Modeling Evidence and Provenance Information for G2P Data Integration and Analysis

Matthew Brush
OHSU Ontology Development Group

Evidence and Provenance

. . . explains how we know claims to be true (or not)

Evidence is information used to evaluate the validity of a claim

- primary data, derived statistical scores, figures/summaries, other statements or claims

Provenance describes the process history of a claim

- who made the claim, how, when
- who generated data used as evidence, how, when

Evidence and Provenance

- **Evidence** and **provenance** metadata for scientific claims is critical to evaluate their credibility, assess utility, and extend findings.
- Such metadata is **inadequately** and **inconsistently** described in most biomedical and clinical databases.
- Result is a lack of **computable metadata** to support **integration**, **discovery**, **evaluation**, and **analysis**.

Goals and Use Cases

Develop a generic and extensible data model for representing evidence and provenance metadata across biomedical databases

1. **Integration:** Support curation and aggregation of evidence and provenance metadata across biomedical databases
2. **Evaluation:** Support manual and computational evaluation of claims based on its lines of evidence and provenance.
3. **Attribution:** Enable tracking agent provenance through distinct assertions and lines of evidence
4. **Communication:** Facilitate shared understanding and communication around claims, evidence, provenance
5. **Discovery:** Support finding and filtering claims based on various aspects of their evidence and provenance
6. **Analysis:** Leverage evidence and provenance metadata to understand how we know what we know, and why we don't know what we don't

Competency Questions

1. *Find all genetic variants associated with disease X, based on functional evidence from mouse model systems.*
2. *Show me what types of data and experiments support a given claim?*
3. *Given two conflicting claims, which is more likely to be correct?*
4. *Find variants of uncertain significance where my new variant population frequency data may be useful in making a definitive classification.*
5. *Are there certain factors (data types, methods, researchers, classification guidelines) that are common to disputed claims (i.e. conflict with other claims)?*
6. *What researchers are most widely attributed for methods or data used as evidence supporting BRCA variant classifications.*

SEPIO

Scientific Evidence and Provenance Information Ontology

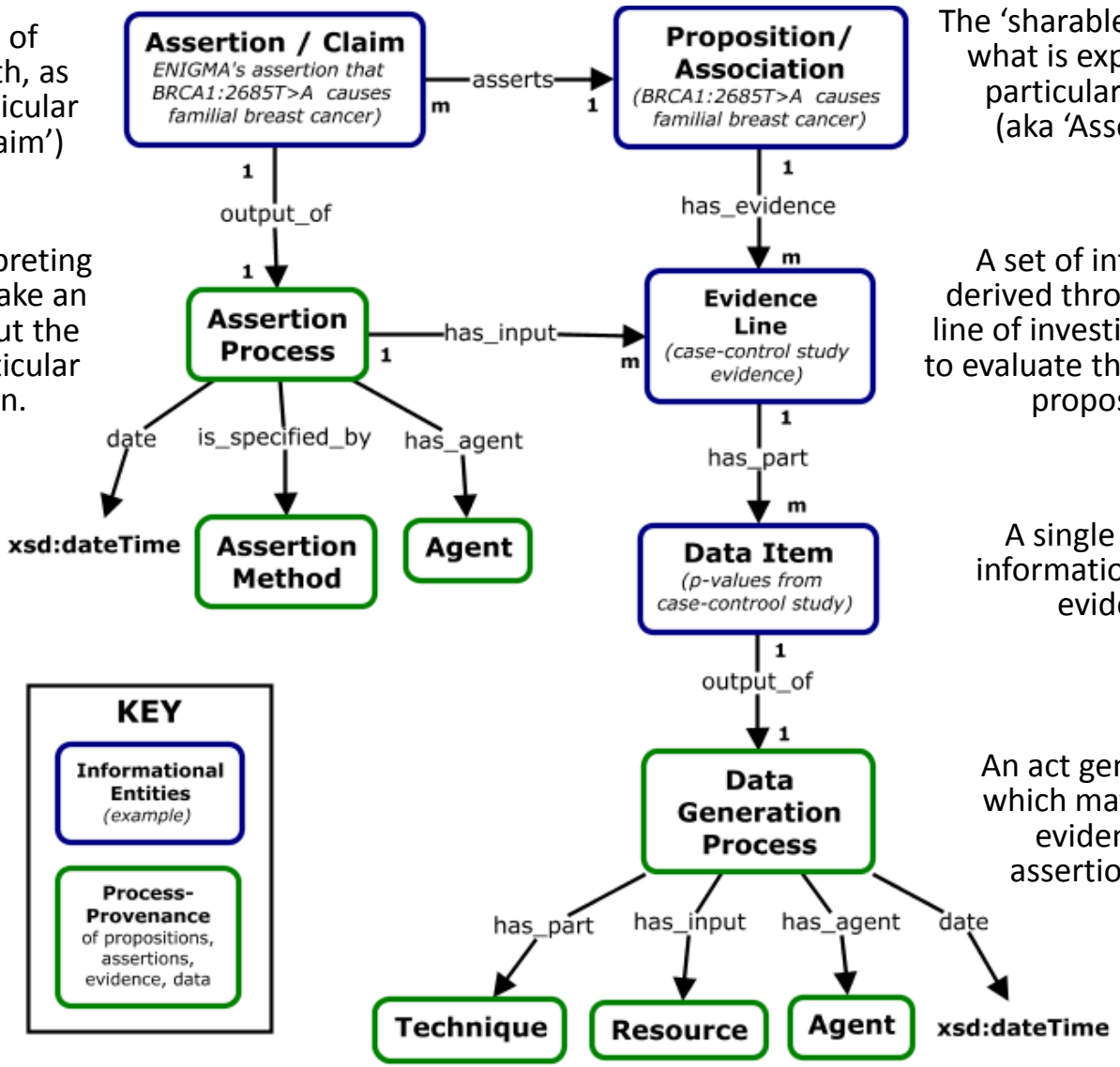
- OWL2 ontology and data model specification to structure evidence and provenance metadata for scientific claims
- Represents the relationships between **propositions**, **assertions** of belief in a particular proposition, **data** used as **evidence** by such assertions, **methods** and **tools** used to generate this data, and the **agents** attributable for all of these activities.
- Core model is domain independent, and extensible to represent any type of claim and its evidence and provenance information

https://github.com/monarch-initiative/SEPIO-ontology

Conceptual Model

A statement of purported truth, as made by a particular agent (aka 'Claim')

An act of interpreting evidence to make an assertion about the truth of a particular proposition.



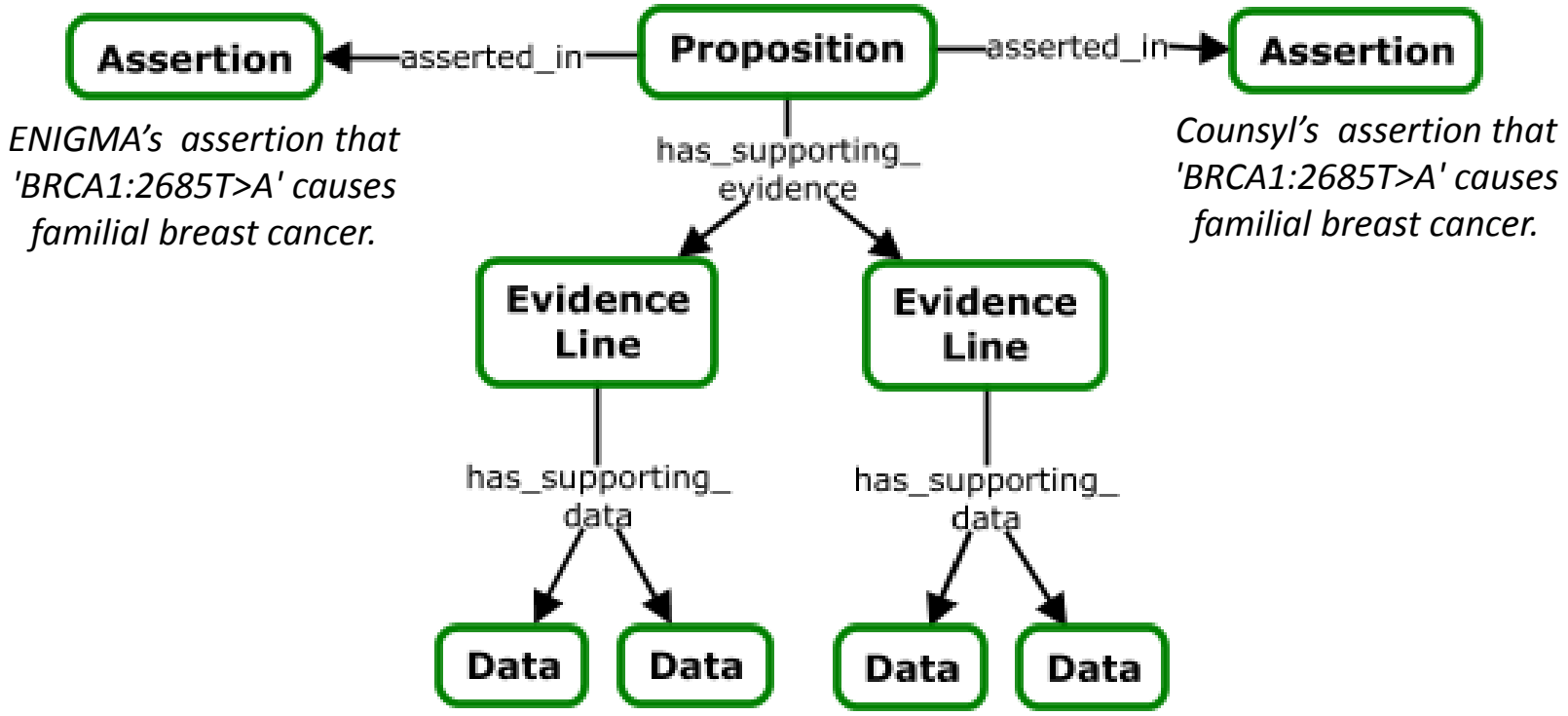
The 'sharable' meaning of what is expressed in a particular assertion (aka 'Association')

A set of information derived through a single line of investigation, used to evaluate the validity of a proposition.

A single piece of information used as evidence

An act generating data which may be used as evidence in an assertion process.

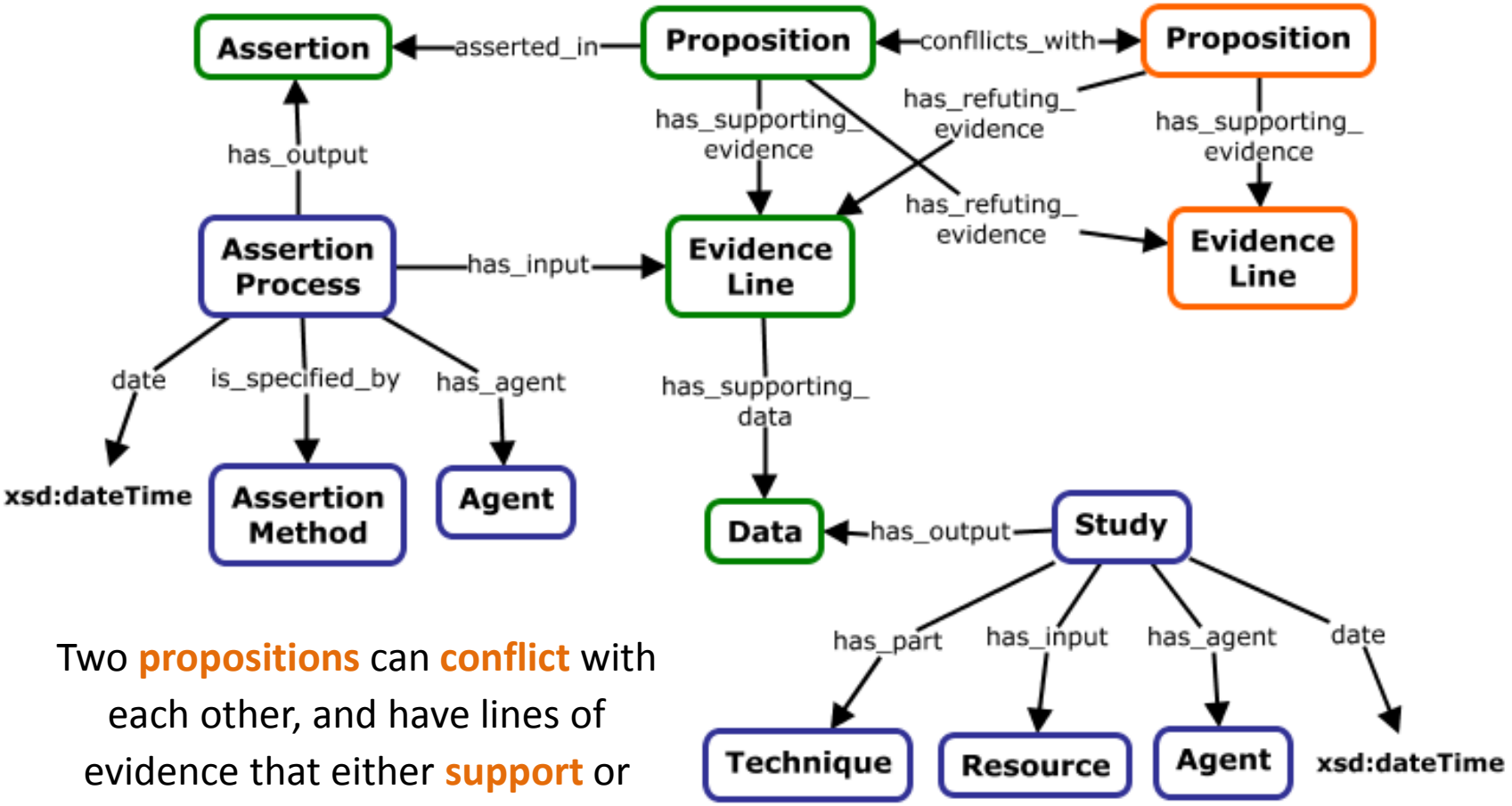
Conceptual Model



A given **proposition** can be stated in many **assertions** and have many **lines of evidence**, each of which may be supported by many pieces of **data**.

Evidence lines consist of evidence information derived from one line of inquiry.

Conceptual Model



Two **propositions** can **conflict** with each other, and have lines of evidence that either **support** or **refute** them

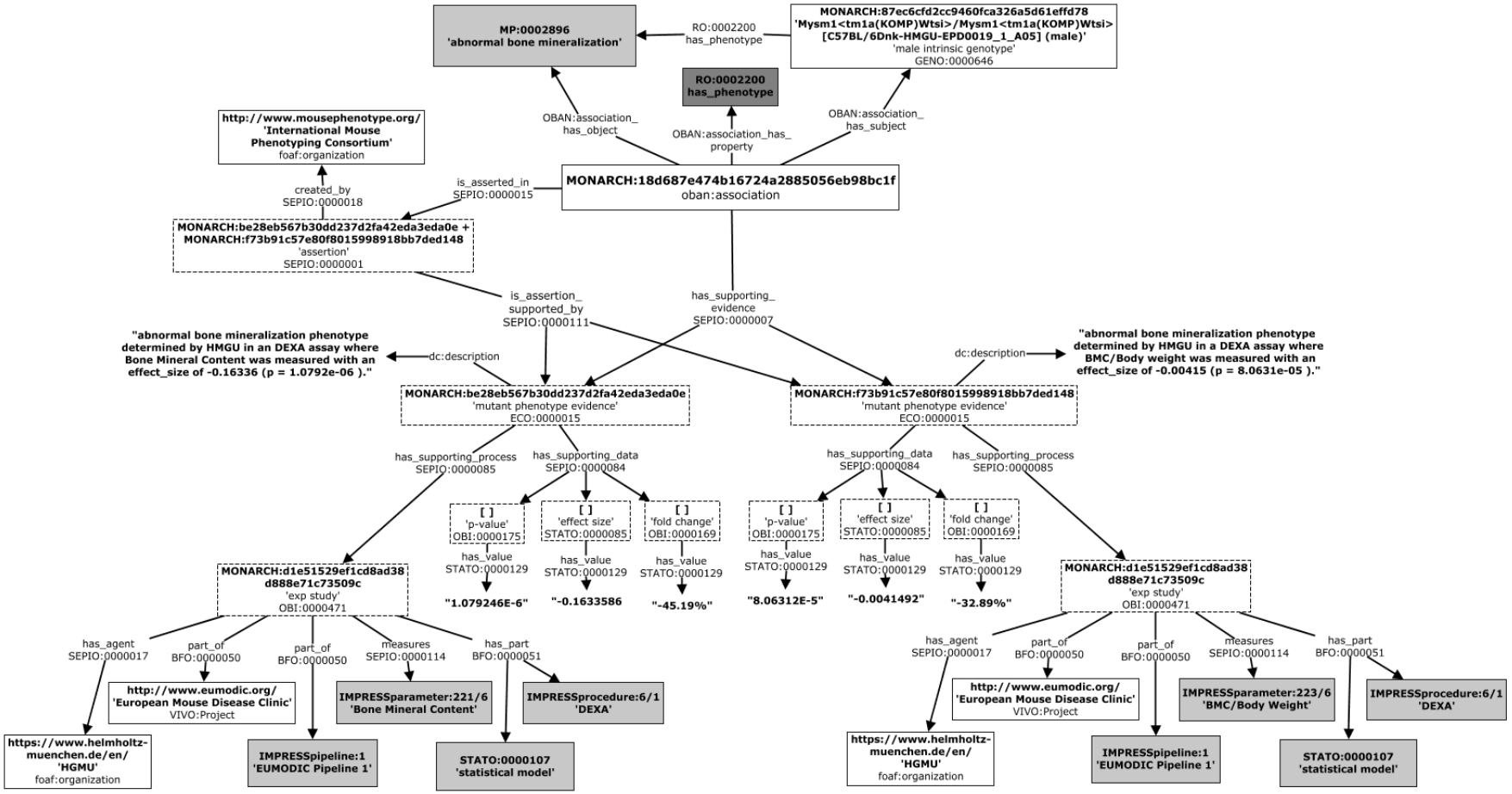
Initial Applications

- Pilot in domain of **genotype-phenotype (G2P) associations**
  - Describe how variation impacts gene function or expression, contributes to disease, affects efficacy of treatments, etc.
- Integrating association and evidence/provenance data from three sources: **ClinVar**, **IMPC**, and **neXtProt**
  - different types of G2P claims/associations
  - different sources of data
  - different types of evidence and provenance metadata collected
  - different data and curation models
- Status:
  - initial pass at **ClinVar** and **IMPC** data ingested into a SciGraph database
  - waiting on **neXtProt** to refactor data model and metadata coverage

IMPC Data

- **Claims**: mouse genotype causes phenotype
  - e.g. Mysm1$^{tm1a(KOMP)/tm1a(KOMP)}$[C57BL/6] causes 'abnormal bone mineralization'
- **Source**: direct from centrally-coordinated phenotyping efforts (single source of consistent and standardized data)
- **Evidence:** primary data, statistical summaries and scores (p-values, effect sizes)
- **Provenance:**
  - **For the Assertion**: agent who made claim
  - **For the Supporting Evidence:** agent, procedure/parameters, stat. methods
- **Challenges:**
  - reliance on internal vocabularies
  - incomplete coverage of genes and phenotypes (no BRCA mutants, no cancer phenotypes assessed)

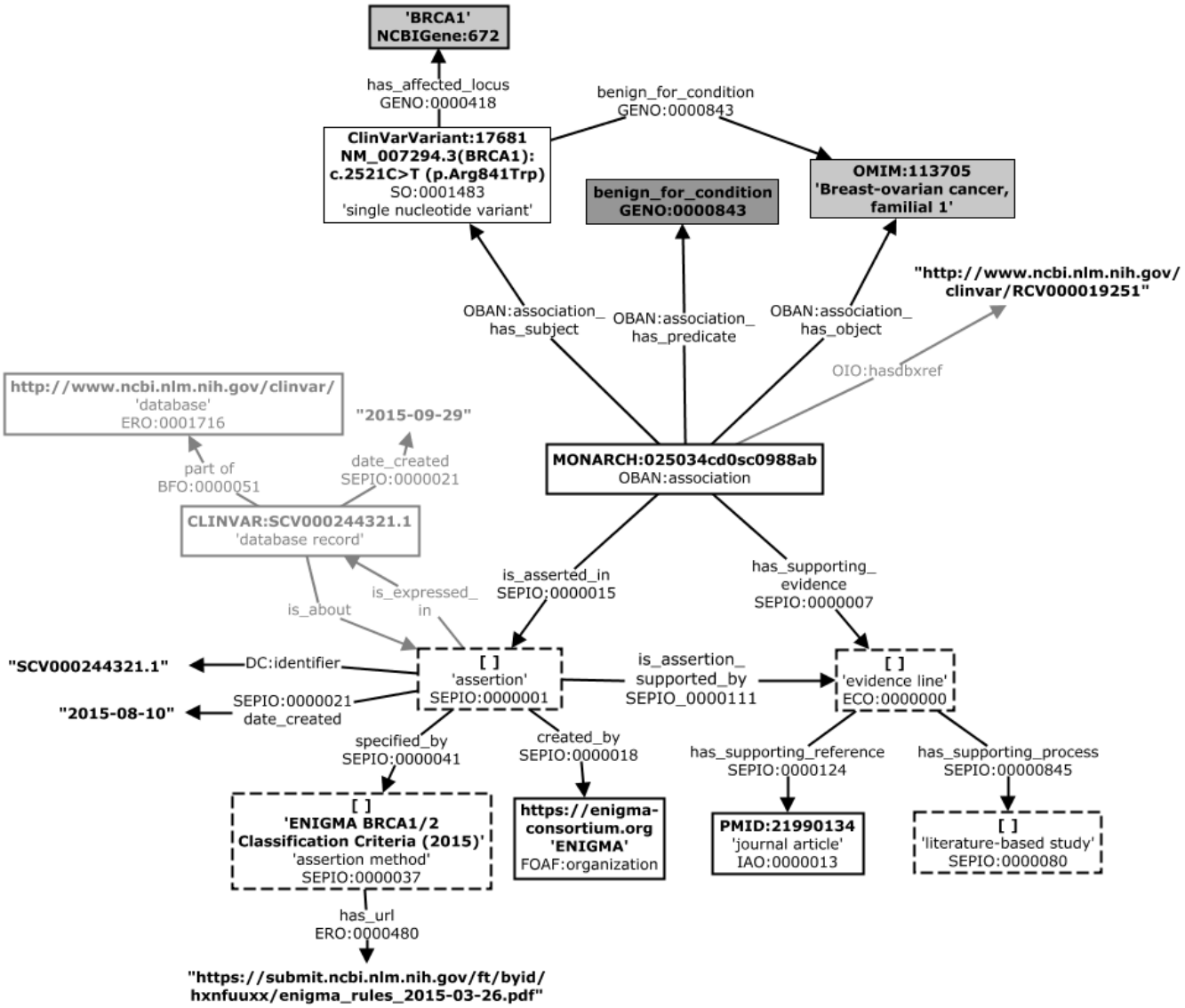
IMPC Data Example



ClinVar Data

- **Claims**: human variant pathogenicity for a specific disease
  - e.g. (BRCA1):c.2521C>T (p.Arg841Trp) pathogenic for 'Breast-ovarian cancer'
- **Sources**: aggregate claims from direct submissions and existing databases – many assertions of same classification
- **Evidence:** minimal (only references to publications)
- **Provenance:**
  - **For the Assertion**: agent making claim, when, classification method used
  - **For the Supporting Evidence:** type of study, and sparse metadata about study participants (e.g. ethnicity, family history) and genotyped specimens (e.g. source tissue, genotyping method)
- **Challenges:**
  - rich data model but sparsely and inconsistently populated data
  - lack of useful evidence metadata, inability to distinguish evidence lines
  - reliance on internal vocabularies

ClinVar Data Example



NeXtProt Data

- **Claims**: human variant 'functional annotations' describing molecular defects of a variant protein
  - e.g. 'BRCA1-p.Arg1699Gln' has decreased localization to the nucleus
- **Source**: curated from literature
- **Evidence:** ECO codes, publications, relevant figures
- **Provenance:**
  - **For Assertion**: agent making claim
  - **For Supporting Evidence:** methods from ECO codes, biological model used (e.g. species, cell line, strain/mutations), free text description
- **Challenges:**
  - undergoing data model refactor , reliance on internal vocabularies
  - limited supporting evidence data
  - different biological scale vs traditional G2P data

Take Homes

- Available data insufficient for meeting many critical use cases and competency questions
- Valuable information is ‘left behind’ in curation pipelines
  - The reported outputs of **variant classification** and **evidence code assignment** are only fraction of information processed in the curation process
- Rich data models do not amount to rich data
- Standard vocabularies and data models desperately needed
- Conceptual and terminological divides are significant

Community Outreach and Coordination

1. Talk at Biocuration Conference, Geneva (April 2016)
2. Paper/Presentation for ICBO (Corvallis, August 2016)
3. Organized a Scientific Evidence workshop attended by members of ontology, informatics, and database communities (Baltimore, May 2016)
4. Consultation and outreach to various databases to facilitate collection and modeling of evidence metadata (BRCAexchange, Facebase, neXtProt)
5. Contracted to make comprehensive improvements to structure and usability of ECO
6. Joined working groups developing shared, extensible representations of experimental techniques and workflows
7. Planning outreach to individual variant classification efforts to obtain lost metadata (need help identifying partners)

Data Demo: ClinVar and IMPC

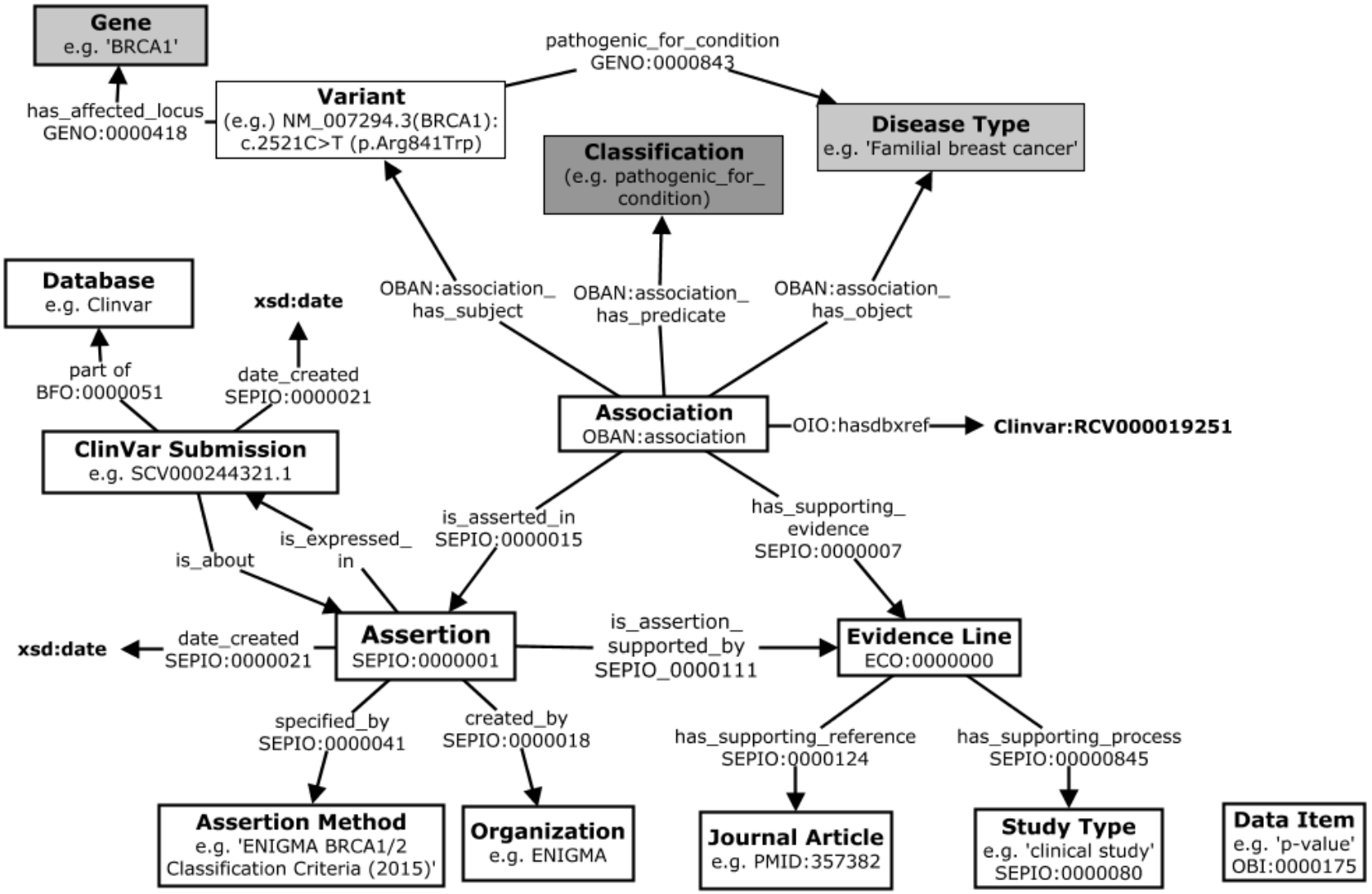
ClinVar Query Set 1: Basic discovery and filtering of variant classifications (BRCA1 gene)

1. Find all BRCA 1 variants that are asserted to play pathogenic role in a disease
2. Show counts of assertions for each disease classification and order by the counts
3. Facet to variants contributing to breast cancers
4. Show each assertion and its evidence and provenance metadata

ClinVar Query Set 2: Exploring conflicting assertions/evidence (VHL gene)

1. Show classifications that have strongly conflicting assertions (pathogenic vs benign)
2. Facet to conflicting assertions for VHL gene variants
3. View evidence and provenance metadata for these assertions
4. Show other classifications involving the disputed variant

ClinVar Record Example



SEPIO Applied in the Domain of Medical Genetics

Modeling ClinVar Data: Aggregated assertions of variant pathogenicity

Evidence in Medical Genetics

- Domain: diagnosis & management of hereditary disorders
- G2P = Variant-Disease Associations ('**classifications**')
  - classify the pathogenicity of a variant for a given disease, or its effect on disease course or response to therapeutics
- Variant -Disease relationship is more complex/nuanced than other types of associations
  - evidence is less direct relative to gene function annotations, or model organism G2P associations
  - reliance on numerous, diverse, and often inconsistent lines of evidence
  - need to critically evaluate strength of claims is critical to research and patient care

Variant Classification Guidelines

- **Variant classification guidelines** provide standards for weighting evidence and using it to assert variant pathogenicity in consistent and rigorous way.
- Most use a 5 point scale: Pathogenic, Likely Pathogenic, Uncertain Significance, Likely Benign, Benign
- The **ACMG Guidelines** are most rigorous and widely-used

16 criteria that support a variant being pathogenic

- ranked as Very Strong (**PVS**), Strong (**PS**), Moderate (**PM**), or Supporting (**PP**)

12 criteria that support a variant being benign

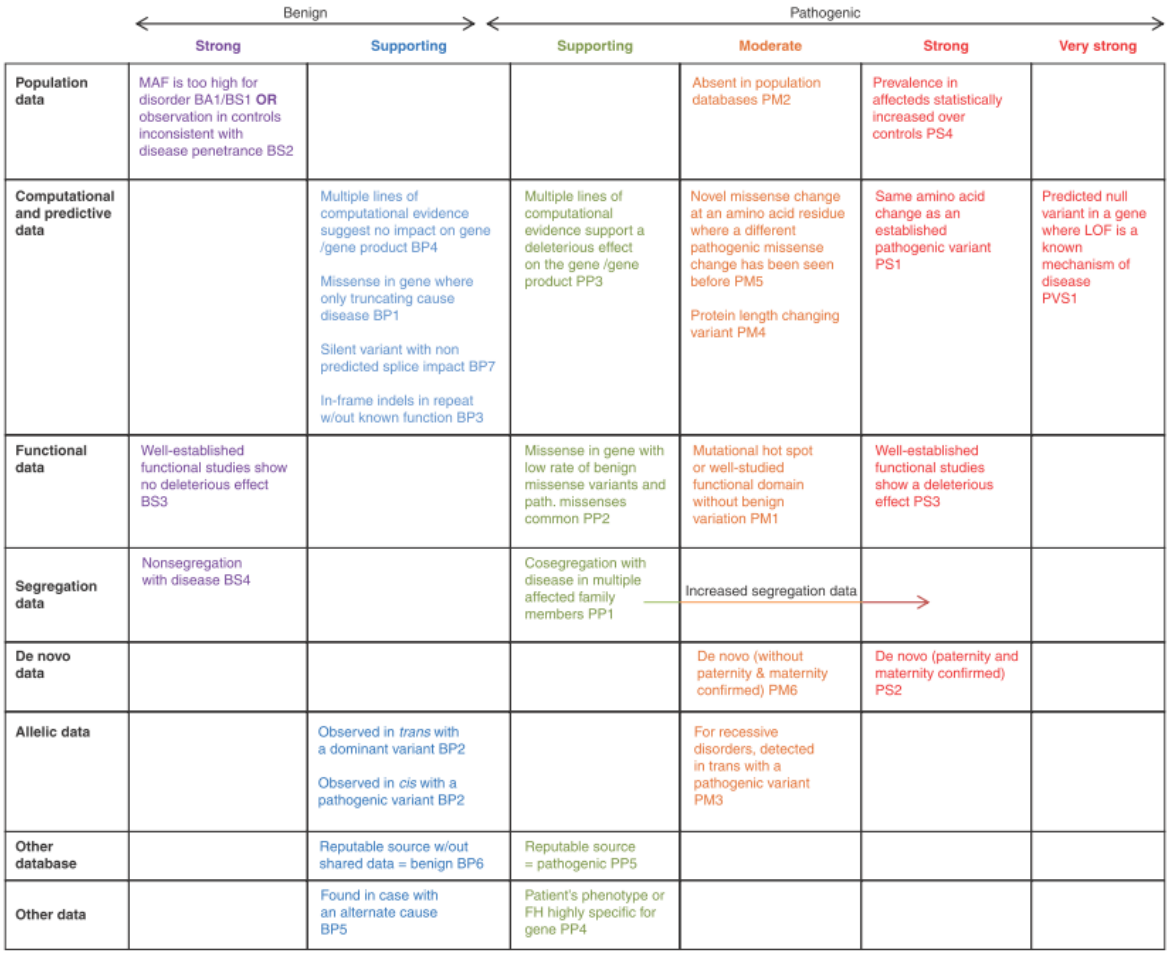
- ranked as Stand-alone (**BA**), Strong (**BS**), or Supporting (**BP**)

ACMG Variant Classification Guidelines

Criteria represent evidence of different types/sources

- **Computational Evidence** - predicted impact on gene structure and function
  - **PVS1:** Variant is predicted to be a null variant *of a gene where LOF is a known mechanism of disease*
  - **BP4:** Variant not expected to have deleterious affect based on multiple lines of computational evidence
- **Functional Evidence** - experimentally validated effect of variant on gene function
  - **PS3**: Variant shown deleterious by functional studies in vitro/in vivo (PS3)
- **Population evidence -** based on frequency of variant in healthy vs affected populations
  - **PS4:** Variant prevalence in affected individuals is statistically increased over controls
  - **BS1**: Variant allele frequency is too high for disorder (>5%/too high for disorder)
- **Genetic Patterns of Trait Heritability/Penetrance**
  - **PP1**: Variant co-segregates with disease in multiple affected family members

ACMG Evidence Framework



	Benign		Pathogenic			
	Strong	Supporting	Supporting	Moderate	Strong	Very strong
Population data	MAF is too high for disorder BA1/BS1 **OR** observation in controls inconsistent with disease penetrance BS2			Absent in population databases PM2	Prevalence in affecteds statistically increased over controls PS4	
Computational and predictive data		Multiple lines of computational evidence suggest no impact on gene /gene product BP4 Missense in gene where only truncating cause disease BP1 Silent variant with non predicted splice impact BP7 In-frame indels in repeat w/out known function BP3	Multiple lines of computational evidence support a deleterious effect on the gene /gene product PP3	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before PM5 Protein length changing variant PM4	Same amino acid change as an established pathogenic variant PS1	Predicted null variant in a gene where LOF is a known mechanism of disease PVS1
Functional data	Well-established functional studies show no deleterious effect BS3		Missense in gene with low rate of benign missense variants and path. missenses common PP2	Mutational hot spot or well-studied functional domain without benign variation PM1	Well-established functional studies show a deleterious effect PS3	
Segregation data	Nonsegregation with disease BS4		Cosegregation with disease in multiple affected family members PP1	Increased segregation data →		
De novo data				De novo (without paternity & maternity confirmed) PM6	De novo (paternity and maternity confirmed) PS2	
Allelic data		Observed in *trans* with a dominant variant BP2 Observed in *cis* with a pathogenic variant BP2		For recessive disorders, detected in trans with a pathogenic variant PM3		
Other database		Reputable source w/out shared data = benign BP6	Reputable source = pathogenic PP5			
Other data		Found in case with an alternate cause BP5	Patient's phenotype or FH highly specific for gene PP4			

Figure 1 Evidence framework. This chart organizes each of the criteria by the type of evidence as well as the strength of the criteria for a benign (left side) or pathogenic (right side) assertion. Evidence code descriptions can be found in **Tables 3** and **4**. BS, benign strong; BP, benign supporting; FH, family history; LOF, loss of function; MAF, minor allele frequency; path., pathogenic; PM, pathogenic moderate; PP, pathogenic supporting; PS, pathogenic strong; PVS, pathogenic very strong.

Table 5 Rules for combining criteria to classify sequence variants

Pathogenic	(i) 1 Very strong (PVS1) *AND* (a) ≥1 Strong (PS1–PS4) *OR* (b) ≥2 Moderate (PM1–PM6) *OR* (c) 1 Moderate (PM1–PM6) and 1 supporting (PP1–PP5) *OR* (d) ≥2 Supporting (PP1–PP5) (ii) ≥2 Strong (PS1–PS4) *OR* (iii) 1 Strong (PS1–PS4) *AND* (a)≥3 Moderate (PM1–PM6) *OR* (b)2 Moderate (PM1–PM6) *AND* ≥2 Supporting (PP1–PP5) *OR* (c)1 Moderate (PM1–PM6) *AND* ≥4 supporting (PP1–PP5)
Likely pathogenic	(i) 1 Very strong (PVS1) *AND* 1 moderate (PM1–PM6) *OR* (ii) 1 Strong (PS1–PS4) *AND* 1–2 moderate (PM1–PM6) *OR* (iii) 1 Strong (PS1–PS4) *AND* ≥2 supporting (PP1–PP5) *OR* (iv) ≥3 Moderate (PM1–PM6) *OR* (v) 2 Moderate (PM1–PM6) *AND* ≥2 supporting (PP1–PP5) *OR* (vi) 1 Moderate (PM1–PM6) *AND* ≥4 supporting (PP1–PP5)
Benign	(i) 1 Stand-alone (BA1) *OR* (ii) ≥2 Strong (BS1–BS4)
Likely benign	(i) 1 Strong (BS1–BS4) and 1 supporting (BP1–BP7) *OR* (ii) ≥2 Supporting (BP1–BP7)
Uncertain significance	(i) Other criteria shown above are not met OR (ii) the criteria for benign and pathogenic are contradictory

Pathogenic | Likely Pathogenic | Benign | Likely Benign | Uncertain Significance

Modeling Test Case

Evaluating the pathogenicity of galactosidase gene variant
NM_000169.2(GLA):c.639+919G>A *for* ***Fabry disease***

Findings/Evidence from relevant studies:

1. Six affected individuals with the variant had reduced GLA enzyme activity.
   PS3*: functional study shows detrimental effect*

2. Variant absent from 528 unaffected race matched controls
   PS4*: prevalence in cases > controls*

3. Variant predicted to cause abnormal splicing that inserts additional sequence
   PM4*: protein length changing variant*

4. Family history showed Fabry-associated phenotypes segregating with variant
   PP1*: co-segregate with disease in family members*

5. Population database shows high frequency of individuals homozygous for variant
   BA1*: variant population frequency inconsistent with role for variant in disorder*

Classification Outcomes

Evidence	PS3	PS4	PM4	PP1	BA1	Assertion
Agent 1	x	x	x			Pathogenic
Agent 2		x	x	x		Pathogenic
Agent 3			x		x	Benign

Three agents evaluate data to classify the pathogenicity of the variant, leading to three assertions.

Why get different assertions?

1. Use different data (lack of discovery/access)
2. Interpret the same data differently
3. Use different classification guidelines to assign classification

Modeling the Evidence and Provenance

Evidence	PS3	PS4	PM4	PP1	BA1	Assertion
Agent 1	x	x	x			Pathogenic
Agent 2		x	x	x		Pathogenic
Agent 3			x		x	Benign

Five studies (:s1, :s2, :s3, :s4, :s5)

. . . generated many pieces of data (:d1, :d2, . . . , :dn)

. . . .that was evaluated by three agents (:ag1, :ag2, :ag3)

. . . using three assertion methods: (:am1, :am2, :am3)

. . . to make three assertions (:a1-p, :a2-p, :a3-lb)

. . . that express two distinct claims (:c1-p, :c2-lb)

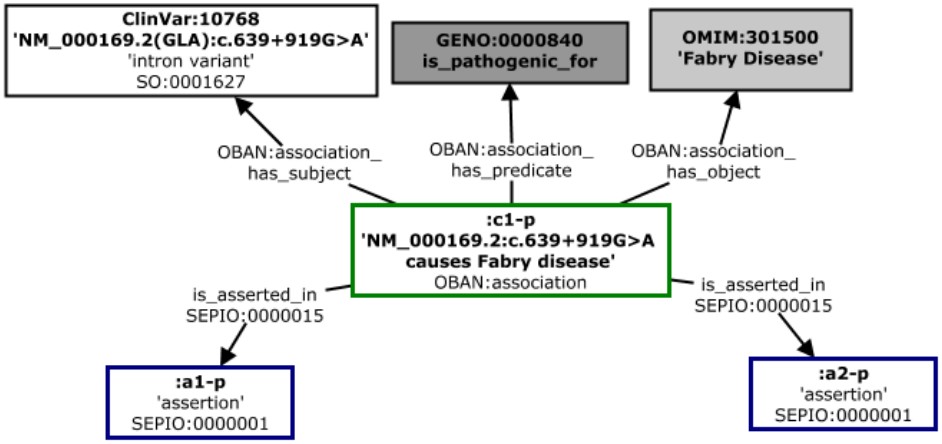
. . . based on five distinct lines of evidence (:e1, :e2, :e3, :e4, :e5)

The Data, Modeled

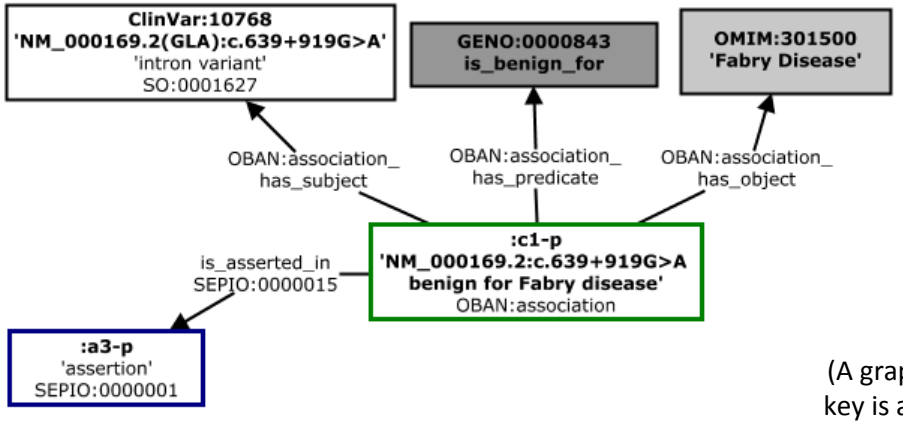
Instance-level illustration of the exemplar scenario as an rdf graph.

We start by viewing the two associations emerging from this scenario.

Association :c1-p classifies the variant as pathogenic, and is asserted in two **assertions**.

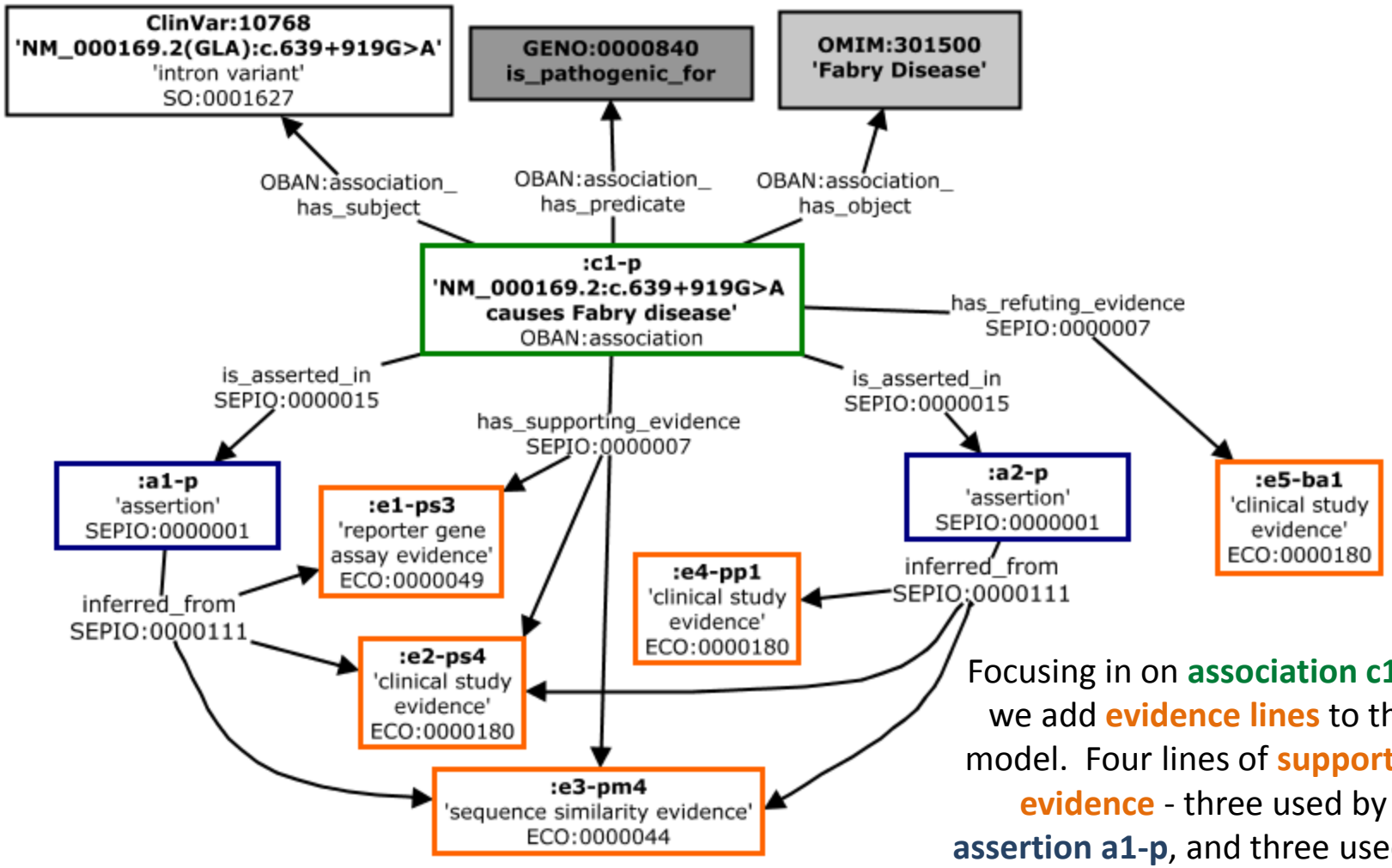


Association :c2-b classifies the variant as benign, and is asserted in one **assertion**.



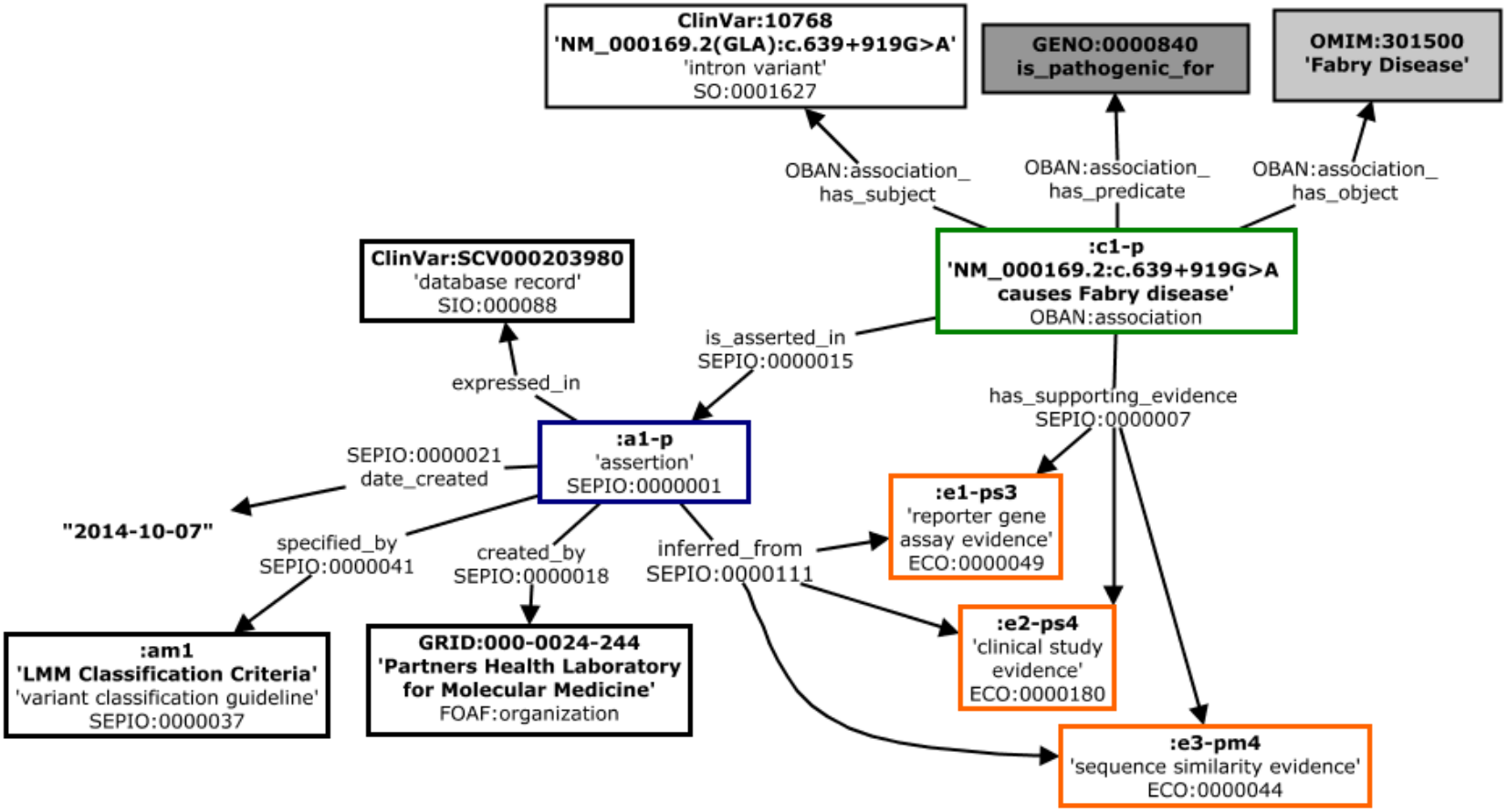
(A graphical notation key is at end of deck)

The Data, Modeled



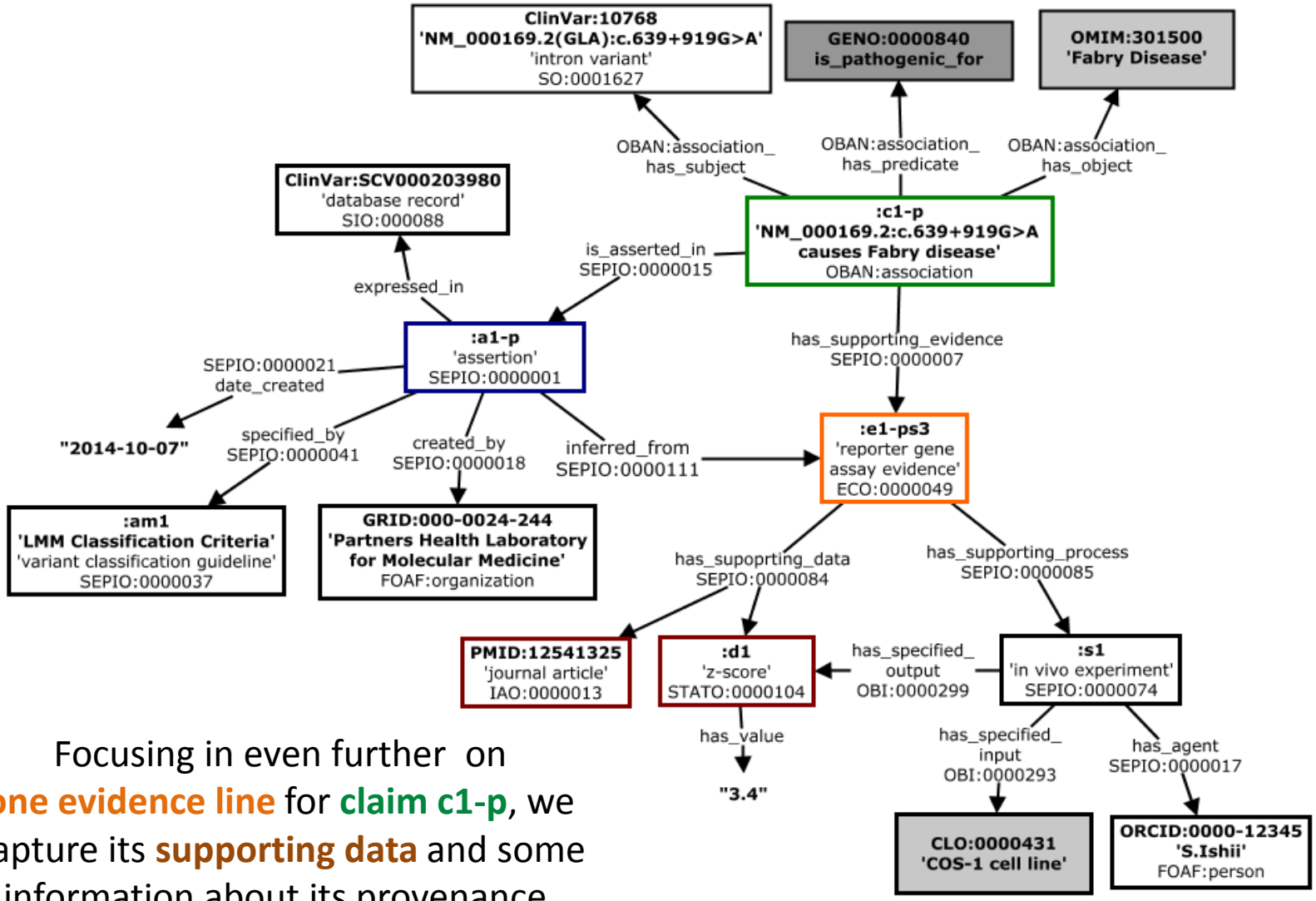
Focusing in on **association c1-p**, we add **evidence lines** to the model. Four lines of **supporting evidence** - three used by **assertion a1-p**, and three used by **assertion a2-p**. There is also one line of **refuting evidence** (**:e5**)

The Data, Modeled



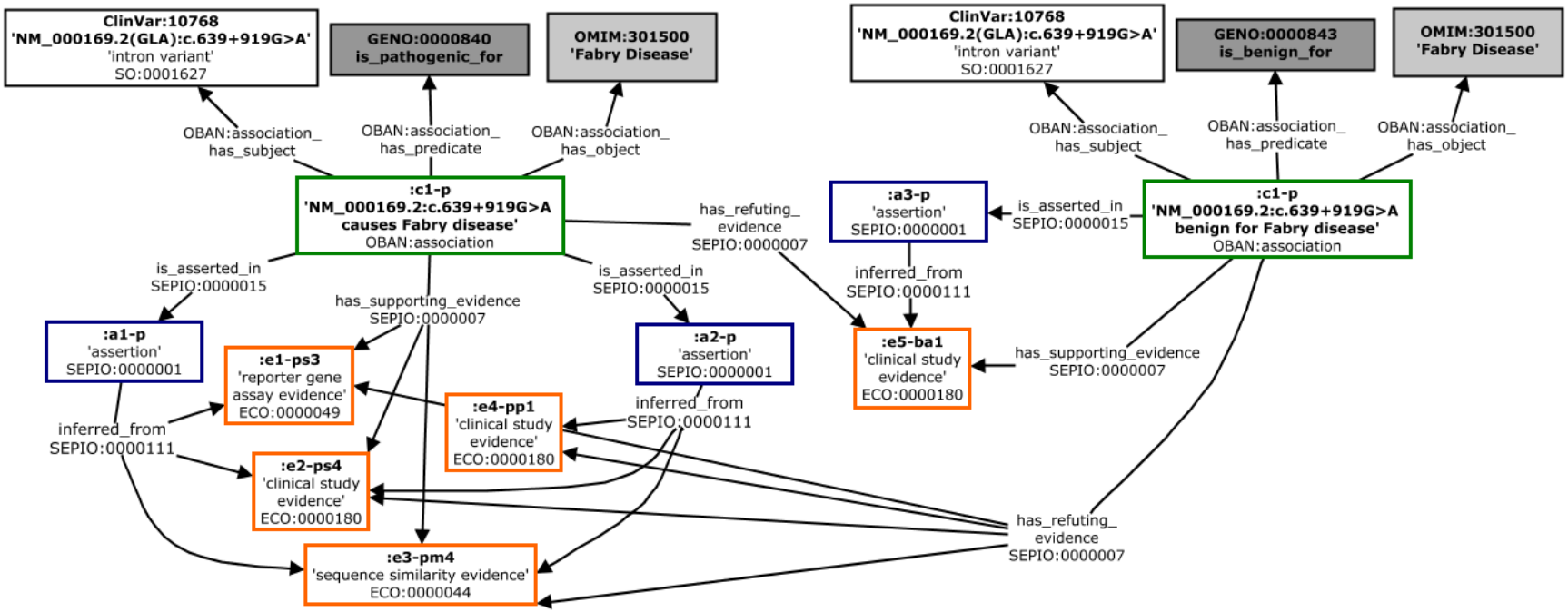
Focusing in further on **assertion a1-p** used by **association c1-p**, we can capture the **provenance** of this assertion including the **agent** who made it (**:ag1**), the **criteria** used (**:am1**), and the **date** it was asserted.

The Data, Modeled



Focusing in even further on **one evidence line** for **claim c1-p**, we capture its **supporting data** and some information about its provenance.

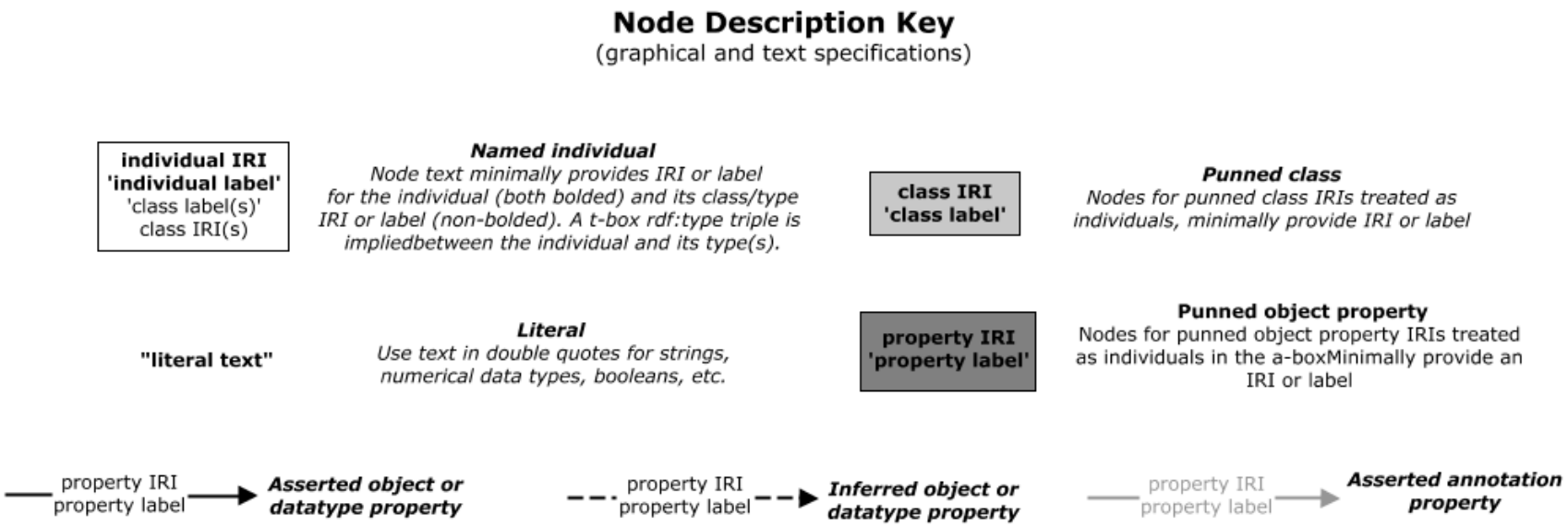
The Data, Modeled



A more complete picture of the relationships between the **two associations**, **three assertions**, and **five evidence lines** in the example scenario.

The Data, Modeled

Graphical Notation and Syntax



Key explaining the notation and syntax used in the diagrams for the Fabry Disease example.