The DISK Hypothesis Ontology: Capturing Hypothesis Evolution for Automated Discovery

Daniel Garijo, Yolanda Gil and Varun Ratnakar
Information Sciences Institute, University of Southern California, Marina del Rey, CA, U.S.A
{dgarijo, gil, varunr}@isi.edu

ABSTRACT

Automated discovery systems can formulate and revise hypotheses by gathering and analyzing data. In order to generate new hypotheses and provide explanations of their new findings, these systems need a language to represent hypotheses, their revisions, and their provenance. This paper describes the DISK hypothesis ontology which fulfills these requirements. The paper then presents a survey of existing models for representing hypotheses along with their features and tradeoffs. We compare these hypothesis models in the context of automated discovery and hypothesis evolution.

CCS CONCEPTS

• **Information systems** → **Artificial intelligence**; *Knowledge representation and reasoning*

KEYWORDS

Hypothesis representation, hypothesis evolution, nanopublications, micropublications, automated discovery, ontologies.

1 INTRODUCTION

Formal representations of scientific hypotheses would be useful in many contexts. For instance, in order to keep up with the latest updates on a research area, scientists need to quickly understand the contributions of an article and how it was derived from others. However, the vast amount of new scientific publications makes this task increasingly complex. If scientists represented hypotheses formally in publications, related literature could be easily searched for hypotheses of interest. Alternatively, machine reading systems could also extract hypotheses from text in articles, and generate these formal representations.

Formal representations of hypotheses may also be used to improve reproducibility. Community initiatives on reproducibility promote registering hypotheses and methods before conducting the research [Munafo et al 2017]. Hypotheses are stated in textual form, which can express arbitrarily complex statements about hypotheses. However, text can be imprecise and ambiguous. Creating machine readable representations of research hypotheses would facilitate the organization and management of the literature. To date there is not a standard way of capturing the contents and context of a hypothesis to understand its evolution.

Another important use of formal hypothesis representations is to enable automated discovery systems to do hypothesis testing and revision. Autonomous discovery systems generate hypotheses autonomously based on analysis of relevant data [Pankratius et al 2016; King 2017; Gil et al 2017].

In this paper, we focus on hypothesis representations to capture hypothesis evolution in automated discovery systems. We discuss the requirements that we have found throughout work on the DISK discovery system [Gil et al 2017]. We propose an ontology for hypothesis representation, and compare it to existing models for representing hypotheses.

The rest of the paper is organized as follows. Section 2 describes the DISK automated discovery system, and introduces its hypothesis ontology. Section 3 introduces an evaluation framework for existing models and overviews them. Section 4 discusses the different alternatives for hypothesis representation, and Section 5 concludes the paper.

2 REPRESENTING HYPOTHESES IN THE DISK AUTOMATED DISCOVERY SYSTEM

Our goal is to allow automated discovery systems to test hypotheses provided by users, and revise them based on the results of running computational experiments autonomously.

In prior work, we introduced an approach that captures scientists' strategies for pursuing hypotheses as *lines of inquiry* that specify the data to be retrieved, the experimental workflows to run, and how to combine the results to generate a revised confidence level and in some cases a revised hypothesis [Gil et al 2016]. This approach was implemented in the DISK framework (Automated DIscovery of Scientific Knowledge) and demonstrated for cancer multi-omics [Gil et al 2017]. DISK is given a hypothesis statement, such as whether a protein is associated with a type of cancer, and returns either a confidence level on that hypothesis or a revised hypothesis that refers to a mutation of the protein or a more specific type of cancer. As new data becomes available, DISK re-runs the analysis and continuously revises the original hypothesis. DISK tracks the provenance of revised hypotheses in terms of the original hypotheses and the data analyses that were carried out.

K-CAP2017 Workshops and Tutorials Proceedings,

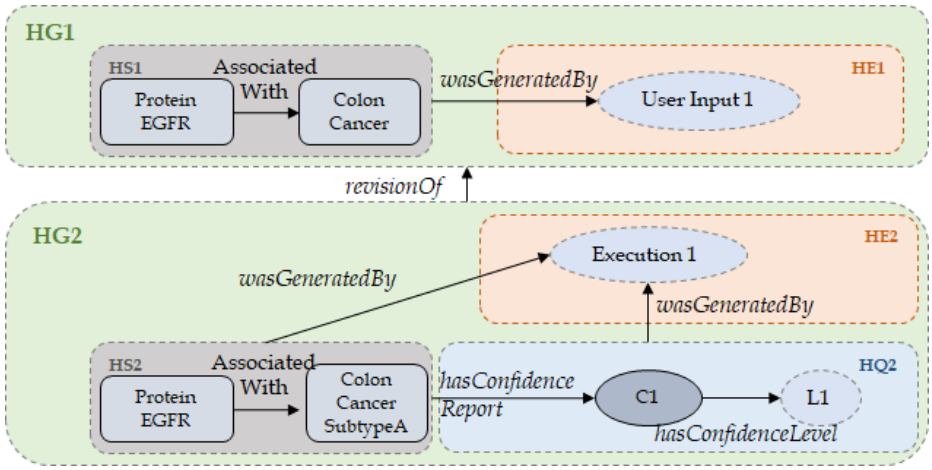

Figure 1. Representing hypotheses in the DISK automated discovery system using the DISK hypothesis ontology. The initial hypothesis statement HS1 is provided by the user. It is then tested through data analysis, which provides evidence HE2 for the hypothesis, a new hypothesis statement HS1, and a qualification HQ2 with a confidence level L1. The revised hypothesis HG2 is a revision of HG1, indicated by a link.

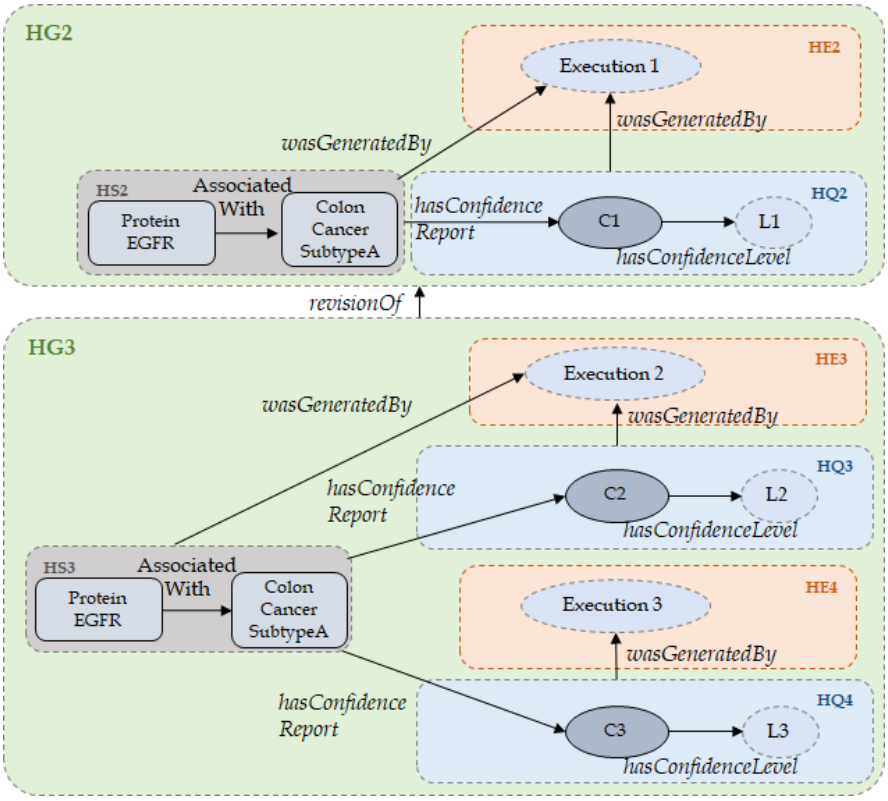

Figure 2. Representing hypothesis evolution in the DISK automated discovery system using the DISK hypothesis ontology. In this example, additional data of two different types becomes available, causing the system to trigger two separate analyses whose results are hard to combine. A revised hypothesis statement HS3 is added with a new confidence level L2 (included as part of HQ3) backed by one of the analyses as evidence HE3. The other analysis HE4 qualifies HS3 with HQ4.

DISK uses a representation of hypotheses that is needed to track their evolution. In DISK, a hypothesis consists of:

1. A **hypothesis statement**, which is a set of structured assertions about entities in the domain. For example, that the protein EGFR is associated with colon cancer.

2. A **hypothesis qualifier**, which represents the veracity of the hypothesis based on the data and the analyses done so far. A typical qualifier is a numeric confidence level. For example, for the hypothesis statement above we could have a confidence level given by a p-value of 0.07.

3. **Hypothesis evidence**, which is a record of the analyses that were carried out to test a hypothesis statement. For example, the evidence of a given hypothesis may include an analysis of mass spectrometry data for 25 patients with colon cancer and 25 healthy controls followed by clustering, cluster metrics and binary hypothesis testing.

4. A **hypothesis history**, which points to prior hypotheses that were revised to generate the current one. In our example, a hypothesis such as the association of protein EGFR with colon cancer SubType A would link back to the original hypothesis statement that protein EGFR is associated with colon cancer.

DISK represents hypothesis statements as a graph, where the nodes are the entities in the hypotheses and the links are their relationships. In our work, a hypothesis statement is represented in RDF as a simple triple, and the triple is linked to its qualifier, evidence, and history. All those assertions are also made in RDF. The hypothesis evidence and hypothesis history both represent different aspects of provenance for the hypothesis. This is captured using the PROV provenance standard [Lebo et al 2013].

Figure 1 illustrates this representation using the running example with protein EGFR. The original hypothesis HG1 had its own statement HS1 and evidence HE1. The revised hypothesis HG2 includes its statement HS2, its confidence level L1 (part of the qualifier HQ2), its evidence HE2, and a link to the original hypothesis HG1. A feature of this representation is the ability to model different confidence levels associated to a hypothesis statement. This often happens when evidence is obtained from analyzing different types of data and it is unclear how to combine the resulting confidence levels. Figure 2 shows an example. HS3 is qualified with two confidence reports (C2 and C3), which have different supporting evidence (HE3 and HE4) each resulting from a different data source.

The DISK hypothesis ontology is available in OWL and documented in [Garijo et al 2017]. A major focus of the DISK hypothesis ontology is capturing hypothesis evolution. The rest of this paper focuses on comparing this ontology to other representations of scientific hypotheses in the literature.

3 A SURVEY OF HYPOTHESIS REPRESENTATIONS

In this section we present a survey of existing models of scientific hypotheses and assess their features to support automated discovery.

3.1 Comparing hypothesis models

In our analysis, we consider the following key aspects, based on the representation presented in Section 2:

1. **Statement:** Does the model have a representation for statements in a hypothesis?
2. **Qualifier:** Does the model have a means to qualify a hypothesis with a confidence level?
3. **Evidence:** Does the model describe the supporting evidence for a hypothesis?
4. **History:** Does the model represent the relationship between hypothesis revisions?

In addition, the following aspects are desirable for flexibility and extensibility:

5. **Classification:** Does the vocabulary support a taxonomy of hypothesis statements?
6. **Standards:** Is the model defined using standards or does it use proprietary or idiosyncratic formats?

3.2 Models for representing hypotheses

This section introduces different approaches to represent hypothesis at different levels of granularity. We group them based according to the level of detail at which they describe hypotheses: coarse-grained and fine-grained representations.

3.2.1 Coarse-grained hypothesis models

We group under this section those vocabularies that include main concepts to identify hypotheses, but do not include the means to qualify them or describe them at a statement level. For example, popular vocabularies like the **Semantic Web for Earth and Environmental Terminology Ontology[1](SWEET)** [Raskin and Pan 2005] contain modules for defining hypotheses as "Experimental Activities". Likewise, the **Ontology for Biomedical Investigations (OBI)**[2] [Brandowski et al 2016] and the **Ontology for Clinical Research (OCRe)**[3] [Sim et al 2014] have concepts to refer to a hypothesis in the context of a biological experiment.

Other vocabularies include terms to further describe hypotheses. The **EXPO Ontology** aims to define a model for representing scientific experiments, "including generic knowledge about scientific experimental design, methodology and results representation" [Soldatova and King, 2006]. The EXPO Ontology extends common upper level ontologies in order to bridge the gap between domain specific experiment formalization and upper level ontologies. EXPO aims at describing scientific papers, and has a specific part designed for the description of hypotheses. The

[1] http://sweet.jpl.nasa.gov/2.3/reprSciModel.owl
[2] http://purl.obolibrary.org/obo/OBI_0001908
[3] http://purl.org/net/OCRe/OCRe.owl#OCRE400032

focus of EXPO is on how the hypothesis is defined on a research paper (the "part of" relationship between the scientific experiment and the hypothesis), rather than identifying the statements contained by the hypothesis itself. However, different classes of hypothesis are identified in the ontology (i.e., null hypothesis, research hypothesis and scientific hypothesis).

Finally, the **Linked Science Vocabulary** [4] proposes a lightweight model to express support to hypothesis by some research. A hypothesis is represented to make predictions about facts, but it is not described at a statement level.

3.2.2 Fine grained hypothesis models

We group in this section those approaches that provide the means to represent in detail the statements belonging to a hypothesis, along with their metadata.

LABORS [Soldatova and Rzhetsky 2011] is designed to support investigations run by an automated system for the area of Systems Biology and Functional Genomics. LABORS uses EXPO as an upper level ontology, and splits the representation of hypotheses into textual and logical representations, using concepts from OBI and other upper level ontologies. It also allows aggregating hypotheses with multiple statements in *hypothesis sets*, using a Datalog representation for each hypothesis statement.

The **nanopublication model** [5] [Groth et al 2010] aims to represent "the smallest unit of publishable information", i.e., every assertion that is part of a hypothesis graph. Nanopublications are composed of three main graphs: An *assertion graph* containing the assertion or multiple assertions which are part of the nanopublication, a *provenance graph* with the statements that describe the provenance of the assertion graph (e.g., the assertion graph came from a publication, a scientific experiment, etc.); and lastly a *publication info* graph which contains the metadata about the nanopublication itself. (e.g., who created the nanopublication, date when the nanopublication was created, etc.). Each of the graphs is represented using a named graph,[6] so as to be able to describe it properly with metadata from any of the other graphs. An example can be seen in the snippet below, where a hypothesis H1 as in Figure 1 is represented with its provenance (*sub:provenance)*, assertion (*sub:hypothesisAssertion*) and publication (*sub:pubInfo*) graphs.

```
@prefix sub: <http://example.org/hypothesis#> .
@prefix np:  <http://www.nanopub.org/nschema#> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix xsd:  <http://www.w3.org/2001/XMLSchema#> .
@prefix ex:  <http://example.org#>
sub:defaultGraph {
 sub:n1 np:hasAssertion sub: hypothesisAssertion;
        np:hasProvenance sub:provenance ;
        np:hasPublicationInfo sub:pubInfo ;
        a np:Nanopublication, ex:Hypothesis .
}
sub:hypothesisAssertion {##statements contained in the
hypothesis graph
 ex:EGFR ex:associatedWith ex:ColonCancer .}
sub:provenance { ##provenance of the assertion graph
 sub: hypothesisAssertion prov:generatedAtTime "2012-02-
03T14:38:00Z"^^xsd:dateTime ;
            ex:hasConfidenceReport ex:conf1.
            prov:wasAttributedTo ex:experimentScientist .
ex:conf1 a ex:ConfidenceReport;
            ex:hasConfidenceLevel "0.6".
            prov:wasGeneratedBy ex:execution1.
}
sub:pubInfo {##publication information of the user who
performed the hypothesis
: prov:generatedAtTime "2016-03-26T12:45:00Z"^^xsd:dateTime;
            prov:wasAttributedTo ex:user1 .
}
```

The **ovopublication model** proposes a simple approach designed to capture the provenance of assertions [Callahan and Dumontier 2013]. When contrasted with nanopublications, "the ovopub is simpler as it consists of only a single named graph with key provenance information directly contained in and associated with the ovopub graph" [Callahan and Dumontier 2013]. Ovopublications mix the notion of named graphs with reification to refer to the different components and relationships of the own ovopublication. The Ovopub model is integrated as part of the Semanticscience Integrated Ontology (SIO)[7], which also provides the means to describe hypothesis as literals

The **Semantic Web Applications in Neuromedicine (SWAN) ontology**[8] **[**Ciccarese et al 2008**]** aims to represent the scientific discourse of bio-medicine papers in general and neuro-medicine papers in particular. The model is composed of several modules for representing discourse elements and their relationships, different types of agents, the roles, provenance and versioning of a given statement and bibliographic references. SWAN was designed to describe statements in papers (along with the evidence supporting them). If we consider a hypothesis as a text statement, the following example illustrates the SWAN model:

```
@prefix swande: <http://purl.org/swan/1.2/discourse-
elements/> .
@prefix swanco:<http://purl.org/swan/1.2/swan-commons/> .
@prefix swanqs: <http://purl.org/swan/1.2/qualifiers/> .
@prefix swandr: <http://purl.org/swan/1.2/discourse-
relationships/> .
@prefix swanpav: <http://purl.org/swan/1.2/pav/> .
@prefix swanci: <http://purl.org/swan/1.2/citations/> .

ex:hypothesis a swande:ResearchStatement ;
    swande:title "EGFR is associated with colon cancer
subtype A"@en;
    swanco:researchStatementQualifiedAs
<http://swan.mindinformatics.org/ontologies/1.2/rsqualifiers/
hypothesis>;
    swanci:derivedFrom ex:execution1;
    ex:hasConfidenceReport ex:c1;
    swanpav:authoredBy ex:experimentScientist;
    swanpav:createdOn 2012-02-03T14:38:00Z"^^xsd:dateTime .
```

In the example, a hypothesis is extracted from a research article. The hypothesis is represented as a statement, which can be further described with SWAN. The provenance of the hypothesis is represented as well by representing the agents who created the hypothesis statement.

[4] http://linkedscience.org/lsc/ns/

[5] http://www.nanopub.org/nschema#

[6] https://www.w3.org/TR/rdf11-concepts/

[7] http://semanticscience.org/ontology/sio.owl

[8] https://www.w3.org/TR/hcls-swan/

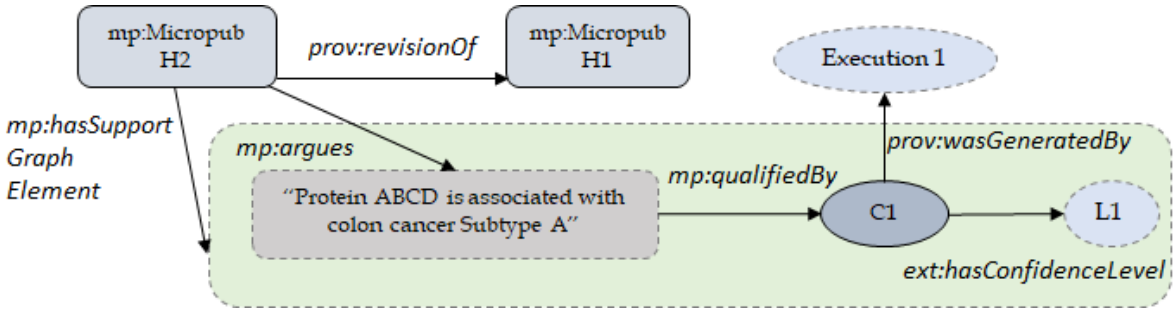

Figure 3: The example from Figure 1 adapted to the micropublication model, following [Clark et al 2014]. The namespaces indicate the ontology used: mp for micropublications, prov for the PROV ontology, and ext for the extension that would need to be added.

Finally, **micropublications**[9] [Clark et al 2014] are derived from the SWAN model and can be considered a refinement of the nanopublication model. Micropublications propose a semantic model of scientific argumentation and evidence that supports natural language statements, data and materials specifications, discussion, etc. Figure 3 shows an illustrative example, where a micropublication uses a mechanism similar to an assertion graph to represent the claim of a protein being associated with a subtype of colon cancer, along with its supporting evidence. The micropublication model uses the Web Annotation Ontology[10] to associate a micropublication and its contents with text from articles.

4 DISCUSSION

Table 1 summarizes the different candidate models for hypothesis representation in automated discovery systems, according to the features described in Section 3.1. Most models lack support for qualifying a given hypothesis with confidence levels. In order to overcome this issue, we may follow an approach similar to Figure 1: extend the target model with a class (*confidence Report*) and two properties (*hasConfidenceReport* and *hasConfidenceLevel*) linking them together. A reason why the confidence level may not be directly linked to a hypothesis is that the same hypothesis may be evaluated at different points in time, resulting in multiple confidence levels with different provenance information each included in a separate confidence report.

The upper half of Table 1 corresponds to the models for coarse grained hypothesis representation. These models include a main concept to refer to a hypothesis, but lack the means to describe hypothesis statements. Therefore, they do not meet the majority of requirements that DISK requires for representing hypothesis statements, qualifiers, history and evidence. However, the LinkedScience, OBI and EXPO vocabularies define different types of hypotheses, and may be potential candidates for reuse if we need to define a hypothesis taxonomy.

The lower half of Table 1 corresponds to fine-grained models to describe hypotheses, either defining classes and properties to qualify hypothesis statements with provenance metadata or relating its different parts together. Among these, the nanopublication and micropublication models are the most flexible approaches, compliant with most of the requirements of the DISK model (in the last row). LABORS uses a datalog representation for describing hypothesis statements and is domain specific. The ovopublications model is a simplification of the nanopublication model to include provenance of assertions or collections of assertions. Although it could be used for hypothesis representation, we consider that the model would need to be thoroughly extended. Similarly, the SWAN model is extended in the micropublication approach to represent argumentation of facts in publications. Therefore, the nanopublication and micropublication models provide a richer initial framework.

A major difference between micropublications and nanopublications is the scope of the domain. For instance, micropublications was explicitly designed to model facts and argumentation of text statements. If an automated discovery system aims to represent single assertions of hypotheses and their evolution, then an argumentation framework such as the one proposed in the micropublication model is not necessary. In contrast, if the provenance trace includes all evidence to support a particular claim made in a hypothesis, then micropublications are an appropriate model to use.

Another aspect to consider is the support from the communities that are using these models. The nanopublication model has been discussed for some time, and has available tooling, documentation and examples.[11] The micropublication model has been documented in detail with examples [Clark et al 2014], but it has not yet reached the level of adoption and tooling that nanopublications have.

[9] http://purl.org/mp
[10] https://www.w3.org/ns/oa
[11] http://nanopub.org/

Table 1: Overview of models for hypothesis representation.

Hypothesis Model	Hypothesis statement	Hypothesis qualifier	Hypothesis evidence	Hypothesis history	Hypothesis classification	Use of standards
SWEET [Raskin and Pan 2005]	No	No	No	No	No	Yes (OWL)
OBI [Brandowski et al 2016]	No	No	No	No	Yes	Yes (OWL)
EXPO [Soldatova and King 2006]	No	No	No	No	Yes	Yes (OWL)
OCR [Sim et al 2014]	No	No	No	No	No	Yes (OWL)
Linked Science Vocabulary	No	No	Partly	No	No	Yes (OWL)
LABORS [Soldatova and Rzhetsky 2011]	No	No	Yes	No	Yes	Yes (OWL)
Nanopublications [Groth et al 2010]	Text/ structured	No	Yes	Yes	No	Yes (OWL), named graphs
Ovopublications [Callahan and Dumontier 2013]	Text/ structured	No	No	Yes	No	Yes (OWL), named graphs
SWAN [Ciccarese et al 2008]	Text	No	Yes	Yes	No	Yes (OWL)
Micropublications [Clark et al 2014]	Text	Yes	Yes	No	No	Yes (OWL), named graphs
DISK [Garijo et al 2017]	Structured	Yes	Yes	Yes	No	Yes (OWL), named graphs

Finally, both the nanopublication and micropublication models present an important limitation for representing hypotheses: they have been designed to describe simple facts, i.e., single statements or a single collection of statements as part of their claim. In the nanopublication model this is reflected by having a unique assertion graph per nanopublication, containing one or more statements. If we wanted to describe a hypothesis composed of multiple statements, each with confidence levels assigned independently by different experiments, we would have to extend the nanopublication model. A possibility may be creating a new class (a hypothesis composition concept such as the "hypotheses-set" in LABORS) that aggregates each of its statements as an individual nanopublication. Likewise, each micropublication contains a main claim graph and its support. A mechanism for extending and aggregating micropublications would also be needed to represent hypothesis with multiple statements. Note that the extension would only be necessary in both models if we wanted to keep the provenance for each statement of the hypothesis. Otherwise they can be included in the assertion graph in the case of nanopublications or the claim graph in the case of micropublications.

5 CONCLUSIONS AND FUTURE WORK

In this paper we introduced the DISK hypothesis ontology for representing hypotheses evolution, which was developed for the DISK automated discovery system. We also presented a survey of existing vocabularies to represent hypotheses, and assessed their suitability in the context of automated knowledge discovery. Future work includes extending the DISK ontology to align with these models.

ACKNOWLEDGMENTS

We gratefully acknowledge support from the Defense Advanced Research Projects Agency through the SIMPLEX program with award W911NF-15-1-0555, and from the National Institutes of Health under award 1R01GM117097. We also thank our collaborators in the DISK project, especially Parag Mallick, Ravali Adusumilli, and Hunter Boyce for their useful feedback on this work.

REFERENCES

[Callahan and Dumontier 2013] Alison Callahan and Michel Dumontier. Ovopub: Modular data publication with minimal provenance. arXiv preprint arXiv:1305.6800, 2013.

[Brandrowski et al 2016] Bandrowski A, Brinkman R, Brochhausen M, Brush MH, Bug B, et al. (2016) The Ontology for Biomedical Investigations. PLOS ONE 11(4): e0154556. https://doi.org/10.1371/journal.pone.0154556

[Clark et al 2014] Tim Clark, Paolo N. Ciccarese and Carole A. Goble. Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications. Journal of Biomedical Semantics 2014, 5:28.

[Ciccarese et al 2008] Ciccarese P, Wu E, Kinoshita J, et al. The SWAN Scientific Discourse Ontology. Journal of biomedical informatics. 2008;41(5):739-751. doi:10.1016/j.jbi.2008.04.010.

[Garijo et al 2017] The DISK Hypothesis Ontology. Version 1.0.0. Available from http://disk-project.org/ontology/disk#

[Gil et al 2016] Gil, Y.; Garijo, D.; Ratnakar, V.; Mayani, R.; Adusumilli, R.; and Boyce, H. Automated Hypothesis Testing with Large Scientific Data Repositories. In Proceedings of the Fourth Annual Conference on Advances in Cognitive Systems (ACS), pages 1-6, 2016.

[Gil et al 2017] Gil, Y.; Garijo, D.; Ratnakar, V.; Mayani, R.; Adusumilli, R.; Boyce, H.; Srivastava, A.; and Mallick, P. Towards Continuous Scientific Data Analysis and Hypothesis Evolution. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17), 2017.

[Groth et al 2010] Groth, Paul; Gibson, Andrew; Velterop, Jan. The anatomy of a nanopublication. Information Services and Use, 30, 1-2: 52-56, 2010.

[King 2017] Ross King. The Adam and Eve Robot Scientists for the Automated Discovery of Scientific Knowledge. Bulletin of the American Physical Society, 2017

[Lebo et al 2013] Lebo, T., McGuiness, D., Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., and Zhao, J. (2013). The PROV ontology, W3C recommendation. Technical report, World Wide Web Consortium (W3C), 30th April 2013.

[Munafo et al 2017] Marcus R. Munafò, Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware & John P. A. Ioannidis. A manifesto for reproducible science. Nature Human Behaviour 1, Article number: 0021 (2017). doi:10.1038/s41562-016-0021

[Pankratius et al 2016] V. Pankratius, J. Li, M. Gowanlock, D. Blair, C. Rude, T. Herring, F. Lind, P. Erickson, C. Lonsdale, Computer-Aided Discovery: Towards Scientific Insight Generation with Machine Support. IEEE Intelligent Systems 31(4), pp. 3-10, Jul/Aug 2016.

[Raskin and Pan 2005] Robert G. Raskin and Michael J. Pan. Knowledge representation in the semantic web for Earth and environmental terminology (SWEET). Computers & Geosciences 31(9):1119-1125, November 2005. doi:10.1016/j.cageo.2004.12.004.

[Sim et al 2014] Sim I, Tu SW, Carini S, et al. The Ontology of Clinical Research (OCRe): An Informatics Foundation for the Science of Clinical Research. Journal of biomedical informatics. 2014;52:78-91. doi:10.1016/j.jbi.2013.11.002.

[Soldatova and King 2006]: Soldatova, LN & King, RD. (2006) An Ontology of Scientific Experiments. Journal of the Royal Society Interface, 3(11):795-803, 2006. doi:10.1098/rsif.2006.0134.

[Soldatova and Rzhetsky 2011]: Soldatova, LN and Rzhetsky, A. Representation of research hypotheses. Journal of Biomedical Semantics20112(Suppl 2):S9. 2011. https://doi.org/10.1186/2041-1480-2-S2-S9