.
3. We publish all of our scripts as well as evaluation data and results in a publicly available GitHub repository[4] for reproducibility.

2 Related Work

Literature-based Discovery is a well-known and highly discussed topic, i. e. inferring new knowledge based on the current state of literature [16]. In this work, we focus on the application of scientific knowledge graphs for digital libraries. Contextualisation of data can be realised by adding additional contextual information to an individual statement or fact. Regarding RDF, this means to incorporate triples into the knowledge graphs that capture information about a specific triple already existent in the data. Ideas on how to represent contextual information in RDF are provided in [13]. This process is called reification of RDF data [8]. It is realised by introducing a new resource, referencing the reified triple in other statements.

Qualifiers for Contextualising Knowledge. Wikidata, the most extensive open knowledge base on the Web, tries to reify pure RDF facts by using so-called *qualifiers* [19]. Qualifiers add information to a fact by appending a property-value pair directly to it. An example fact (`simvastatin`, `causes`, `rhabdomyolysis`) may further be described by an additional qualifier, namely `when simultaneously used with` along with the respective value `amiodarone`. The qualifiers claim that *simvastatin causes rhabdomyolysis* only, in a simultaneous treatment with *simvastatin* and *amiodarone*. Thus, qualifiers may be used to add additional provenance and sometimes contextual information to simple RDF facts [9]. Even though Wikidata comprises around 30 million qualifier statements (10-2018), they are hardly used to express context for scientific facts, i. e. drug-disease treatments. Even more, only about 5% of all statements are qualifiers (573 million statements). Qualifiers are often restricting the statement they are referring to in a temporal manner, e. g. using the `start time` qualifier. Besides, they may add some provenance information such as references or citations to the statements. In other cases they state information that has no impact on the validity of the fact in question, e. g. the `determination method` is simply used with qualifier values like *chronometry* or *questionnaire* without affecting the validity of its fact. Using qualifiers in joining facts has no precise semantics, e.g. how can we decide whether two qualifiers describe the same context? The curation of explicit contexts is a huge task and moreover, working with explicit context models in practice is unclear.

[4] https://github.com/HermannKroll/ContextInformationFusion.

N-ary Fact Extraction. An extension of extracting binary facts is to harvest n-ary facts [6]. In a large-scale experiment, the authors prove that n-ary facts are more precise than just using binary facts [6]. Thereby, it is possible to explicitly extract and store the context of relations in a higher level relation. For our previous drug and side effect scenario, we may easily design a ternary relation capturing drug, the cause as well as the interacting drug: $causes \subseteq drug \times sideeffect \times interacting\ drug$. However, how good is n-ary fact extraction in practice? Ernst et al. extracted the relation *AthleteWonAward* from a news corpus consisting of 2.8 million documents with about 112 million sentences [6]. They mined 3804 binary, 1089 ternary, 224 4-ary, 23 5-ary and two 6-ary instances of this relation with their best configuration regarding precision. Even though n-ary facts are a promising idea to capture the context of facts, obtaining such n-ary facts is a difficult task, because it requires manually defining the context for every single relation by defining its arity, its domains and its semantics upfront. This is a very strong restriction because considering any possible context of some relation a priori is close to impossible.

Provenance. Another understanding of contexts is provenance, which mainly focuses on storing information attached to the actual fact [17]. The scope of provenance thereby ranges from storing only the explicit source document over additionally storing information related to its creation process such as the author or release date [20]. Provenance can then help to argue about the quality and trustworthiness of the statement in question. Provenance can be integrated into knowledge graphs by using Named Graphs [5]. These are linked to individual facts by extending RDF triples to form N-Quads [4]. In the last years, much work was spent on developing the so-called Prov-O Ontology Description [12]. Prov-O enables knowledge graph designers to encode and store arbitrary information, such as context, for knowledge graph facts. Unfortunately, Prov-O requires users to spend much work on manually providing this additional information, i. e. Prov-O comes with a similar problem as qualifiers in Wikidata. There is yet no solution to automatically reuse context information in the fusion process of knowledge graphs. As far as we know, there exists no practical evaluation of using contexts in typical knowledge graph tasks. With the introduction of our document-based implicit context model and evaluation on a real-world scenario, we extend the current state of literature by giving a practical solution to retain context for digital libraries. Therefore, already applied techniques like Prov-O, Named Graphs, as well as reification, may simply be used as an implementation providing the necessary context in the form of document references for our implicit context model.

3 Implicit Context

Instead of modelling contexts explicitly, textual documents (i. e. research papers) serve as contexts for knowledge graph facts. A scientific publication interweaves facts in assumptions, methods, observations and conclusions. Thus, the argumentative story of a scientific document provides all relevant context variables

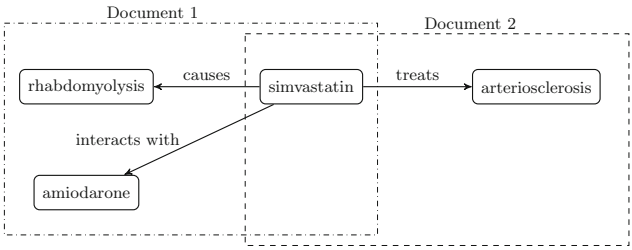


Fig. 1. Implicit context representation for a knowledge graph

implicitly, validating its contained facts. We assume scientific documents to come with a single context, e. g. clinical trials analyse drugs under stable conditions. Indeed, surveys and scientific papers might include several contexts, e. g. describing related work. For this paper, we assume that scientific knowledge graphs should be built by extracting facts out of the paper's main argumentation, i. e. skipping sections such as related work in the extraction process. For our running example, the document provides vital information that simvastatin only causes rhabdomyolysis, when the person is simultaneously treated with amiodarone. Here, the document itself implicitly defines and, thereby, determines the context of interest, because we assume the extracted facts to participate in the main argumentation of the paper. If we mine facts from a single document, then all extracted facts from this document naturally share the same context. The information fusion process by combining/joining facts from the same document to answer a query automatically leads to valid facts because they stem from the same context. In the scientific domain, this context often boils down to conclusions being observed under the same experimental conditions. Therefore, returning to our running example, we define the implicit context of a fact as the document it stems from, see Fig. 1 as an example.

When using a **strict implicit context**, we restrict the combination of facts to those facts within the same **context**, i. e. to facts extracted from the exact same document. Applied to our example, we obtain either that simvastatin treats arteriosclerosis, or that simvastatin causes rhabdomyolysis. We would not obtain the wrong side effect rhabdomyolysis in an arteriosclerosis treatment because there is not a single document validating it.

3.1 Context Compatibility

Obviously, restricting the fusion process of knowledge graphs to strict implicit context will have a substantial impact on the number of obtained results, because we combine facts stemming from the same document only. In addition to strict implicit contexts, we may assume that two scientific documents on simvastatin share the same context, e. g. they describe clinical trials analysing an arteriosclerosis treatment using simvastatin. Since both papers are clinical trials with the same experimental conditions, it seems promising that a combination of facts

from both documents leads to valid query results. Hence, inferring new knowledge between different documents may also be possible. Our idea extends the restriction on pure document contexts to context compatibility ranging over sets of documents. This will lead to broader contexts and allows for a less restrictive combination of facts. Two documents d_1 and d_2, sharing the same context in the above-mentioned sense, will be denoted as **context compatible**: $d_1 \sim d_2$. Thereby, we require $\sim$ to be a reflexive binary relation over the document collection, i. e. one document is always compatible with itself. Combining facts from different but context compatible documents shall yield valid query results.

Comparing the contexts spanned by two or more documents directly is a tedious and time-consuming task that requires a deep understanding of documents' domains. Here, we use different metrics to approximate the context compatibility of documents. In digital libraries, a collection of documents typically provides valuable metadata information. Subsequently, we design two different kinds of similarity metrics to assess the context compatibility of documents: 1. metrics, which directly work on metadata information like authors and curated keywords, and 2. metrics, which build upon textual similarities for titles and abstracts. We choose a threshold-based classification approach to estimate whether two documents are context compatible or not. If the similarity value, computed by a metric, between two documents is above a threshold t, we assume the documents to have a compatible context. Thus, we can safely fuse the facts of two context compatible documents to form a valid answer.

Definition 1. *Let sim be a similarity metric between documents and $t \in \mathcal{R}$ a threshold value. Two documents d_1 and d_2 are context compatible, denoted by $d_1 \sim d_2$, if $sim(d_1, d_2) \geq t$.*

Metadata-Based Similarity Metrics. In scientific contexts, researchers typically work on a specific research field, e. g. a group of medical experts are researching *drug interactions with simvastatin*. They might write several publications about their findings based on similar assumptions like *experimental conditions*. Thus, we assume papers, written by the same authors, to have compatible contexts. We formulate the first metric sim_{author} to estimate context compatibility by using the Jaccard coefficient between the authors of documents. Since contexts of facts should be compatible, if they comprise similar assumptions or experimental designs, we try to capture this intuition by relying on the valuable manually curated metadata available for medical documents. In PubMed, documents are annotated with manual curated mesh headings and chemicals. A mesh heading is a mesh term describing medical entities, actors, processes and concepts like *humans*, *pain*, *trial* and *simvastatin*. The mesh headings, therefore, might capture the context that is given by a document. The second metric sim_{mesh} is defined as the Jaccard coefficient of the documents' mesh headings. Similarly to the mesh terms, we use the chemicals annotated to documents as an approximation for context compatibility. Therefore, $sim_{chemical}$ is defined as the Jaccard coefficient of the documents' chemicals.

Text-Based Similarity Metrics. In addition to the metadata-based approaches, we also try to capture the context compatibility by measuring textual similarities among the documents' texts. Here, sim_{title} is defined as the Jaccard coefficient between the titles of two documents to estimate the text-similarity between documents. The previous similarity metrics can only be applied to pairs of documents for determining context compatibility. To further extend fact fusions to more than a pair of documents, we suggest to also directly determine the compatibility between multiple documents by clustering documents into **context compatible sets** such that all documents inside such a set are pairwise context compatible. Given the respective documents the facts in the knowledge graph stem from, we use a clustering method to produce groups of documents with compatible contexts. Here, we use textual information, i. e. titles and abstracts of documents. We select a common method to cluster documents to understand whether compatible document sets are helpful: 1. We extract the titles and abstracts of documents. Thereby, we remove stop words and apply stemming. 2. We compute the TF-IDF matrix upon the texts. Words which occur very frequently or words which occur very rarely are removed. 3. Clustering documents with various texts requires much computational power. Thus, we use a principal component analysis (PCA) to reduce the number of dimensions to 300. 4. Finally, we apply a k-means++ clustering on the reduced matrix with different k values.

4 Analysis on SemMedDB

In the following experiments, we evaluate whether restricting fact combinations to their document contexts is capable of producing valid facts for typical medical queries. We perform a comparison to querying a knowledge graph without contextual information, allowing us to join arbitrary facts. In our expectations, using implicit context should increase the quality of query results substantially, while reducing the overall number of results. For the evaluation, we compare the number and quality of results for typical queries on a large medical knowledge graph called *SemMedDB* by using no context as a baseline and our implicit context models.

SemMedDB is a fact-based database consisting of medical entities and relations between them [11]. A fact mining process automatically extracted all facts from abstracts and titles of documents in PubMed. For each extracted fact in *SemMedDB*, a reference to its source document is retained. Hence, *SemMedDB* provides provenance information. We use *SemMedDB* 2019[5] in version *semmedVER40R*. This version comprises 20,124,700 distinct facts extracted 97,972,561 times. We design three experiments to compare the usage of *SemMedDB* as a knowledge graph without context on the one hand and with implicit context on the other. The experiments are built on three scientific queries, and are also depicted in Fig. 2: 1. Knowledge discovery via querying using the `causes` relation, 2. Predicting drug-drug interactions via a gene (like already performed by

[5] https://skr3.nlm.nih.gov/SemMedDB/.

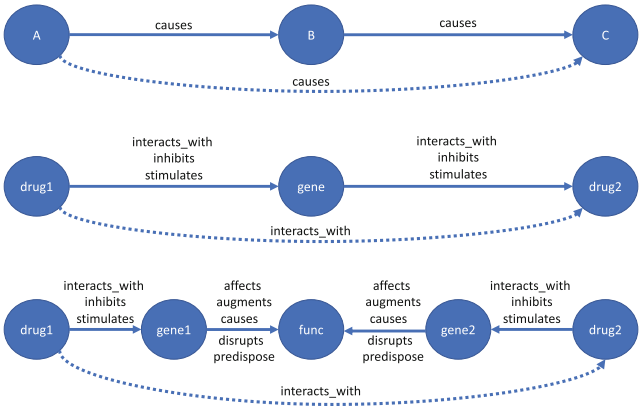


Fig. 2. Graph Patterns to derive new facts in SemMedDB. The dotted edge depicts the new derived fact

domain experts [22]) and 3. Predicting drug-drug interactions via a biological function (like already performed by domain experts [22]).

Transitive Causal Relation (Causes). `Causes` is used to express a relation between a cause and an effect of medical concepts, e. g. a drug and a disease. Since this relation is usually assumed to be transitive, the goal in this knowledge discovery task is to query for new facts by joining two existing causal facts from the knowledge graph. As an example, the facts (`simvastatin`, `causes`, `risk of heart disease`) and (`risk of heart disease`, `causes`, `heart failure`) may be joined to obtain the new fact `(simvastatin, causes, heart failure)`. To increase the quality of these facts, we select only facts appearing in at least three documents, yielding 153,024 distinct facts extracted 1,584,676 times from documents.

Predicting Drug-Drug Interactions (DDI). In a second experiment, we rely on a known approach for finding drug-drug interactions using *SemMedDB* [22]. Such an interaction may cause several side effects in a patient's treatment. Thus, finding these new interactions is a relevant task for medical experts that can be easily supported by knowledge graphs. Drug-drug interactions are discovered using two queries as described in [22]. We call these interactions *DDI-G*, a drug-drug interaction via a gene and DDI-F, a drug-drug interaction via a function.

Estimating the Result Quality. To be able to perform the evaluation, we take *SemMedDB* as the gold standard of medical knowledge and assume that it is 100% correct and also complete. As far as we know, there is no medical source comprising more medical domain knowledge than *SemMedDB*. *SemMedDB* contains a dedicated `causes` predicate and `interacts with` predicate between drugs. Thus, we count how many derived facts are contained in *SemMedDB* already and how many of them are correct. To estimate the recall, we take the

Table 1. Number and quality of newly distinct obtained facts by querying a knowledge graph without context and with strict implicit context

Graph	#Obtained facts	#Correct	Precision	Recall
Knowledge Graph (Causes)	7,978,099	95,037	1.19%	**100%**
Strict Implicit Context (Causes)	11,478	5,544	**48.3%**	5.83%
Knowledge Graph (DDI-G)	753,899	55,370	7.34%	**100%**
Strict Implicit Context (DDI-G)	1,311	909	**69.3%**	1.64%
Knowledge Graph (DDI-F)	18,685,416	148,346	0.79%	**100%**
Strict Implicit Context (DDI-F)	2,138	1,352	**63.2%**	0.9%

number of query answers on the knowledge graph without restricting fact combinations as an overestimation of the number of all correct results. Thereby, we overestimate the recall of the knowledge graph as being 100% and compare the remaining approach to that number. We underestimate the precision, because there may exist correctly derived facts, which are not included in our ground truth (the knowledge graph itself).

4.1 Strict Implicit Context

For the knowledge graph query experiments, we have no restrictions when joining facts and just perform a simple pattern matching from the query to the knowledge graph. In contrast, when using strict implicit context, we restrict fact combinations to the document contexts, i. e. combinations of facts are only possible within the context of a document. The number and quality of obtained results by using no context in comparison to using strict implicit context for all three tasks (causes, DDI-G and DDI-F) are listed in Table 1. The number of facts obtained from the baseline, a knowledge graph without context, differs by orders of magnitude compared to the knowledge graph with strict implicit context in all three experiments. However, the results only come with a precision of 1.19% (causes), 7.34% (DDI-G) and 0.79% (DDI-F) by using no context and 48.3% (causes), 69.3% (DDI-G) and 63.2% (DDI-F) by using strict implicit context. The recall decreases from 100% to 5.83% (causes), 1.64% (DDI-G) and 0.9% (DDI-F).

Discussion. In sum, using strict implicit document-based contexts outperforms the plain knowledge graph (no context) approach for all three experiments with regard to the precision. However, strict implicit context restricts the derivation process of facts to single document contexts, and thus a considerable amount of incorrect, but also some correct results are not returned. This leads to a lower recall in comparison to joining arbitrary facts. When querying a knowledge graph, a high degree of correctness is often needed. Particularly if medical experts need to verify drug-drug interactions in studies, high-quality results are desired.

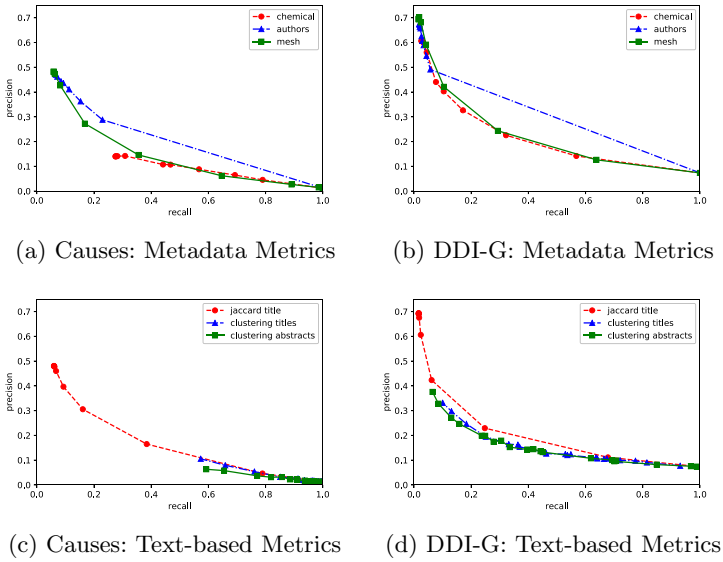
4.2 Context Compatibility

We design context compatibility to increase the recall for different tasks in comparison to strict implicit context by allowing the fusion of facts stemming from compatible document contexts. Our evaluation comprises six different approaches for context compatibility on two different medical queries. Three of the approaches work purely on the metadata (i. e. chemical, mesh headings and authors) and three approaches work with textual measures (i. e. Jaccard coefficient between titles, clustering of titles and abstracts). The two queries are the causes query from Fig. 2 at the top and the DDI-G query depicted in Fig. 2 in the middle. Unfortunately, we have to skip the third experiment (DDI-F) here due to performance issues. In the DDI-F experiment, the knowledge graph produces around 18 million facts. Checking the context compatibility between documents, validating a fact derivation, leads to too many different combinations. For all our experiments, we evaluate different thresholds and k-values to report our findings as precision-recall curves. We check different thresholds (0 to 1.0 by a step size of 0.1) and 20 different k values ranging from 2 to 100,000. Additionally to the results presented in this paper, more experimental results can be found on our GitHub repository. To perform our experiments, we have accessed the metadata and texts of PubMed documents by downloading the latest version of the PubMed Medline 2019 as an XML dump[6], which provides title, abstracts and valuable metadata.

Causes Experiment. Figure 3 (a) depicts the precision-recall curve for the cause experiment using metadata similarity metrics. Note that selecting a threshold of 0.0 leads to the same result as using the knowledge graph approach without contextual restrictions and 1.0 leads to similar results as using strict implicit context. We achieve the best possible precision of about 48% with a recall of about 6% by using a threshold of 1.0 for sim_{mesh} and $sim_{authors}$. A higher recall is achieved when using $sim_{chemicals}$ because 53% of all documents provide curated chemicals, whereas the other metadata is less common. We obtain the best F1-Score of 25.5% (28.8% precision and 23% recall) for $sim_{authors}$ with a threshold of 0.1. Although sim_{author} outperforms the other metrics regarding precision and recall, sim_{author} provides only a small recall range. 9 of 10 thresholds for sim_{author} yield a recall below 23% and the last threshold yields 100% recall. Computing more fine-grained thresholds would not help here, because most of the papers have only a few authors yielding a small range of different Jaccard coefficients.

The results of our text-based approaches for context compatibility are depicted in Fig. 3 (c). Here, the clustering methods on titles and abstracts share a similar shape; hence they have a comparable performance. Variations of the number of clusters can cover a range of recall values between 0.6 and 1.0 while keeping an acceptable precision of around 10%. Hence, the methods can boost the precision of the knowledge graph 10-fold, while only sacrificing around 40%

[6] https://www.nlm.nih.gov/databases/download/pubmed_medline.html.



(a) Causes: Metadata Metrics (b) DDI-G: Metadata Metrics

(c) Causes: Text-based Metrics (d) DDI-G: Text-based Metrics

Fig. 3. Precision-recall curve of the experiments (Causes and DDI-G) by using different metrics to estimate the context compatibility between documents

of recall. In contrast, the Jaccard-based similarity sim_{title} outperforms the clustering methods (denoted as jaccard title in the plot). The approach achieves a comparable precision for high recall values. Besides, it is possible to achieve even higher precision, for sacrificing some correct results at lower recall values by achieving a precision of almost 50% at a recall of 10%.

Overall, we can summarise that sim_{author} and sim_{title} achieve the best results for the causes experiment. While sim_{author} performs better regarding precision, sim_{title} offers to select a broader range of recall values.

DDI Gene Experiment. Figure 3 (b) depicts the precision-recall curve for the DDI-G experiment using metadata similarity metrics. Again, $sim_{authors}$ outperforms the other metrics, e. g. selecting a threshold of 0.1 yields a precision of 49% and a recall of 6%. Compared to strict implicit context, the precision decreases from 69% to 49%, while the recall increases from 1.6% to 6%. Thereby, 9 of 10 thresholds for $sim_{authors}$ yield a recall below 6%. In this experiment, $sim_{chemical}$ performs better than in the causes experiment. We obtain the best F1-Score of 26.5% (22.6% precision and 32.1% recall) for $sim_{chemicals}$ with a threshold of 0.2. We assume that a chemical-based similarity fits best for a drug-based query.

We depict the precision-recall curve for the DDI-G experiment using text-based similarities in Fig. 3 (d). Again, the clustering methods on titles and abstracts share a similar shape. In comparison to the causes experiment, the clustering approaches provide a broader range of recall values with higher precision. The Jaccard-based similarity sim_{title} outperforms the clustering methods.

Similar to our previous experiments, all approaches boost the precision of the knowledge graph, which was around 7%, while keeping good recall values. Overall, for the DDI-G experiment, we can summarise that sim_{author} and sim_{title} achieve best results.

Discussion. All techniques for context compatibility can boost the poor quality of query answers on knowledge graphs by at least one order of magnitude while being able to retain high recall. Furthermore, the techniques offer much more flexibility than the knowledge graph without context and with strict implicit context alone by providing the possibility of choosing between precision and recall, depending on the application.

5 Conclusion

In this paper, we highlighted the importance of retaining document contexts for supporting typical knowledge graph tasks for digital libraries. Indeed, document context proves crucial for proving the validity of facts, especially, in scientific domains such as biomedicine or pharmacy. Moreover, we introduced *implicit contexts* using documents as an approximation of contexts and evaluated them in combination with compatible contexts for different tasks. Our experiments show the applicability and feasibility of document-driven contextualisation for tasks like knowledge discovery and querying in practice. Approximating contexts at the document-level offers an easy-to-use and, likewise, high-quality opportunity to maintain context in knowledge graphs. Storing techniques like Prov-O, Named Graphs and N-Quads are already ready-to-use and established fact mining processes may easily be extended by maintaining a reference for each fact to its source document, but nothing more. Providing context compatibility between documents might be as simple as designing metrics for already available metadata in digital libraries. This technique leads to an apparent increase of recall when using implicit contexts, but would not deny the valuable context given by librarian documents.

As future work, we would like to investigate measures for *story-based* similarity between documents and to evaluate their usefulness for context compatibility. The *story* of a document is related to its argumentation plus their contextual settings. We believe that a story-based similarity measure would improve the previously described similarity metrics in different tasks.

References

1. Auer, S., Kovtun, V., Prinz, M., Kasprzik, A., Stocker, M., Vidal, M.E.: Towards a knowledge graph for science. In: Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics. WIMS 2018. ACM (2018)
2. Bechhofer, S., et al.: Why linked data is not enough for scientists. Fut. Gener. Comput. Syst. **29**(2), 599–611 (2013)
3. Candan, K.S., Liu, H., Suvarna, R.: Resource description framework: metadata and its applications. SIGKDD Expl. **3**(1), 6–19 (2001)

4. Carothers, G.: RDF 1.1 N-Quads. https://www.w3.org/TR/n-quads/ (2014)
5. Carroll, J.J., Bizer, C., Hayes, P., Stickler, P.: Named graphs, provenance and trust. In: Proceedings of the 14th International Conference on WWW, WWW 2005, pp. 613–622. ACM (2005)
6. Ernst, P., Siu, A., Weikum, G.: Highlife: higher-arity fact harvesting. In: Proceedings of the 2018 World Wide Web Conference, WWW 2018, International World Wide Web Conference on Steering Committee, pp. 1013–1022 (2018)
7. Fathalla, S., Vahdati, S., Auer, S., Lange, C.: Towards a knowledge graph representing research findings by semantifying survey articles. In: Kamps, J., Tsakonas, G., Manolopoulos, Y., Iliadis, L., Karydis, I. (eds.) TPDL 2017. LNCS, vol. 10450, pp. 315–327. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67008-9_25
8. Hayes, P.J., Patel-Schneider, P.F.: RDF 1.1 Semantics. https://www.w3.org/TR/rdf11-mt/##whatnot (2014)
9. Hernández, D., Hogan, A., Krötzsch, M.: Reifying RDF: what works well with Wikidata? In: Proceedings of the 11th International Work. on Scalable Semantic Web Knowledge Base Systems. CEUR Working Proceedings, vol. 1457, pp. 32–47. CEUR-WS.org (2015)
10. Kalo, J.C., Homoceanu, S., Rose, J., Balke, W.T.: Avoiding Chinese Whispers: controlling end-to-end join quality in linked open data stores. In: Proceedings of the ACM Web Science Conference, WebSci 2015, pp. 5:1–5:10. ACM (2015)
11. Kilicoglu, H., Shin, D., Fiszman, M., Rosemblat, G., Rindflesch, T.C.: SemMedDB: a PubMed-scale repository of biomedical semantic predications. Bioinformatics **28**(23), 3158–3160 (2012)
12. Lebo, T., Sahoo, S., McGuinness, D.: PROV-O: The PROV Ontology. https://www.w3.org/TR/prov-o/ (2013)
13. Patel-Schneider, P.: Contextualization via qualifiers. In: Workshop on Contextualized Knowledge Graphs co-located with 17th International Semantic Web Conference on, CKG@ISWC 2018 (2018). http://wiki.knoesis.org/index.php/CKG2018
14. Pinto, J.M.G., Balke, W.-T.: Can plausibility help to support high quality content in digital libraries? In: Kamps, J., Tsakonas, G., Manolopoulos, Y., Iliadis, L., Karydis, I. (eds.) TPDL 2017. LNCS, vol. 10450, pp. 169–180. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67008-9_14
15. Shen, W., Wang, J., Han, J.: Entity linking with a knowledge base: issues, techniques, and solutions. IEEE Trans. Knowl. Data Eng. **27**(2), 443–460 (2015)
16. Swanson, D.R.: Complementary structures in disjoint science literatures. In: Proc. of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 280–289. SIGIR 1991, ACM (1991)
17. Tan, W.C.: Provenance in databases: past, current, and future. Bull. IEEE Comput. Soc. Techn. Committee Data Eng. **30**(4), 3–12 (2007)
18. Vahdati, S., Palma, G., Nath, R.J., Lange, C., Auer, S., Vidal, M.-E.: Unveiling scholarly communities over knowledge graphs. In: Méndez, E., Crestani, F., Ribeiro, C., David, G., Lopes, J.C. (eds.) TPDL 2018. LNCS, vol. 11057, pp. 103–115. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00066-0_9
19. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Commun. ACM **57**(10), 78–85 (2014)

20. Wylot, M., Cudré-Mauroux, P., Hauswirth, M., Groth, P.: Storing, tracking, and querying provenance in linked data. IEEE Trans. Knowl. Data Eng. **29**(8), 1751–1764 (2017)
21. Xia, F., Wang, W., Bekele, T.M., Liu, H.: Big scholarly data: a survey. IEEE Trans. Big Data **3**(1), 18–35 (2017)
22. Zhang, R., et al.: Using semantic predications to uncover drug-drug interactions in clinical data. J. Biomed. Inform. **49**, 134–147 (2014)