Building and Exploiting a Web of Machine-Readable Scientific Facts to Make Discoveries

Nicola Raffaele Di Matteo[1,*], Andrea Schimmenti[2], Fabio Vitali[2] and James Blustein[1]

[1]Dalhosusie University, 6299 South St, Halifax, NS, Canada
[2]Università di Bologna, Via Zamboni, Bologna, Italy

Abstract

We propose a method and the motivation for constructing a Web of machine-readable scientific claims, which computers can use for comparisons and to make new discoveries through the creation of a public structure of Nanopublications representing claims extracted from published papers.

A scientific claim can be represented by a semantic predication, a triple in the form *(subject, predicate, object)*. Information such as provenance can be associated with Nanopublications, making it a valid, atomic, scientific publication. By connecting these claims, discoveries can be made, and machines do this automatically with our structure. As a result, scientific articles published in hypertext are linked to increasing the objective knowledge of the world with valuable new hypotheses.

While semantic predications from previously published scientific papers can be created using NLP tools, humans can extract them from the new articles. Authors, publishers, and readers can accurately identify core statements. They are valuable resources that must be encouraged to create or improve nanopublications.

After having introduced how to make discoveries by connecting claims in the literature and having delineated the structure that makes them readable by machines, we introduce *Desx* as a conceptual method to enable authors to extract semantic predications from their published hypertext manuscripts and contribute to the global collection of nanopublications efficiently and accurately. We also discuss Desx's envisioned potential to linearize RDF into quasi-plain text for effective human readability, highlighting its future possibilities in knowledge presentation.

Keywords

Semantic Web,, RDF to text,, text to RDF,, nanopublications,, Literature-based Discoveries

1. Introduction

Scientific literature is vast and grows rapidly. More than 50 million scientific papers have been published, and around 2.5 million see the light of day every year [1]. In the last decades, we have seen a substantial increase in publications. It is an extremely fragmented but highly reliable knowledge, thanks to the activity of more than 28,000 peer-reviewed journals [2] and innumerable authors that make accessible their work in different formats, particularly

IRCDL 2024: 20th conference on Information and Research science Connecting to Digital and Library science, February 22–23, 2024, Bressanone, Brixen, Italy

*Corresponding author.
†These authors contributed equally.
nicola.dimatteo@dal.ca (N. R. D. Matteo); andrea.schimmenti2@unibo.it (A. Schimmenti); fabio.vitali@unibo.it (F. Vitali); jamie@cs.dal.ca (J. Blustein)
0001-9618-3830 (N. R. D. Matteo); 0000-0001-7865-7537 (A. Schimmenti); 0000-0002-7562-5203 (F. Vitali); 0000-0003-4347-054X (J. Blustein)

hypertext in recent times. Such literature contains assertions that describe the world and all of our knowledge of it as a collection of (often unrelated, fragmented, and subjective) assertions that are the consequences of the experiments reported and reviewed by peers, and represent objective knowledge built interactively by researchers. A scientific paper in literature is an autonomous part of knowledge that brings elements, assertions, fragments of a puzzle that form a pattern that oftentimes may reveal new hypotheses [3].

With its load of facts and, in general, assertions that can validate new hypotheses, theories to connect, and experiment results, the scientific literature is itself a lode to dig and explore so as to extend the scientific discovery process. Following Karl Popper [4], theories originate from free creations of the mind and speculations of which scientists should imagine and test counter-evidence to assess their robustness. Thus scientific theories are scholarly conjectures that need to be thoroughly tested: those that can be tested in the real world are stronger, and the ones that pass more validations positively are *better theories* than those that have failed the tests or that were not fully and independently validated. In short, published arguments and public criticism are crucial to the advancement of science.

In Popper's view, there are three worlds of reality: physical, objective, and subjective. The objective world is the collection of theories with their validations performed by humans; on the other hand, the subjective world is what is known and experienced by individuals. The scientific literature is objective knowledge: it is created by humans and contains theories, experiments to verify them, and inventions. However, what is contained in literature and the part that each of us know of it is different. There is no certain method to retrieve all the (published) knowledge: there will always be something to discover for individuals, including the novel problems brought on by new theories and inventions. In conclusion, in the literature exists "undiscovered public knowledge" [3, p. 108]. It is sometimes possible for crucial pieces of information to remain undiscovered, e.g., that not all swans are white: a researcher might be unaware of the existence of black swans if no trustworthy source has confirmed their presence. Similarly, a researcher may not be able to establish the hypothesis that A causes C if they know of a paper reporting that A causes B, but misses a connection to another paper affirming that B causes C.

Objective knowledge is a key to finding answers to scientific questions. To achieve it, researchers need not only hypotheses and potential refutations, but also the necessary connections between sources of information. Researchers can arrive at objective knowledge not only via scientific experiments, but also and ever more frequently by exploring and connecting various sources of information, and reach conclusive answers through information networks. Although no systems can be created to retrieve all the knowledge [3], any step to building better tools to recognize and combine hypotheses is a step that increases the subjective and objective knowledge of the world and, therefore, Science.

Studying methods to improve such discoveries and build efficient systems is the scope of the field of Literature-Based Discovery (LBD), a research area born when Swanson [5] showed that connecting results from independently published articles can reveal new and unexpected hypotheses. Navigating objective, published knowledge is possible and beneficial, and studies have proposed different methods and tools to do it automatically. Systems based on the recently proposed relation/predicate-based approach that reason on semantic predications can suggest new hypotheses based on well-defined logical consequences [6] and show the associations that connects assertions [7] efficiently [8, 9]. An example of a discovery deducted by claims

in different papers is reported in Figure 1. Here, the claims EPA (Fish Oil) STIMULATES Epoprostenol, Epoprostenol TREATS RD/RP (Raynaud Disease), are premises for an interesting syllogism that concludes that Fish Oil is highly beneficial to treat Raynaud Disease, a medical condition of reduced blood flow in extremities such as fingers and toes. Notably, the hypothesis was not stated in any published articles [7]. Claims could have also originated from papers belonging to different scientific domains, so they likely being unbeknownst to researchers in either field. As shown in Figure 1, other claims found in different articles suggest another interesting hypothesis: because both Epoprostenol and APA disrupt platelet aggregation, platelet aggregation could cause Raynaud Disease. These consequences could also explain the biological mechanisms. In summary, claims represented in triples and ontologies to associate terms such as Epoprostenol and Prostaglandin make deductions that suggest hypotheses possible.

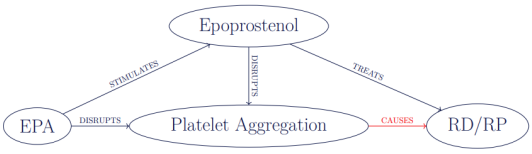

Figure 1: Graphic representations of the syllogisms that conclude that Fish Oil treats Raynaud Disease and that this is possible because the platelet aggregation is reduced.

Despite the effectiveness of making discoveries by reasoning about semantic predicates as reported in several studies (e.g., [8, 9]), almost all the tools presented in these studies are not available anymore. This could suggest difficulties in using them and having an active community that maintains and upgrades them. Indeed, the problem of making discoveries could be positioned at a higher level through an efficient tools' architecture that encourages a divide-and-conquer approach. Such tools can create a heuristic space where interconnected and accountable information can show new questions and unseen correlations for researchers to study.

In the systems proposed in the literature, almost all the LBD components are implemented from scratch: from the analysis of a collection of scientific papers to retrieve significant terms, to the algorithms that find relationships, to the user interfaces, to the methods to rank the results [10]. Using available prepackaged parts such as NLP tools and domain ontologies seems not to be considered in the studies; this leads to considerable, sometimes fruitless, efforts to create and recreate usable tools. Also, this approach does not encourage collaboration between experts in different fields.

Building tools that make discoveries could be straightforward by exploiting semantic predicates on the Web. Decoupling the extractions of assertions in articles from the reasoning about them opens up the opportunity to use standard software components and allow different teams to work together on a specific problem. Publishers and authors could produce publicly accessible assertions [11]. With an ecosystem in which articles as collections of machine-readable facts — predications — and ontologies for the field of interest are published, the objective knowledge represented by the literature could be extracted, analyzed, and discoveries could be made. Also, researchers could contribute to collecting facts with new hypotheses suggested and, eventually, validated in such an environment. The environment can be built and managed with tools defined

for this purpose, to make human knowledge accessible by machines; within the Semantic Web: a collection of technologies and methods to define vocabularies, store, manage, share, query data on the Web, and reason over them [12].

We hereby envision a *Web of Facts*, a publicly accessible collection of (scientific) assertions and hypotheses that computers use to make discoveries. We see it as a collection of nanopublications [13], i.e. RDF Named Graphs that contain an atomic scientific statement together with annotations and provenance. By removing rhetorics and style, a scientific article can be represented as a collection of scientific assertions that can be published as a collection of semantic predications `<subject, predicate, object>`, alongside provenance metadata, ratings of certainty, and so on, to ensure the truth value of such sentences. A scientific claim can therefore be expressed as a triple, e.g. `<aspirin, treats, cancer>`, and annotations report the provenance and epistemic value of the claim itself.

The source of the Web of Facts, the infrastructure of assertions that machines can use, are scientific publications. NLP tools can extract assertions from published articles and represent extracted statements as nanopublications. NLP tools combine techniques such as Transformers-based models fine-tuned for assertions extraction, Named Entity Recognition like SemRep [14] developed by the National Library of Medicine (NLM), Abstract Meaning Representation, and Relationship Extraction. In our view, the Web of Facts is a large-scale repository of nanopublications from different scientific disciplines. In this space, scientists can find, explore and validate new questions. New hypotheses can be generated and, in turn, published as nanopublications.

Both machines and authors, publishers, and readers can participate in identifying the core facts and generate nanopublications.

We propose identifying fundamental methods of a process for making discoveries automatically from the literature, in which the Web of Facts is the layer that permits decoupling the effort to extract predications from papers and, therefore, encourages collaborations between different entities. The methods involve:

(i) Extracting assertions from existing papers, articles, and documents to publish them as nanopublications; (ii) Using logic to derive new and unexpected inferences from such nanopublications; (iii) Publishing the newly discovered knowledge in the form of new nanopublications; (iv) Making this knowledge accessible to humans in a readable format for inspection and evaluation of relevance by the domain experts, and as seeds for further explorations and new research opportunities;

An easy-to-use tool that authors can use to feed the Web of Facts is part of our vision of the Web of Facts Publication Cycle. Through it, publishers can offer open services that convey new users to their premium offerings, giving access to a collection of assertions published alongside the version for humans.

Our contribution is to (i) delineate a methodology to make discoveries automatically, connecting information expressed in scientific articles published in hypertext for humans, (ii) lay the foundation of a Web of scientific facts, an essential part of the view that permits decoupling the effort to extract predications from papers and, therefore, encourages collaborations between different entities, (iii) suggest a new publishing cycle that opens opportunities for publishers while contributing to disseminate machine-processable scientific knowledge, and (iv) propose a method to support the creation of nanopublications manually and visualize to humans the knowledge generally represented in RDF.

2. Previous works and background

The Literature-Based Discovery (LBD) methodology, initially proposed by Swanson [5], connects information expressed in different papers to propose new hypotheses to the researcher. Subsequently, methods to make discoveries using semantic predications extracted from papers automatically have been proposed. In our view, discoveries emerge more efficiently if the different phases of the process are well-distinct and dedicated tools and methodologies are used. Such parts are:

(i) The collection/generation of large quantities of predications representing atomic scientific results extracted from, and deeply grounded in, the scientific articles they were first introduced in. (ii) The adoption of meaningful and applicable relational models between concepts (e.g., ontologies) that supply the possibility to discover and boost unexpected and non-obvious relationships between disconnected predications. (iii) Tools to integrate distributed collections of predications, their comparison and integration, especially in a setting where intellectual property of the scientific articles containing the source facts belong to private enterprises. (iv) Tools to make aware scholars of the newly discovered potential predications in an operable manner, allowing the humans to receive, understand, evaluate, edit and improve of the discovered hypotheses.

Different contributions in the literature about methods that use semantic predications to make discoveries other than for each aspect of interest are given.

Discoveries exploting semantic predications Different methods to make discoveries from the literature have been proposed. Methodologies that exploit semantic predications clearly represent the connections that bring discoveries, leading to better accuracy [8]. For instance, in showing results to determine the most effective chemotherapy for lung cancer treatment, Li et al. [15] points out the convenience of using semantic predications to evaluate hypotheses and make discoveries. Semantic predications are "basic knowledge unit" [15, p. 5] representing the assertions in papers. They can also be filtered considering their level of certainty, including controversial and contradictory assertions.

Using predications extracted from the MEDLINE corpus and grouping the results considering the context defined by the MeSH descriptors is the approach Obvio [9] uses. Relations between a concept of interest and possible causes are searched, connecting predications in a graph and considering as more relevant the shortest paths. The method can easily infer relations not explicitly stated, and discoveries can be grouped by categories such as *cellular activity*, *pharmaceutical*, and *lipids*. The authors conclude that more hypotheses could be recovered with predications extracted from the full-text [7] other than from abstracts and titles.

Melodi [16] is another tool that analyses predications extracted by the MEDLINE corpus with SemRep [17], a natural language processing tool developed by the National Library of Medicine (NLM). Triples that express relations between common terms, i.e., frequently occurring in the corpus, are not considered. On the other hand, paths that include connections between rarer concepts, i.e., less frequent in the corpus and more often stated in the set of articles, are considered more relevant. An improved version of the tool [18] filters the predications by the semantic type of subjects, objects, and predicates.

Semantic Web for discoveries Can Semantic Web triples and ontologies help? Some scholars are positive. For instance, in [19] it is suggested that the Semantic Web has all the resources needed to support finding new drugs. The authors advocate a new Semantic Web Stack in which relevant conclusions from papers and ontologies are published in RDF, knowledge graphs are built, and discoveries are made by reasoning on them.

The authors of [20] point out that because Semantic Web enables knowledge sharing, systems based on it can improve scientific discoveries. By sharing findings in machine-readable format, biological models can be represented, and hypotheses can emerge through logical inferences. With ontologies published in RDF on the Web, scientists can specify complex relations between concepts allowing the combination of knowledge from different fields. Study conclusions can then be described with RDF triples and included in articles, making the integration of RDF statements a part of the publishing process itself. We strongly agree with the idea that including RDF that describes relevant assertions expressed in a paper should be intrinsic part of the publishing process. With such a mechanism, researchers can use previous studies more efficiently to make new discoveries and generate and validate hypotheses.

With the prototype of their tool named HyQue [21], [22] shows that Semantic Web technologies can be used to validate hypotheses. Using ontologies published on bio2RDF [23], the tool validates speculations in the biology fields. Also, they suggest a mechanism to associate provenance to the assertions published on the Web that should reduce the complexity of the nanopublication model.

[24] suggests that Semantic Web technologies can satisfy the need to access knowledge from different sources. In their view, a “virtual knowledge broker service” [24, p. 429] extracts assertions and metadata from heterogeneous sources such as full-text literature and repositories of protein sequences, genes and phenotypes. They point out that managing complex and voluminous data, adopting Semantic Web standards, creating simple user interfaces, having experts in data providing are the challenges to resolve to make, one day, the infrastructure they propose.

Extracting semantic predications Extracting predications from text is a complex task. When approaching Text-to-RDF models, it is essential to consider that the text to be converted must present a statement where a "variable" is given a "value" or, even better, something is described. Not all text is worth converting into RDF, as some may lack meaningful information or structured data that can be effectively represented in a knowledge graph. At an initial level, the abstract of a paper is a good example. The text must be cleaned of *uninteresting* sections, and multiple sentences must be grouped together to summarize the semantics of it. This brings us to two other NLP topics: Abstract Meaning Representation and Natural Language Understanding. Attempts to generate RDF from NL are not new [25], but to date, most approaches use NN. Generating *good* RDF from text has seen its first successful attempts from simple sentences. AMR2FRED, [26] generates RDF with NER and DBpedia. Similar tools employ both NER and AMR to generate RDF. Still, a generally accepted solution has not been adopted generally by the community. Other approaches include adaptations of seq2seq models using graph embeddings, such as CycleGT. [27].

Presenting discoveries: RDF-to-Text RDF-to-Text is an essential task in natural language generation, aimed at converting knowledge graphs into natural language. This field is continually evolving, exploring various methodologies including rule-based, template-based, and neural network-based approaches, each with unique strengths and challenges. No standard solution has yet gained widespread acceptance, a situation influenced by several factors: (i) *Diversity in knowledge graphs*: Varying structures and complexities in RDF graphs make it challenging to develop a universal solution. Machine Learning and Neural Network models, though adaptable in general contexts, struggle with fine-tuning for specific domains. (ii) *Distance between RDF and NL*: The structure of RDF triples, despite having a natural language-like label, differs significantly from natural language constructs, complicating direct verbalization ([28]). (iii) *Advancements in NL generation*: The evolution of techniques, such as those introduced by GPT-3, contribute to the lack of standardization ([29]). (iv) *Differences in output requirements*: The variability in desired outputs based on application or end-user preferences, combined with the dominance of English in these technologies, adds to the complexity ([30]). (v) *Performance vs Precision*: The balance between the general applicability of ML and NN models like GPT-3 and Transformer ([29, 31]) and the precision of template-based models like RDF2PT ([32, 33, 34]) remains a key challenge.

The WebNLG challenges have significantly contributed to the growth of this field. Notably, the 2020 Challenge highlighted the Graph2Seq approach as a promising method for bridging the RDF-NL gap ([28]). Initial tests with GPT-4 for generating text from RDF triples show potential but also reveal issues like the inclusion or omission of key information. Future research might explore fine-tuning models to adhere more closely to input ontologies and graphs.

3. Making discoveries exploiting the Web of Facts

The Web of Facts is a global collection of machine-readable scientific (nano)publications. Objective knowledge is made accessible by machines using such a structure, similar to how it is readable by humans through HTML pages and PDFs. Nanopublications are extracted from papers by NLP tools and humans; they feed a collection made accessible by machines that can generate well-founded and likely valuable hypotheses, being based on transparent deductions from valid scientific assertions. Figure 2 shows the cycle we envision: the Web of Facts is fed by nanopublications, machines generate syllogisms (reproducing what suggested by the LBD methodology), and results are presented to humans in natural language. The results themself, in RDF, can contribute to feeding the Web of Facts with newly generated nanopublications. Visualization can then be used to show networks of papers that state similar things (or, by using OWL properties, opposite or conflicting nanopublications).

With a global collection accessible, it could retrieve more nanopublications that represent identical semantic predication (i.e., the triple subject, predicate, and object are equal). This fact can increase accuracy, thanks to the redundancy. Tools that summarize, reason, and make discoveries exploit such a collection.

The model is summarized in Figure 3.

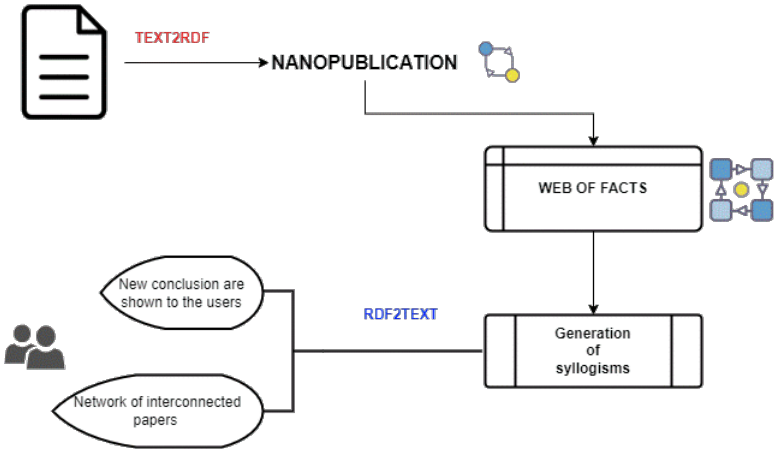

Figure 2: Cycle of nanopublication extraction and visualization.

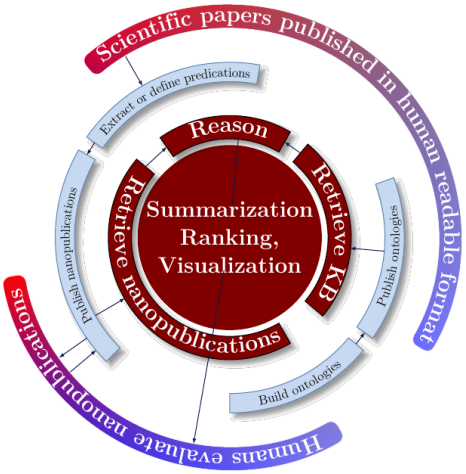

Figure 3: The model we envision: assertions in documents are curated and published on the Web and made available to machines that can reason on them.

4. Feeding the Web of Facts with author-curated predications

The Web of Fact can be feed by the authors themself. However, a straightforward tool to use is necessary. We claim that annotating documents in HTML facililate the extraction of semantic predications and the creation of nanopublications. Elements to include in a nanopublication are (i) a semantic predication representing an assertion expressed in the paper, (ii) a level of truth of the claim, such as speculation, hypothesis, claim, fact, or observation, (iii) provenance, i.e., author and any reference to the scientific publication, and (iv) information about the intellectual property.

We envision a method, that we named *DesX* to annotate natural language statements in papers, that is complemented by the counterpart to represent the RDF output by LBD tools in

natural language. Based on the author's input, DesX inserts annotations inside attributes of the *span* elements in the HTML documents to indicate entities and predicates. An example of an annotated sentence is:

```
<span data-desx-tpl="wd:Q41567 wdt:P50 wd:Q692">
    <span data-desx-entity="wd:Q692" title="Shakespeare">The bard</span>
    wrote '<span data-desx-entity="wd:Q41567">Hamlet</span>'
</span>
```

With such information, DesX can automatically compile the semantic predications needed to create the nanopublication to feed the Web of Facts.

Also, DesX can suggest to the author triples worthy of being extracted. In this case, the identification is organized into three steps: (i) Identifying the template/pattern (ii) Identifying the entities mentioned in the template/pattern (iii) Identifying the role of such entities (subject, object - depending on the template) A template is defined as a list of terms with their role. It can be specific to a particular field and could originate from ontologies. With this mechanism in place, the interface helps the author(s) publish nanopublications alongside their paper. Once a predication is created, the author revises the triple, improves the predication, and adds necessary information. The result is a nanopublication ready to feed the collection of machine-readable assertions — the Web of Facts. An example of a sentence that contains a claim from which a nanopublication is extracted is shown in Figure 4. Provenance can be added automatically to

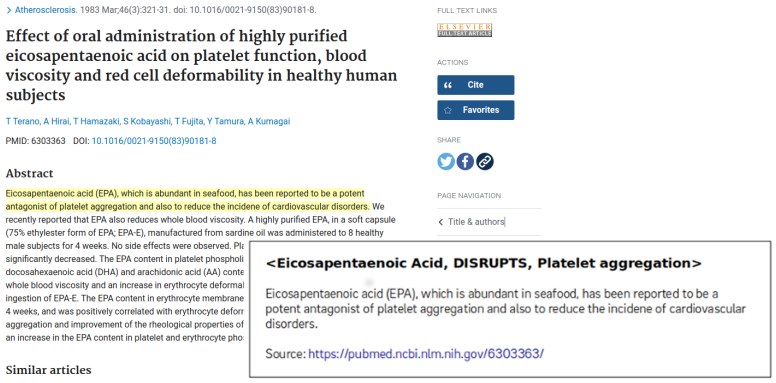

Figure 4: A nanopublication extracted by the author from the HTML published version of the article.

the nanopublication as a triple and the corresponding assertion in NL, such as "doi:xxxx claims that Eicosapentaenoic acid disrupts platelet aggregation"
Apart from stating the correct provenance, the simple sentence(s) summarizes the paper's outcome or work. Queries can be performed for searching nanopublications with similar subjects, objects, or both, or papers with similar (or identical) nanopublications to filter the number of statements involved in reasonings.

This process includes identifying and reconnecting entities already in the network, such as authors, institutions, and other relevant resources. When the new nanopublication is added to the Web of Facts, mechanisms that leverage unique identifiers, such as DOIs, and established

ontologies, can ensure that the newly added RDF nanopublications are correctly connected to the existing entities, preserving the integrity and consistency of the Web of Facts, especially to avoid duplication of entities and loss of new possible discoveries.

5. Presenting discoveries to Humans

Machines make discoveries using the Web of Facts infrastructure and must present them to humans. A linearization of RDF in the quasi-plain text becomes necessary to represent discoveries and make them readable to a wide audience. With DesX we propose an alternative serialization of RDF meant for being easier to read for humans. It is a template and rule-based model we are developing for Text-to-RDF tasks (e.g. NL realization of nanopublications), which leverages RDF properties to store and describe templates. A sub-property of rdfs:label stores the generative template so that specific property can be verbalized at any time. They can be easily implemented directly in any knowledge base and are highly customizable. The choice was also made to keep consistency between different graphs without risking any divergent interpretation of the graph by using e.g. LLMs. A DesX template represents a pattern of one or two variables and a set of natural language tokens. It is similar to a SPARQL basic graph pattern, mutatis mutandis. DesX templates consist of: (i) One or two variables (object, subject); the variables are denoted by the $ prefix ($object, $subject). The variable resolution is a substitution function with the matched triple's labels on subject and object entities (or data types). The variables work as "placeholders" for the entities' labels. (ii) A set of tokens which represent the property. Any property can be verbalized using a default template: (i) Object properties: $subject is in relationship $propertyLabel with $object (ii) Data properties: $subject's relationship $propertyLabel has value "xsd:DataType $value" E.g.: Template: dcterms:title desx:template "$subject has the title $object" Realization: "doi:10.1234/jneuro.2022.001 has title "Gene therapy improves motor and cognitive function in a mouse model of Huntington's disease" DesX preferred output is in HTML: by using elements with attributes e.g. "data-desx-tpl" to store the originally extracted triple(s), "data-desx-src" to store the source of the triple(s), so reverse conversion and extraction can be easily performed, as well as editing.

5.1. From nanopublications to text

5.1.1. Nanopublications

We show an example starting from a mock nanopublication about Huntington's disease[1].

```
med:np001 {
  <> a np:Nanopublication ;
     np:hasAssertion med:assertion001 ;
     np:hasProvenance med:provenance001 ;
     np:hasPublicationInfo med:pubinfo001 .
}
med:assertion001 {
   med:therapy_genic_med001 a med:Treatment ;
   med:treats med:Huntington_disease ;
```

[1]We assume to use standard prefixes for the triples as shown

```
   med:hasMethod med:gene_therapy ;
   med:hasEvidence med:study001 .
}
med:provenance001 {
    med:study001 a prov:Entity ;
   dcterms:title "Gene therapy improves motor and cognitive function in a mouse model of Huntington's disease
    dcterms:creator "Doe J., Smith A."@en ;
    dcterms:date "2022-01-15"^^xsd:date ;
    foaf:homepage <http://example.com/study001> ;
    prov:wasDerivedFrom <http://example.com/mouse_model_001> .
}
med:pubinfo001 {
  dcterms:title "Gene therapy for Huntington's disease" ;
  dcterms:publisher "Journal of Neuroscience" ;
  dcterms:identifier "doi:10.1234/jneuro.2022.001" .
}
```

5.1.2. Templates

The properties would have templates similar to ones in table 1:

Property	Template	Example
med:Treatment	`med:Treatment desx:template "$subject can be used to treat $object"`	Gene therapy can be used to treat Huntington's disease
med:treats	`med:treats desx:template "$subject is a treatment for $object"`	med:treats desx:template "Gene therapy is a treatment for Huntington's disease"
med:hasMethod	`med:hasMethod desx:template "$subject was administered using the method $object"`	"Gene therapy was administered using the method gene therapy
med:hasEvidence	`:med:hasEvidence desx:template "There is evidence supporting the claim that $subject $object"`	There is evidence supporting the claim that gene therapy improves motor and cognitive function in a mouse model of Huntington's disease
dcterms:title	`dcterms:title desx:template "$subject has the title $object"`	The publication has the title Gene therapy for Huntington's disease
dcterms:publisher	`dcterms:publisher desx:template "The publication was published by $object"`	The publication was published by Journal of Neuroscience
dcterms:identifier	`dcterms:identifier desx:template "The publication has the identifier $object"`	The publication has the identifier doi:10.1234/jneuro.2022.001

Table 1
Desx template examples

The resulting text, without much post-processing, would result as: "Gene therapy, a type of treatment, which is based on the method of gene therapy, can treat Huntington's disease, as supported by the evidence presented in `med:study001`, "Gene therapy improves motor and cognitive function in a mouse model of Huntington's disease". It was published by Journal of Neuroscience", it was created on 2022-01-15 by Doe J. and Smith A. It has the identifier

doi:10.1234/jneuro.2022.001, and can be found at http://example.com/study001. The study was derived from the resource http://example.com/mouse_model_001".

A visualization method based on templates, while it has its drawbacks (the template themselves, difficulty in making the text seem natural), offers several advantages. First, it is accountable, meaning the resulting text can be traced back to the RDF data and the template used to generate it. This is important for ensuring the accuracy and reliability of the information presented in the text. Second, the method is transparent in the way it generates the text. By using predefined templates, the process of creating the text is clear and can be easily understood by others. This is particularly important in cases where the text is used for decision-making purposes, as it enables stakeholders to understand how the text was generated and assess its validity. Third, the method relies on provenance information, which is stored in the RDF data, to ensure that the information presented in the text is based on reliable sources. This means the text can be trusted to reflect the underlying data accurately. Fourth, the method is consistent across platforms. The templates are defined as platform-independent, enabling the same template to generate text across different systems and applications. This consistency ensures that the text remains the same, regardless of the platform used to generate it, while being adaptable to different languages.

6. Conclusions

More than 50 million scientific publications contain hidden, implicit hypotheses that could accelerate science if retrieved. Although making discoveries automatically from the scientific literature is arduous, succeeding in this would have a relevant impact on humanity to justify the highest effort. The Literature-based discoveries approach generates new hypotheses by analyzing and combining different scientific publications. Computers can execute this systematic, pragmatic, and well-defined methodology reasoning on semantic predications representing article assertions. If such statements are published with attributes such as provenance, they are valid (nano) scientific publications that can be combined to generate valuable hypotheses. From this, the idea: create a Worldwide collection of nanopublications that machines can consume to suggest new hypotheses. Nanopublications are extracted from hypertext and published next to them: they are the machine-readable collection of the relevant assertions in the paper. Our view is to generate nanopublications to feed such Web of Fact, curating them, having software that makes conclusions exploiting the collection, and generating new hypertext for humans. Tools and methodologies exist to show that with syllogisms from semantic predications discoveries are rendered, and we have presented examples.

In this paper, we have given an overview of the model we envision and described Desx, a relevant part of the model that facilitates the feeding of the Web of Facts with manually generated nanopublications and the representation of the discoveries in natural language.

In conclusion, by reducing the sparseness of literature, and discovering new hypotheses within the scientific literature, offers significant potential for improving Science.

References

[1] A. Jinha, Article 50 million: An estimate of the number of scholarly articles in existence, Learned Publishing 23 (2010) 258–263. doi:10.1087/20100308.

[2] M. Ware, M. Mabe, The STM Report: An overview of scientific and scholarly journal publishing, Copyright, Fair Use, Scholarly Communication, etc. (2015). URL: https://digitalcommons.unl.edu/scholcom/9.

[3] D. R. Swanson, Undiscovered Public Knowledge, The Library Quarterly 56 (1986) 103–118. URL: https://www.journals.uchicago.edu/doi/10.1086/601720. doi:10.1086/601720.

[4] K. Popper, The Logic of Scientific Discovery, Hutchinson, London, 1959.

[5] D. R. Swanson, Fish Oil, Raynaud's Syndrome, and Undiscovered Public Knowledge, Perspectives in Biology and Medicine 30 (1986) 7–18. URL: https://muse-jhu-edu.ezproxy.library.dal.ca/article/403510/pdf.

[6] J. Preiss, M. Stevenson, M. H. McClures, Towards Semantic Literature Based Discovery (2012) 2.

[7] D. Cameron, O. Bodenreider, H. Yalamanchili, T. Danh, S. Vallabhaneni, K. Thirunarayan, A. P. Sheth, T. C. Rindflesch, A graph-based recovery and decomposition of Swanson's hypothesis using semantic predications, Journal of Biomedical Informatics 46 (2013) 238–251. URL: https://www.sciencedirect.com/science/article/pii/S1532046412001517. doi:10.1016/j.jbi.2012.09.004.

[8] M. Thilakaratne, K. Falkner, T. Atapattu, A Systematic Review on Literature-based Discovery: General Overview, Methodology, & Statistical Analysis, ACM Computing Surveys 52 (2020) 1–34. URL: https://dl.acm.org/doi/10.1145/3365756. doi:10.1145/3365756.

[9] D. Cameron, R. Kavuluru, T. C. Rindflesch, A. P. Sheth, K. Thirunarayan, O. Bodenreider, Context-driven automatic subgraph creation for literature-based discovery, Journal of Biomedical Informatics 54 (2015) 141–157. URL: https://www.sciencedirect.com/science/article/pii/S1532046415000167. doi:10.1016/j.jbi.2015.01.014.

[10] M. Thilakaratne, K. Falkner, T. Atapattu, A systematic review on literature-based discovery workflow, PeerJ Computer Science 5 (2019) e235. URL: https://peerj.com/articles/cs-235. doi:10.7717/peerj-cs.235, publisher: PeerJ Inc.

[11] T. Kuhn, P. E. Barbano, M. L. Nagy, M. Krauthammer, Broadening the Scope of Nanopublications, in: P. Cimiano, O. Corcho, V. Presutti, L. Hollink, S. Rudolph (Eds.), The Semantic Web: Semantics and Big Data, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2013, pp. 487–501. doi:10.1007/978-3-642-38288-8_33.

[12] W3C, Semantic Web - W3C, 2015. URL: https://www.w3.org/standards/semanticweb/.

[13] P. Groth, A. Gibson, J. Velterop, The anatomy of a nanopublication, Information Services & Use 30 (2010) 51–56. URL: https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/ISU-2010-0613. doi:10.3233/ISU-2010-0613.

[14] H. Kilicoglu, G. Rosemblat, M. Fiszman, D. Shin, Broad-coverage biomedical relation extraction with SemRep, BMC Bioinformatics 21 (2020) 188. URL: https://doi.org/10.1186/s12859-020-3517-7. doi:10.1186/s12859-020-3517-7.

[15] X. Li, S. Peng, J. Du, Towards medical knowmetrics: representing and computing medical knowledge using semantic predications as the knowledge unit and the uncertainty as the knowledge context, Scientometrics (2021). URL: https://doi.org/10.1007/

s11192-021-03880-8. doi:10.1007/s11192-021-03880-8.

[16] B. Elsworth, K. Dawe, E. E. Vincent, R. Langdon, B. M. Lynch, R. M. Martin, C. Relton, J. P. T. Higgins, T. R. Gaunt, MELODI: Mining Enriched Literature Objects to Derive Intermediates, International Journal of Epidemiology 47 (2018) 369–379. URL: https://academic.oup.com/ije/article/47/2/369/4803214. doi:10.1093/ije/dyx251.

[17] T. C. Rindflesch, M. Fiszman, The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text, Journal of Biomedical Informatics 36 (2003) 462–477. URL: https://linkinghub.elsevier.com/retrieve/pii/S1532046403001175. doi:10.1016/j.jbi.2003.11.003.

[18] B. Elsworth, T. R. Gaunt, MELODI Presto: a fast and agile tool to explore semantic triples derived from biomedical literature, Bioinformatics (2020). URL: https://doi.org/10.1093/bioinformatics/btaa726. doi:10.1093/bioinformatics/btaa726.

[19] S. Kanza, J. G. Frey, A new wave of innovation in Semantic web tools for drug discovery, Expert Opinion on Drug Discovery 14 (2019) 433–444. URL: https://doi.org/10.1080/17460441.2019.1586880. doi:10.1080/17460441.2019.1586880, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/17460441.2019.1586880.

[20] E. K. Neumann, E. Miller, J. Wilbanks, What the semantic web could do for the life sciences, Drug Discovery Today: BIOSILICO 2 (2004) 228–236. URL: https://linkinghub.elsevier.com/retrieve/pii/S1741836404024205. doi:10.1016/S1741-8364(04)02420-5.

[21] A. Callahan, M. Dumontier, N. H. Shah, HyQue: evaluating hypotheses using Semantic Web technologies, Journal of Biomedical Semantics 2 (2011) S3. URL: https://doi.org/10.1186/2041-1480-2-S2-S3. doi:10.1186/2041-1480-2-S2-S3.

[22] A. V. Callahan, Semi-automated Hypothesis Evaluation Using Semantic Technologies, Text, Carleton University, 2014. URL: https://curve.carleton.ca/20d96321-a493-4a92-8e46-151b80fef6e6, last Modified: 2015-07-03T16:20-04:00.

[23] F. Belleau, M.-A. Nolin, N. Tourigny, P. Rigault, J. Morissette, Bio2RDF: Towards a mashup to build bioinformatics knowledge systems, Journal of Biomedical Informatics 41 (2008) 706–716. URL: http://www.sciencedirect.com/science/article/pii/S1532046408000415. doi:10.1016/j.jbi.2008.03.004.

[24] I. Harrow, W. Filsell, P. Woollard, I. Dix, M. Braxenthaler, R. Gedye, D. Hoole, R. Kidd, J. Wilson, D. Rebholz-Schuhmann, Towards Virtual Knowledge Broker services for semantic integration of life science literature and data sources, Drug Discovery Today 18 (2013) 428–434. URL: https://www.sciencedirect.com/science/article/pii/S1359644612004011. doi:10.1016/j.drudis.2012.11.012.

[25] I. Augenstein, S. Padó, S. Rudolph, LODifier: Generating Linked Data from Unstructured Text, in: E. Simperl, P. Cimiano, A. Polleres, O. Corcho, V. Presutti (Eds.), The Semantic Web: Research and Applications, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2012, pp. 210–224. doi:10.1007/978-3-642-30284-8_21.

[26] A. Meloni, D. R. Recupero, A. Gangemi, Amr2fred, a tool for translating abstract meaning representation to motif-based linguistic knowledge graphs, in: Extended Semantic Web Conference, 2017.

[27] Q. Guo, Z. Jin, X. Qiu, W. Zhang, D. Wipf, Z. Zhang, CycleGT: Unsupervised Graph-to-Text and Text-to-Graph Generation via Cycle Training, in: Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+),

Association for Computational Linguistics, Dublin, Ireland (Virtual), 2020, pp. 77–88. URL: https://aclanthology.org/2020.webnlg-1.8.

[28] Q. Guo, Z. Jin, N. Dai, X. Qiu, X. Xue, D. Wipf, Z. Zhang, A Plan-and-Pretrain Approach for Knowledge Graph-to-Text Generation, in: Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+), Association for Computational Linguistics, Dublin, Ireland (Virtual), 2020, pp. 100–106. URL: https://aclanthology.org/2020.webnlg-1.10.

[29] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language Models are Few-Shot Learners, 2020. URL: http://arxiv.org/abs/2005.14165. doi:10.48550/arXiv.2005.14165, arXiv:2005.14165 [cs].

[30] T. Castro Ferreira, C. Gardent, N. Ilinykh, C. van der Lee, S. Mille, D. Moussallem, A. Shimorina, The 2020 Bilingual, Bi-Directional WebNLG+ Shared Task: Overview and Evaluation Results (WebNLG+ 2020), in: Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+), Association for Computational Linguistics, Dublin, Ireland (Virtual), 2020, pp. 55–76.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention Is All You Need, 2017. URL: http://arxiv.org/abs/1706.03762. doi:10.48550/arXiv.1706.03762, arXiv:1706.03762 [cs].

[32] D. Moussallem, T. C. Ferreira, M. Zampieri, M. C. Cavalcanti, G. Xexéo, M. Neves, A.-C. N. Ngomo, RDF2PT: Generating Brazilian Portuguese Texts from RDF Data, 2018. URL: http://arxiv.org/abs/1802.08150. doi:10.48550/arXiv.1802.08150, arXiv:1802.08150 [cs].

[33] D. Moussallem, Knowledge Graphs for Multilingual Language Translation and Generation (2020). URL: http://arxiv.org/abs/2009.07715. doi:10.17619/UNIPB/1-980, arXiv:2009.07715 [cs].

[34] A.-C. N. Ngomo, D. Moussallem, L. Bühmann, A Holistic Natural Language Generation Framework for the Semantic Web, Technical Report arXiv:1911.01248, arXiv, 2019. URL: http://arxiv.org/abs/1911.01248. doi:10.48550/arXiv.1911.01248, arXiv:1911.01248 [cs] type: article.