Capturing Hypothesis Evolution in Automated Discovery Systems

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 data and analyzing it. 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 these requirements, and presents a survey of existing models for representing hypotheses along with their features and tradeoffs. We assess 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, ontologies.

1 INTRODUCTION

Formal representations of scientific hypotheses would be useful in many contexts. 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, the literature could be easily searched for hypotheses of interest. Alternatively, machine reading systems could also extract hypotheses from the text in articles, and generate these formal representations.

Formal representation of hypotheses could also be used to improve reproducibility. Community initiatives on reproducibility promote the registration of hypotheses and methods ahead of conducting the research [Munafo et al 2017]. At the moment, the process of creating, modifying and evaluating hypotheses is done by hand by scientists. Creating machine readable representations of research hypotheses would facilitate the organization and management of hypotheses. To date there is not an agreed standard way of capturing the contents and context of a hypothesis to understand its evolution.

2017 Workshop on Capturing Scientific Knowledge, held in conjunction with the ACM International Conference on Knowledge Capture (K-CAP), December 4-6, 2017, Austin, TX.

Another important use of formal hypothesis representations is to enable automated discovery systems to do hypothesis testing and revision, including our own work [Pankratius et al 2016; King 2017; Gil et al 2017]. These systems analyze hypothesis statements, generate or find relevant data to be analyzed, and generate new hypotheses based on the analyses.

In this paper, we focus on the development of hypothesis representations for automated discovery systems. We discuss requirements for hypothesis representations, and present an overview of existing models for representing hypotheses and their advantages and tradeoffs for describing hypotheses at different levels of granularity.

The rest of the paper is organized as follows. Section 2 motivates the different aspects to take into account when representing hypotheses. 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 AUTOMATED DISCOVERY SYSTEMS

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 workflows to run, and how to combine the results to generate a revised confidence value 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 in [Gil et al 2017]. DISK is given a hypothesis statement, such as whether a protein is associated with a type of cancer, and returns a confidence value on that hypothesis or a revised hypothesis that refers to a more specific type of cancer. As new data becomes available, DISK re-runs the analysis and continuously revises the original hypothesis.

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 assertions about entities in the domain. For example, they express assertions such as "Protein EGFR is associated with colon cancer."

2. A **hypothesis qualifier**, which qualifies the veracity of the hypothesis based on the data and the analyses done so far. A typical qualifier is a numeric confidence value. For example, for

the hypothesis statement above we could have a confidence value of 0.7.

3. A **hypothesis provenance**, which is a record of the analyses that were carried out to test the hypothesis statement. For example, the provenance 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 prior hypothesis could have a statement such as "Protein EGFR is associated with 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, provenance, and history. All those assertions are also made in RDF. The DISK hypothesis ontology is available in OWL.[1]

Figure 1 shows an example of the previous elements by illustrating hypothesis H2 and including its confidence value, its provenance, and the pointer to H1, its original version. Note that the hypothesis revisions may indicate a new confidence value, a new analysis ("Execution 2" in the figure), and a new hypothesis statement (a subtype of colon cancer in this case).

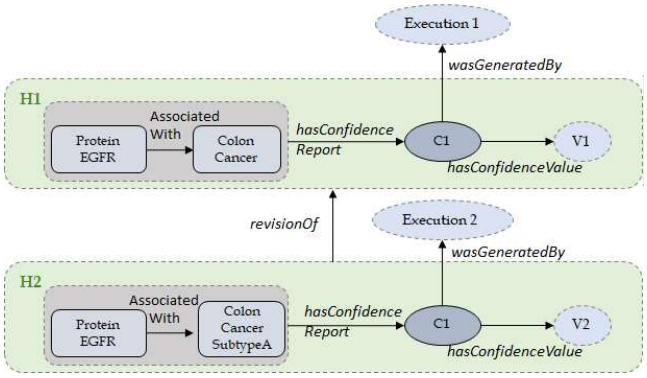

Figure 1: Hypothesis H1 is tested through an execution, resulting in a revised hypothesis H2 that has a confidence value, provenance, and is linked to H1 as its revision.

3 REPRESENTING HYPOTHESES: AN OVERVIEW

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

3.1 Comparing hypothesis representation models

In our analysis, we consider the following key aspects:

1. **Ability to describe and relate the elements in a hypothesis:** Does the vocabulary provide the means to represent entities in a hypothesis, and connect them to a record of the hypothesis provenance and revisions?
2. **Confidence model:** Does the model have a mean to describe the confidence associated with a hypothesis? Does the model describe the provenance of the confidence value in terms of how this value was obtained?
3. **Statement description:** Does the model support describing each of the statements included within a hypothesis?

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

4. **Hypothesis organization:** Does the vocabulary support a taxonomy of hypotheses?
5. **Standards:** Is the model defined using standards or does it use proprietary or idiosyncratic formats?
6. **Domain generality:** Is the model designed to be used in a specific domain, or is it a general-purpose model?

3.2 Models for representing hypotheses

This section introduces different approaches to represent hypothesis at different levels of granularity. We group them based on their granularity: coarse-grained and fine-grained representations.

3.2.1 Coarse grained hypothesis representation

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[2](SWEET) [Raskin and Pan] contain modules for defining hypotheses as "Experimental Activities". Likewise, the **Ontology for Biomedical Investigations** (OBI)[3] [Brandowski et al 2016] and the **Ontology for Clinical Research** (OCRe)[4] [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 focus of EXPO is on how the hypothesis is defined on a research paper (the "part of" relationship between the scientific experiment

[1] http://disk-project.org/ontology/disk#

[2] http://sweet.jpl.nasa.gov/2.3/reprSciModel.owl

[3] http://purl.obolibrary.org/obo/OBI_0001908

[4] http://purl.org/net/OCRe/OCRe.owl#OCRE400032

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** [5] 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 representation

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 aggrefgating hypotheses with multiple statements in *hypothesis sets*, using a Datalog representation for each hypothesis statement.

The **nanopublication model** [6] [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,[7] 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:hasConfidencevalue "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)[8], which also provides the means to describe hypothesis as literals.

The **Semantic Web Applications in Neuromedicine** (SWAN) ontology[9] **[**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 becomes 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.

[5] http://linkedscience.org/lsc/ns/
[6] http://www.nanopub.org/nschema#
[7] https://www.w3.org/TR/rdf11-concepts/
[8] http://semanticscience.org/ontology/sio.owl
[9] https://www.w3.org/TR/hcls-swan/

Table 1: Overview of the models for hypothesis representation.

Model Name	Relate hypoth. elements	Confidence model	Statement description	Hypothesis classification	Uses Standards	Domain generality
SWEET	No	No	No	No	Yes (OWL)	No
OBI	No	No	No	Yes	Yes (OWL)	No
EXPO	No	No	No	Yes	Yes (OWL)	Yes
OCR	No	No	No	No	Yes (OWL)	No
LinkedScience	No	No	No	No	Yes (OWL)	No
LABORS	Yes	No	No	Yes	Yes (OWL)	No
Nanopublications	Yes	No	Yes	No	Yes (OWL) and named graphs	Yes
Ovopublications	No	No	Yes	No	Yes (OWL), named graphs	Yes
SWAN	Yes	No	Yes	No	Yes (OWL)	No*
Micropublications	Yes	No	Yes	No	Yes (OWL) and named graphs	Yes

*The model was designed for a particular domain, but has been adopted in other domains as well.

Finally, **Micropublications**[10] [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. An illustrative example can be seen in Figure 2, 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 is compatible with the Web Annotation Ontology, [11] which allows associating a micropublication and its contents with text from articles.

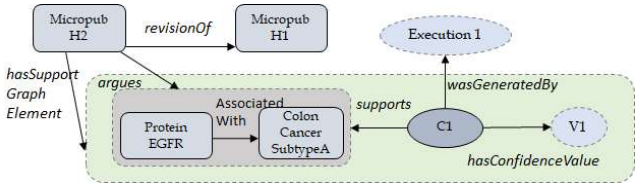

Figure 2: The example from Figure 1 adapted to the micropublication model, following [Clark et al 2014].

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. The first commonality for all the models is the lack of support for confidence on a given hypothesis. 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 *hasConfidenceValue*) linking them together. A reason why the confidence value 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 values 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 qualify hypothesis statements. Therefore, they do not meet the majority of requirements that we require for proper hypothesis representation. 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. 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 differentiator between micropublications and nanopublications is the scope of the domain. For instance, micropublications was explicitly designed to work with text statements and model their facts and argumentation. If an

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

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 take into consideration 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. [12] 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.

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 aggregating different statements, each with confidence values 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 have presented a survey of different existing vocabularies to represent hypotheses and assessed their suitability in the context of automated knowledge discovery systems. We have presented a comparison of the different models, discussing their advantages and disadvantages, and suggesting solutions for their limitations. As a result, two models emerge as potential candidates for hypothesis representation. The first one is the nanopublication model, which proposes a minimalistic model to represent facts, and the micropublication model, which aims to support the argumentation in favor or against every fact being represented. Both models provide the appropriate flexible means to represent hypotheses, but the nanopublications model is more widely adopted and provides more tooling support. Future work includes: 1) extending the DISK framework to support and align with these models, in order to publish new findings generated by the system in a more standard format, 2) extending the nanopublications model to incorporate additional requirements posed by complex hypotheses in automated discovery systems.

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 awards 1U01CA196387 and 1R01GM117097. We also acknowledge support from the Canary Foundation.

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.

[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. (30th April 2013.). The PROV ontology, w3c recommendation. Technical report, WWW Consortium.

[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 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

[12] http://nanopub.org/