SEPIO: A Semantic Model for the Integration and Analysis of Scientific Evidence

Matthew H. Brush, Kent Shefchek, Melissa Haendel
Oregon Health & Science University
Portland, OR 97239, USA
contact: brusm@ohsu.edu

***Abstract*—The Scientific Evidence and Provenance Information Ontology (SEPIO) was developed to support the description of evidence and provenance information for scientific claims. The core model represents the relationships between scientific claims (aka assertions), the sharable propositions they express belief in, the data they use as evidence, the methods and tools used to generate this data, and the agents attributable for these activities. Driving requirements for SEPIO derived largely from the data integration and analysis work of the Monarch Initiative and a related pilot project integrating cancer variant classification data. As a result, SEPIO is unique in its suitability for describing the types of evidence and provenance metadata documented in curated biomedical database. However, the SEPIO model itself is domain independent, extensible to represent any type of claim and its associated evidence, and developed in collaboration with diverse community partners in an effort to create a shared community standard. Application of SEPIO in support of curation, data integration, and claim evaluation activities can aid the development computable evidence-based knowledge networks and algorithms in support of diverse applications including variant prioritization and data quality assurance. SEPIO can be found at github.com/monarch-initiative/SEPIO-ontology/blob/master/src/ontology/sepio.owl.**

Keywords—evidence, provenance, scientific claims, ontology, data integration

I. Introduction

The scientific process aims to establish the set of facts that explains the world in which we live. Such facts begin life as hypotheses, and mature into scientific claims as a body of supporting data is generated. As support grows and opinions converge over time, a claim may become accepted as fact in the fabric of scientific knowledge. Throughout this process, the notions of *evidence* and *provenance* explain why a particular claim is believed to assert a true proposition[1] (or not), and help us to assess its proximity to scientific fact. Evidence for a claim includes any information that is used, directly or indirectly, to evaluate the validity of its proposition. Provenance information describes the process history behind a claim, including acts generating supporting data and acts evaluating this data as evidence to make claim. Together, evidence and provenance information help to place a claim in its broader scientific context, supporting improved understanding of its reliability, significance, and relevance.

Historically, the primary venue for sharing scientific claims and presenting supporting evidence has been the published literature. From the perspective of logic and philosophy, publications represent arguments [1], each built from a set of premises meant to support a logical conclusion. The task of the authors is to convey evidence showing each premise to be true, demonstrate the credibility of this evidence by describing its methodological provenance, and convince us that the logical structure of their argument is sound. If successful, there is sufficient reason to believe that the conclusion of the argument must likewise be true.

A panacea for researchers and informaticians is a formal representation of the knowledge networks that emerge by linking such arguments across publications and databases in a way that enables computational access to the complexity and nuance inherent in scientific experimentation and explanation [2]. While the seeds of such efforts are being sown in efforts such as the Micropublication movement [2, 3] and the Semantic EvidencE framework [1], there are substantial technical, pragmatic, social barriers to overcome before such a dream can be realized. At present, established database and curation efforts have succeeded primarily in codifying isolated claims [4], but not their context in broader networks that define relationships to claims they support or by which they are supported. Rather, supporting information for these claims is limited to the variable and limited evidence and provenance metadata that accompany them. At present, such metadata is inadequately and inconsistently described across most biomedical and clinical databases - offering minimal access to the supporting data, experimental processes, and assertion methods that back a claim. For example, many databases provide only references to publications purported to describe evidence for the claim, some offer evidence codes that summarize the types of evidence that exist but without revealing the evidence itself, and a few provide additional metadata about datatypes and methods used in supporting these assertions. Almost none offer comprehensive access to evidence items such as experimental measurement data, statistical confidence scores, and coded representation of assays, experimental parameters, and tools used in generating supporting data.

[1] Propositions represent the abstract, sharable meaning of what is expressed in a claim as made by a particular agent on a particular occasion. They are independent of space and time, and the primary bearers of truth value (i.e. they are either true or false). Propositions are ‘sharable’ in that the same proposition can be expressed in many different assertions.

Underlying this state of affairs is the practical reality that the expense of such deep curation is prohibitive for most systems and communities, but also the fact that no shared conceptual framework or standards exist to support efficient extraction, integration, or analysis of such metadata in curation pipelines. We posit that a necessary first step toward the longer-term vision of computable knowledge networks is the development of a shared model of evidence and provenance information that can be immediately applied to structure metadata that is currently not being leveraged in informatics applications. Toward this end, we have developed the Scientific Evidence and Provenance Information Ontology (SEPIO). SEPIO represents the relationships between scientific claims (aka assertions), the sharable propositions they express belief in, the data they use as evidence, the methods and tools used to generate this data, and the agents attributable for these activities. The core SEPIO model is domain independent, and extensible to represent any type of claim and its associated evidence and provenance information. Its application in support of curation, data integration, and claim evaluation activities is helping to lay the groundwork for richer and computable knowledge networks that will drive a new generation of semantically-enabled research innovations.

II. Development and Use Cases

SEPIO is an OWL2 ontology that is being developed according to OBO foundry principles [5], including use of the Basic Formal Ontology (BFO) as an upper ontological framework [6]. Initial development was informed largely by two driving projects in the area of genotype-to-phenotype (G2P) data integration. The Monarch Initiative[2] integrates data from model organism and human variation databases relating genotypes, phenotypes, diseases, and treatments, and structures it under a common semantic framework to support analysis and discovery using ontology-driven tools. A separate pilot project is exploring the application of similar semantic approaches to integrated analysis of cancer variant classification data, in collaboration with organizations such as the National Cancer Institute and BRCAexchange network[3]. For both of these efforts, a robust model of the evidence and provenance metadata for G2P claims is critical for users to understand, trust, evaluate, and re-use the integrated and semantically enhanced data they provide.

Though initial requirements came from these driving projects, SEPIO aspires to be a shared community model that is re-usable across domains of research, and leverages existing resources. We performed a landscape analysis of existing models, including the Provenance Ontology (PROV-O)[7], the Evidence and Conclusion Ontology (ECO)[8], the Ontology of Biomedical Investigations (OBI)[9], the Semantic EvidencE (SEE) Framework [1], the Micropublication model [2], the Drug-drug Interaction Evidence Ontology (DIDEO)[10], and the Open Biomedical Annotations (OBAN) ontology [11] (the project wiki[4] details how SEPIO relates to these models). We also engaged a diverse set of ontologists, database developers, and researchers to understand how different communities think about concepts in the domain, the terms they use to describe these concepts, and use cases they have for evidence and provenance metadata. This outreach included a Scientific Evidence Workshop[5] organized by developers and users of ontologies in this domain, including ECO, OBI, SEE, DIDEO, and MP, where participants brought use cases from diverse projects dealing in genetic, phenotype, pharmacologic, and biodiversity data.

These landscape analysis and community engagement efforts highlighted diverse use cases and unmet needs that demanded a novel representation of the entities and relationships between experimental data and the scientific claims they support. In particular, the use cases presented below drove the development of the SEPIO model:

1) **Facilitate Shared Domain Understanding and Communication:** Evidence and provenance are discussed across varying disciplines from philosophy and logic to scientific investigation and explanation, but these concepts are inconsistently defined and often conflated. This use case requires that SEPIO represent and clearly define the core concepts common across domains, provide a generic and intuitive conceptual model of the relationships between these concepts, and map terms used to reference these concepts in different communities of practice.

2) **Drive Integration of Evidence and Provenance Metadata:** Biomedical databases provide varying accounts of evidence and provenance metadata for the claims they curate and provide to the community. The 'integration' use case requires that the model support capture of the diversity of scientific claims, evidence, and provenance information across data sources, and unify them under a coherent and extensible semantic framework. SEPIO-based specifications for structuring instance data should define design patterns and modeling conventions, to facilitate consistent use of the model in data collection, integration, and exchange.

3) **Support Critical Evaluation of Scientific Claims**: In order for researchers to trust and effectively use data, it is critical that they know where it came from and how it was produced. The use case here requires that the model support critical evaluation of validity of a claim based on its lines of evidence and provenance – both by humans and through algorithmic calculation of evidence sufficiency. To achieve this, the model should clearly distinguish distinct lines of evidence for a given claim, capture whether they support or refute a claim, and when conflicting lines of evidence exist. It must also track the provenance histories for separate lines of evidence, and separate assertions of a given claim, including the relationships between data, agents, and resources relevant to each.

4) **Facilitate Discovery of Claims Based on their Evidence and Provenance:** It is often the case that scientists want to discover or filter information presented to them based on various aspects of the evidence and provenance of the

[2] https://monarchinitiative.org/

[3] http://brcaexchange.org/

4 https://github.com/monarch-initiative/SEPIO-ontology/wiki/Related-Ontologies-and-Models

[5] http://obi-ontology.org/page/Workshop_OBI-ECO_Baltimore_2016

information. This can include the type of data or studies supporting a claim, the number of evidence lines supporting or refuting it, or specific agents responsible for the claims or their supporting data. The 'discovery' use case here requires that the model is able to support queries, filtering, and presentation of information to users based on such dimensions. For example, a query such as “Find all variants associated with disease X, based on functional evidence from mouse model systems”.

5) **Enable Attribution of Researchers for Diverse Scientific Contributions:** Linked to the provenance of a scientific claim is the notion of attribution of responsible agents. This use case requires that the model support attribution of agents who generate data used as evidence, and those interpret it to support an assertion. It should also support ‘transitive attribution’ - the capacity to credit when data or resources indirectly contribute to a scientific claim.

III. The SEPIO Conceptual Model

SEPIO implements a simple and domain-independent conceptual model that is extensible to represent any type of claim and its associated evidence and provenance information at different levels of granularity. The primary axis of the SEPIO model consists of four informational entities (shown in blue in Fig. 1): assertions, propositions, supporting data items, and evidence lines.

Term: `Assertion` (aka `Claim`)

Definition: A statement of purported truth, as made by a particular agent on a particular occasion.

Example: The ENIGMA [6] consortium’s assertion that BRCA1:2685T>A causes familial breast cancer.

Comments: The identity of a particular assertion is dependent upon (1) what it claims to be true (its semantic content, aka its ‘*proposition’*), (2) the *agent* asserting it, and (3) the *occasion* on which the assertion is made. Many agents can make assertions expressing belief in the same proposition (e.g. ENIGMA’s assertion that that BRCA1:2685T>A causes familial breast cancer is a separate instance than Counsyl’s assertion of the same underlying proposition). Likewise, a single agent can make more than one assertion of belief in the same proposition on different occasions (e.g. ENIGMA may make a separate assertion of the same proposition that BRCA1:2685T>A causes familial breast cancer at a later date, based on additional evidence).

Term: `Proposition`

Definition: The ‘sharable’ meaning of what is expressed in a particular assertion.

Example: The proposition that variant BRCA1:2685T>A causes familial breast cancer

Comments: The notion of a proposition, and its relationship to an assertion, derives from the domain of logic and philosophy [12]. Propositions are abstract entities that, like numbers, are independent of space and time. They represent only the meaning that is expressed in a particular agent’s assertion, and are ‘sharable’ in that the same proposition can be expressed in many different assertions. Propositions are primary bearers of truth value, in that they are true or false.

Term: `Data Item`

Definition: A piece of information that is used to evaluate the truth of a proposition.

Example: The raw count data from the case-control study above, the calculated p-value as a measure of statistical significance, or a publication figure summarizing these data.

Comments: ‘Data item’ as used here is a broad term covering any information interpreted as evidence in evaluating a proposition. This can include primary data values, derived statistical calculations and confidence measures, or artifacts that summarize such data including publications, reports, figures, and evidence codes. As described below, such data items are created in some ‘data generation process’, and subsequently interpreted in an ‘assertion process’ that uses them as evidence to make an assertion about the truth of a proposition.

Term: `Evidence Line`

Definition: Information derived through a single line of inquiry, as used to evaluate the validity of a proposition.

Example: All information derived from a case-control study of the prevalence of the BRCA1:2685T>A in diseased vs healthy individuals, used to evaluate a particular proposition.

Comments: The information contained in an evidence line includes the set of data items generated in a given study, along with contextualizing information about their provenance that is relevant to evaluating the proposition in question. The content of a particular evidence line is defined based on its common origin in a line of investigation. Explicitly organizing all of the information that supports evaluation of a particular proposition around distinct lines of evidence and provenance is a unique and critical feature of the SEPIO model that allows for the evaluation of a given proposition based on the quantity and diversity of data supporting it.

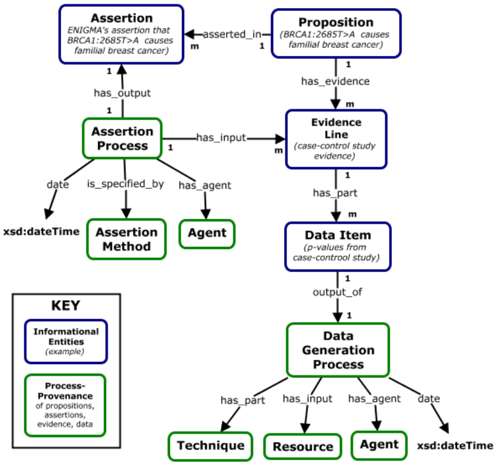


Fig. 1. The SEPIO Conceptual Model

[6] https://enigmaconsortium.org/

The provenance information about the core entities above describes the processes through which they were generated. This information is represented around two general types of processes in the SEPIO framework (green nodes in Fig. 1): an assertion process and a data generation process.

Term: Assertion Process

Definition: An act of interpreting evidence to make an assertion of belief that a particular proposition is true.

Comments: Assertion processes are affected by a particular agent on a particular occasion, and can be specified by formal assertion methods or guidelines. SEPIO implements several OWL individuals representing instances of commonly applied assertion methods, primarily those used in pathogenic variant classification such as the ACMG variant classification criteria [13].

Term: Data Generation Process

Definition: Any activity that generates information which may be used as evidence in an assertion process to evaluate the validity of a claim.

Comments: Data generation process are typically experimental studies or observations, but can include any process generating information used to evaluate a claim. SEPIO defines a hierarchy of more specific subtypes of data generation process that are most commonly used in generating data used as evidence to support claims (e.g. assay, observational study).

The relationships SEPIO defines between these six core concepts are illustrated in the conceptual model shown in Fig. 1, which includes cardinalities indicating where one entity can potentially link to more than one instance of a related entity. Here, a particular proposition can be *asserted_in* one or more assertion artifacts. A proposition *has_evidence* one or more evidence lines, which are comprised of one or more data items supporting evaluation of the proposition's truth. An assertion is the *output_of* an assertion process, which can *have_input* multiple evidence lines, but can *have_output* only a single assertion. An assertion process may be *specified_by* a particular assertion method, such as the ACMG classification guidelines. Modeling of the data generation processes in this diagram is quite minimal, illustrating a few links from a study directly to types of techniques applied and resources used. But more expressive models can be applied here that capture, for example, the temporal workflow and parameters of execution that define the study (see Discussion).

IV. Application of SEPIO Toward Disease Variant Classification

In practice, the full evidence and provenance graph around an assertion or proposition is much richer than the schematic in Fig. 1. A particular proposition is often expressed in many assertions, and can have many lines of evidence which can either support or refute it. Furthermore, each assertion may rely on a different subset of all evidence lines that exist for a given proposition, and each evidence line may be supported by multiple discrete information artifacts. The utility of the SEPIO model for accommodating such complexity is well illustrated by its application in the clinical genetics domain, where we are applying it to represent claims about the pathogenicity of suspected disease variants. Also known as 'variant classifications', these assertions typically use a five category system to describe a variant's its causal relationship with given disease (pathogenic, likely pathogenic, benign, likely benign or uncertain)[13].

Evidence and provenance information for variant classifications are particularly rich, in part because of the high stakes of clinical and research activities where these claims are relied upon, and in part because of the inherent challenge of interrogating the variant-disease relationship. Relative to propositions about gene function or variant-phenotype associations in model organisms where genes can be manipulated to provide direct evidence of a phenotypic effect, clinical genetics deals with more complex biology in experimentally intractable systems (i.e. human patients). Consequently, evidence for propositions is often less direct, more diverse, and requires more nuanced interpretation – and it is common in clinical genetics databases such as ClinVar to find many assertions of a given proposition which are based on diverse lines evidence, and often in conflict with each other.

The scenario we will explore here is modified from an exercise recently conducted by the Clinical Sequencing Exploratory Research (CSER) group [14]. It presents evidence related to the proposition that human galactosidase (GLA) gene variant NM_000169.2(GLA):c.639+919G>A is pathogenic for Fabry Disease (see ClinVar RCV000154318). A simplified account of existing evidence related to this proposition is presented below, presenting summaries of evidence lines (E1-E5) resulting from five studies relevant to the classification of the variant for Fabry Disease:

E1. Six affected individuals with the variant were found to have reduced GLA enzyme activity.

E2. The variant was absent from 528 unaffected controls.

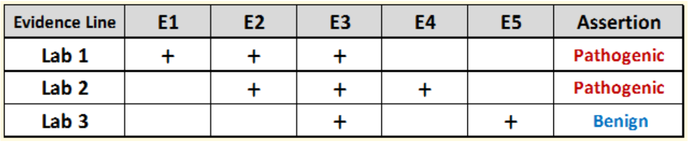
E3. Variant predicted to cause abnormal splicing that inserts additional sequence.

E4. Pedigree analyses showed Fabry Disease phenotypes segregating with the variant.

E5. Population databases show high frequency of individuals homozygous for the variant.

In our scenario, three labs independently evaluate some or all of the evidence above to make an assertion about the pathogenicity of the variant. Table I shows the evidence lines each lab utilized, and their resulting assertion. As is commonly the case, different evidence is used by each lab - either because certain data were not discoverable, or some labs judged certain data to be unreliable or irrelevant to the claim, or some labs interpreted the same data in different ways.

TABLE I: Outcomes of evidence interpretation by three independent labs (a '+' indicates the line was used by a given lab to make their assertion).



Evidence Line	E1	E2	E3	E4	E5	Assertion
Lab 1	+	+	+			Pathogenic
Lab 2		+	+	+		Pathogenic
Lab 3			+		+	Benign

SEPIO translates this scenario into the following narrative, which references particular entities to be represented in a formal model of the data. Five studies (:s1, :s2, :s3, :s4, :s5) generated many pieces of data (:d1, :d2, ... , :dn) using various research resources (:r1, r2, ... , :rn). This data was evaluated by three agents (:ag1, :ag2, :ag3) using three assertion methods: (:am1, :am2, :am3) to make three assertions (:a1, :a2, :a3) that express belief in two distinct propositions (:p1, :p2). Each assertion is based on a subset of five distinct evidence lines (:e1, :e2, :e3, :e4, :e5). The diagram in Fig. 2 shows a portion of the full graph for this scenario. Briefly, proposition :p1 represents the idea that variant NM_000169.2:c.639+919G>A is pathogenic for Fabry Disease. It is supported by evidence lines :e1, :e2, :e3, and :e4, refuted by evidence line :e5, and asserted in assertions :a1 and :a2 which express belief in this proposition. Assertion :a1 is supported by evidence lines :e1, :e2, and :e3, while assertion :a2 is supported by lines :e2, :e3, and :e4. Proposition :p2 conflicts with proposition :p1, holding that variant NM_000169.2:c.639+919G>A is benign for Fabry Disease. It is supported by evidence line :e5, refuted by evidence lines :e1, :e2, :e3, and :e4, and asserted in assertion :a3. Assertion :a3 is supported by only evidence line, :e5.

The portion of the graph described above explicitly captures what propositions exist, what evidence lines support each claim, what assertions express belief in each proposition, and what evidence lines are used by each assertion. It provides a clear picture of what lines of evidence align or refute each other, and where claims contradict each other. This is one critical aspect supporting the ability of researchers or clinicians to assess the credibility and relevance of scientific propositions, particularly when conflicting evidence or assertions exist. The remaining portion of the graph describe the provenance of the assertions, and the provenance of the evidence lines through their supporting data. This information is the second critical component allowing evaluation of scientific propositions – for example by allowing filtering or weighting of results based on who has asserted a belief in a proposition, who provided the data used as evidence in these assertions, or what techniques and resources were used in generating that data.

In Fig. 2, we have space only to illustrate as representative examples the provenance of one assertion (:a1), and one evidence line (:e3). For assertion :a1, the model captures its creation date, agent, and assertion method. Note that the pattern here does not explicitly represent the assertion process itself, instead using shortcut relations to link the assertion to its evidence (*is_assertion_supported_by*), and describe features of the assertion itself (e.g. *created_by*). While SEPIO allows for representation of the assertion process where required (see the conceptual model in Fig. 1), such modeling shortcuts can facilitate more efficient data representation and querying of the data.

Finally, modeling of evidence line :e3 includes a description of key supporting data such as a statistical measure (z-score), as well as a figure highlighting the broader context of this measure. It also captures information about participants in the study that produced supporting data, including the agent who performed it, and a particular cell line that was used. Note here that the model here is quite minimal, and SEPIO can support much more granular representation of a study as desired.

V. Discussion and Conclusions

The SEPIO framework is based on an simple, generic, and carefully defined model built around four informational artifacts (assertions, propositions, evidence lines, data items), and two types of activities that describe their creation and use (assertion and data generation processes). By clearly defining and distinguishing these concepts and supporting mappings to terms across existing models, SEPIO facilitates a shared understanding and communication that will drive development of aligned data models and integration efforts. By defining data model specifications based on this conceptual model and informed by real data use cases from driving projects, we are iteratively developing a new standard for the representation, exchange, and analysis of evidence and provenance metadata for scientific claims.

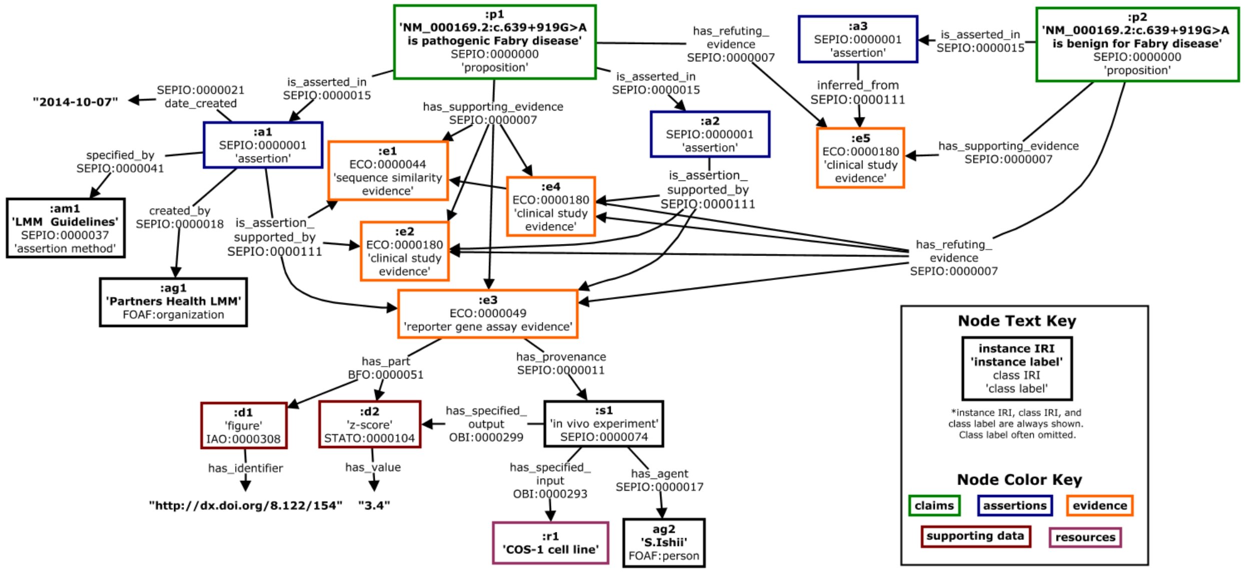

:p1
'NM_000169.2:c.639+919G>A is pathogenic Fabry disease'
SEPIO:0000000
'proposition'
has_refuting_ evidence SEPIO:0000007
:a3
SEPIO:0000001
'assertion'
is_asserted_in SEPIO:0000015
:p2
'NM_000169.2:c.639+919G>A is benign for Fabry disease'
SEPIO:0000000
'proposition'
"2014-10-07"
SEPIO:0000021 date_created
is_asserted_in SEPIO:0000015
is_asserted_in SEPIO:0000015
inferred_from SEPIO:0000111
has_supporting_evidence SEPIO:0000007
:a1
SEPIO:0000001
'assertion'
:a2
SEPIO:0000001
'assertion'
specified_by SEPIO:0000041
:e1
ECO:0000044
'sequence similarity evidence'
:e5
ECO:0000180
'clinical study evidence'
has_supporting_evidence SEPIO:0000007
is_assertion_ supported_by SEPIO:0000111
:e4
ECO:0000180
'clinical study evidence'
:am1
'LMM Guidelines'
SEPIO:0000037
'assertion method'
created_by SEPIO:0000018
is_assertion_ supported_by SEPIO:0000111
:e2
ECO:0000180
'clinical study evidence'
has_refuting_ evidence SEPIO:0000007
:ag1
'Partners Health LMM'
FOAF:organization
:e3
ECO:0000049
'reporter gene assay evidence'
has_part BFO:0000051
has_provenance SEPIO:0000011
:d1
'figure'
IAO:0000308
:d2
'z-score'
STATO:0000104
has_specified_ output OBI:0000299
:s1
'in vivo experiment'
SEPIO:0000074
has_identifier
has_value
has_specified_ input OBI:0000293
has_agent SEPIO:0000017
"http://dx.doi.org/8.122/154"
"3.4"
:r1
'COS-1 cell line'
ag2
'S.Ishii'
FOAF:person
Node Text Key
instance IRI
'instance label'
class IRI
'class label'
*instance IRI, class IRI, and class label are always shown. Class label often omitted.
Node Color Key
claims
assertions
evidence
supporting data
resources

Fig 2: Application of SEPIO toward Modeling Variant Classification Data

A key gap in existing models and practices is support for computational evaluation of claims based on the quality, diversity, and provenance of available evidence. Here SEPIO uses the notion of an **evidence line** to organize data supporting a given claim according to its experimental origins. Evidence lines are described by their 'type' based on the ECO ontology, and by links to OBI terms representing scientific techniques and resources used the creation of supporting data. The structure of the ECO and OBI ontologies can be exploited by semantic similarity algorithms such as OWLSim [15] to understand the diversity and quality of evidence for a given claim. Take for example conflicting assertions about a proposition that a particular variant is causal for a specific disease. One assertion is based on four lines of in vitro evidence based on similar methodology and model systems, and provided by the same lab. The second assertion has two lines of evidence – the first from an in vivo mouse model study, and the second a rigorous statistical analysis of variant frequencies in human populations. Semantic similarity metrics can highlight the superior diversity and quality of evidence for the second assertion based on the distance between evidence types and supporting techniques in the ECO and OBI graph structures, respectively (the assumption being that more diverse and independent lines of evidence provide stronger reason to believe the claim to be true). Furthermore, application-specific rules about the inherent 'quality' of different techniques or research resources could be layered onto ontological graph structures to support an additional means for automated ranking of evidence lines, and generating confidence metrics around scientific claims.

Even with support from computational evaluation methods, human review of evidence for scientific claims will continue to be necessary. Here, the context in which a model captures evidence and provenance metadata needs to support the ability of different communities to customize and weight the types of evidence they want to rely upon for a given application at a granular level. For example, a medical genetics pipeline may want to evaluate disease-variant associations in absence of evidence supported by in vitro data they have decided is not reliable enough to be applied in clinical settings, or eliminate assertions made by a particular organization before running an analysis. The distinctions and links SEPIO draws between propositions, assertions, evidence lines, and supporting data have been expressly developed to support such use cases.

The utility of such automated and manual approaches to evidence and claim evaluation is of course dependent on the creation of rich and consistent metadata in the first place. Here we believe that SEPIO can support innovative curation tools that enable capture of precise evidence and provenance metadata that is currently reviewed in the process of annotating to an ECO code, but not explicitly represented in most curated databases. An intuitive standard for capture and exchange of such data that supports novel and integrative analyses and evaluation use cases, can offer incentive for databases to invest in pipelines and tools that meet these standards.

Finally, an area of future work for SEPIO is to define design patterns for representing the experimental provenance of data used as evidence at different levels of granularity. As noted, this information is critical for understanding and evaluating a given claim, but representing a complete experimental workflow is time and resource intensive, and not necessary for many applications. We are working with related community efforts including OBI and KEfED to provide interoperable representations of experimental provenance ranging from simple links to *types* of techniques and study participant's relevant to a line of evidence, to detailed temporal representations of workflows that specify their *particular* processes and participants, and the experimental variables that parameterize a given study. This interoperable flexibility will be critical for widespread adoption and integrated data analysis use cases supported by the SEPIO framework.

Acknowledgements

The authors would like to thank Mark Diekhans for BRCA use cases; Laura Amendola and Gail Jarvik for information about CSER classification exercises; Tom Conlin for technical assistance with data integration pipelines; and James Overton for his philosophical perspective on evidence and argument. This work was funded by NCI/Leidos #15X143, BD2K U54HG007990-S2 (Haussler).

References

[1] Bölling, Christian, Michael Weidlich, and Hermann-Georg Holzhütter. "SEE: structured representation of scientific evidence in the biomedical domain using Semantic Web techniques." Journal of biomedical semantics 5.1 (2014): 1.

[2] Clark, Tim, Paolo N. Ciccarese, and Carole A. Goble. "Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications." *Journal of biomedical semantics* 5.1 (2014): 1.

[3] Schneider, Jodi, et al. "Using the Micropublications ontology and the Open Annotation Data Model to represent evidence within a drug-drug interaction knowledge base." Proceedings of the 4th International Conference on Linked Science-Volume 1282. CEUR-WS. org, 2014.

[4] Fernández-Suárez XM, Galperin MY: The 2013 Nucleic Acids Research Database Issue and the online molecular biology database collection. Nucleic Acids Res., 41: D1-D7. 10.1093/nar/gks1297. 2013

[5] Smith, Barry, et al. "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration." Nature biotechnology 25.11, 1251-1255, 2007

[6] Arp, Robert, Barry Smith, and Andrew D. Spear. Building ontologies with basic formal ontology. Mit Press, 2015.

[7] PROV-O W3C Recommendation, https://www.w3.org/TR/prov-o/

[8] Chibucos, Marcus C., et al. "Standardized description of scientific evidence using the Evidence Ontology (ECO)." Database 2014

[9] Courtot, Mélanie, et al. "The OWL of Biomedical Investigations." OWLED. Vol. 432. 2008.

[10] Brochhausen, Mathias, et al. "Towards a foundational representation of potential drug-drug interaction knowledge."DIKR, Houston, 2014.

[11] Sarntivijai, Sirarat, et al. "Linking rare and common disease: mapping clinical disease-phenotypes to ontologies in therapeutic target validation." J Biomed Semantics. 2016 Mar 23;7:8.. eCollection 2016.

[12] Stanford Encyclopedia of Philosophy, accessted May 3, 2016 at http://plato.stanford.edu/entries/propositions/

[13] Richards, Sue, et al. "Standards and guidelines for the interpretation of sequence variants." Genetics in Medicine, 2015.

[14] Amendola, Laura M., et al. "Performance of ACMG-AMP Variant-Interpretation Guidelines among Nine Laboratories in the Clinical Sequencing Exploratory Research Consortium." The American Journal of Human Genetics, 2016.

[15] Chen, Chao-Kung, et al. "MouseFinder: candidate disease genes from mouse phenotype data." *Human mutation* 33.5, 858-866, 2012