IDENTIFICATION OF MISSING HIERARCHICAL RELATIONS IN THE VACCINE ONTOLOGY USING ACQUIRED TERM PAIRS

Warren Manuel | ICBO 2022

Joint work with: Rashmie Abeysinghe, Yongqun He, Cui Tao, and Licong Cui

UTHealth Houston

School of Biomedical Informatics

Disclosure

Outline

- Vaccine Ontology (VO)
- Why Quality Assurance for VO is needed?
- Related Work
- Methods
  - Concept Representation
  - Construction of Acquired Term Pairs
  - Detecting potential missing is-a relations
- Results
- Evaluations
- Conclusions and Future Directions

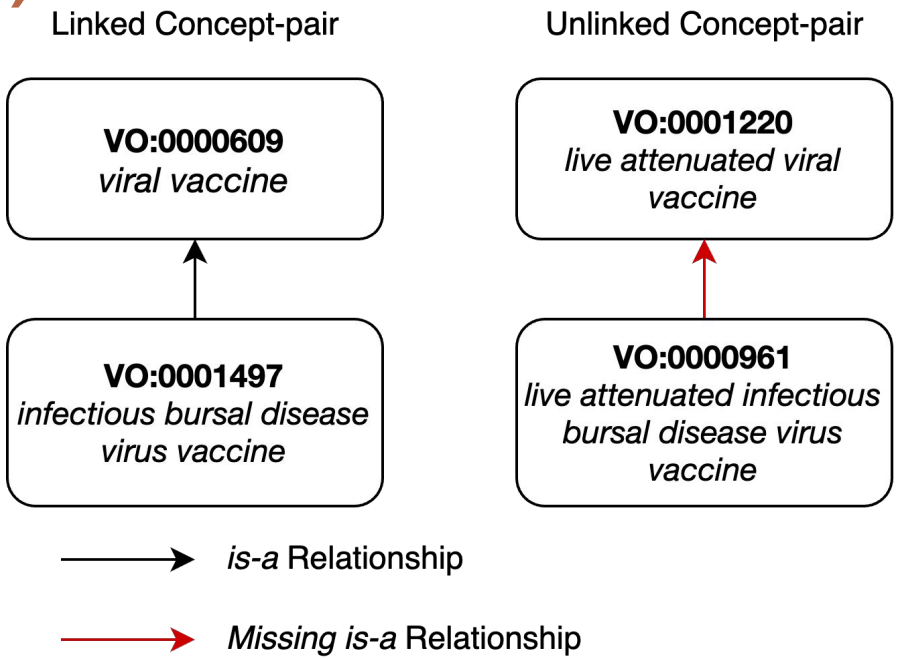
Vaccine Ontology[1] (VO)

- Ontology in the domain of vaccine and vaccination
- Developed as a community-based ontology
- Focus on vaccine categorization, components, quality and vaccine- induced host responses
- Contains over 6800 concepts

[1]He Y, Cowell L, Diehl AD, Mobley H, Peters B, Ruttenberg A, et al.VO: vaccine ontology. In: The 1st International Conference on Biomedical Ontology (ICBO-2009). Buffalo: ICBO; 2009. p. 24–6.

Why Quality Assurance (QA) for VO?

- VO needs to be updated with rapidly evolving biomedical knowledge
- May suffer from incomplete knowledge and inconsistent modelling
- VO has many downstream applications
  - Vaccine Data Integration
  - Literature Mining Systems
  - Used by 15 ontologies[1]
- QA is essential in avoiding error propagation from source
- Manual review: Requires domain experts and time consuming
- Need for [semi] automated methods for QA



A validated missing *is-a* relation identified in the Vaccine Ontology

Jackson, R., Matentzoglu, N., Overton, J. A., Vita, R., Balhoff, J. P., Buttigieg, P. L., Carbon, S., Courtot, M., Diehl, A. D., Dooley, D. M., Duncan, W. D., Harris, N. L., Haendel, M. A., Lewis, S. E., Natale, D. A., Osumi-Sutherland, D., Ruttenberg, A., Schriml, L. M., Smith, B., … Peters, B. (2021). OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies. Database, 2021(0), 1–9. https://doi.org/10.1093/DATABASE/BAAB069

Why Quality Assurance (QA) for VO?

Info: Which ontologies use it?

- ado
- apollo_sv
- cido
- cto
- eupath
- genepio
- oae
- ogsf
- ohpi
- one
- ons
- ontoneo
- opmi
- ovae
- scdo

Image courtesy of: OBO Dashboard (http://dashboard.obofoundry.org/dashboard/vo/dashboard.html)

Related Work

- Abstraction networks for NCIt[1] (where similar concepts are summarized providing a higher-level view of the ontology content)
- Non-Lattice Subgraphs in Gene Ontology[2] (where graph fragments violating the lattice property are extracted)

[1]Min H, Perl Y, Chen Y, Halper M, Geller J, Wang Y. Auditing as part of the terminology design life cycle. J Am Med Inform Assoc. 2006; 13(6):676–90.
[2]Cui L, Zhu W, Tao S, Case JT, Bodenreider O, Zhang GQ. Mining non-lattice subgraphs for detecting missing hierarchical relations and concepts in SNOMED CT. Journal of the American Medical Informatics Association. 2017 Jul 1;24(4):788-98.

Methods

- Concept Representation
- Extract list of linked and unlinked concept pairs
- Generate ATP for each concept pair (linked and unlinked)
  - Representation of concepts as lexical features
- Discovery of potential missing *is-a* relations
- Post processing of results

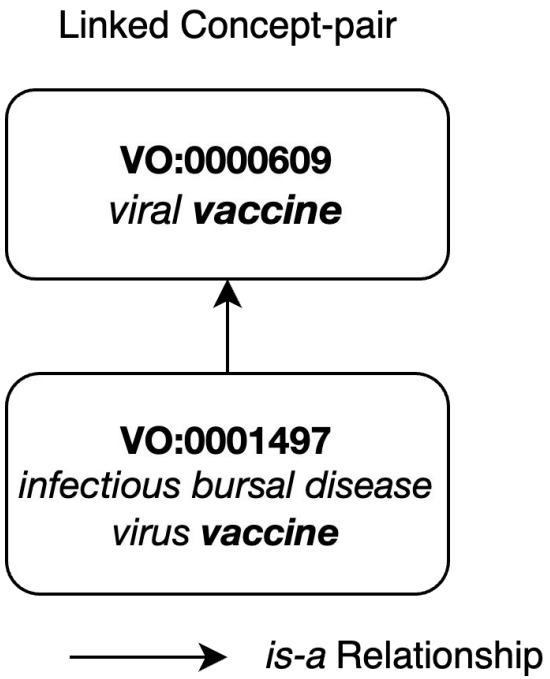
Concept Representation

- Each concept is represented as a set of its lexical features (words)
- C_1 = infectious bursal disease virus vaccine
- $F(C_1)$ = {infectious, bursal, disease, virus, vaccine}

- C_2 = viral vaccine
- $F(C_1)$ = {viral, vaccine}

Extracting Linked Concept Pairs

- A concept pair C and A would form a linked concept pair L(C,A) if:
  - A is an ancestor of C; and
  - C and A have at least a single common lexical feature



Extracting Unlinked Concept Pairs

- A concept pair C and A would form a linked concept pair U(C,D) if C and D are:
  - C ≠ D
  - Not ancestors of each other
  - Have at least a single common lexical feature
  - Belong to the same ontology
  - Fall within the same subhierarchy out of the 19 different subhierarchies under concept “material entity” (BFO:0000040) of VO

Unlinked Concept-pair

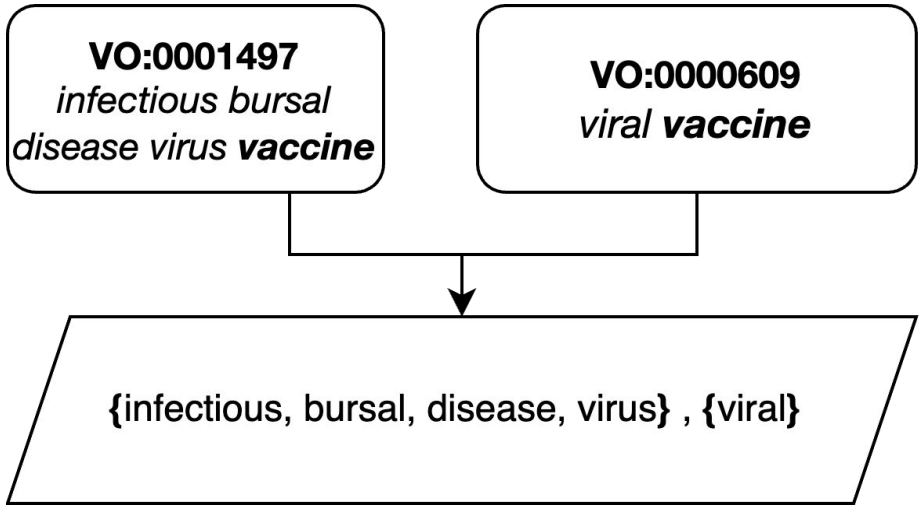
VO:0001220
live attenuated viral vaccine

VO:0000961
live attenuated infectious bursal disease virus vaccine

Generation of Acquired Term Pairs (ATP)

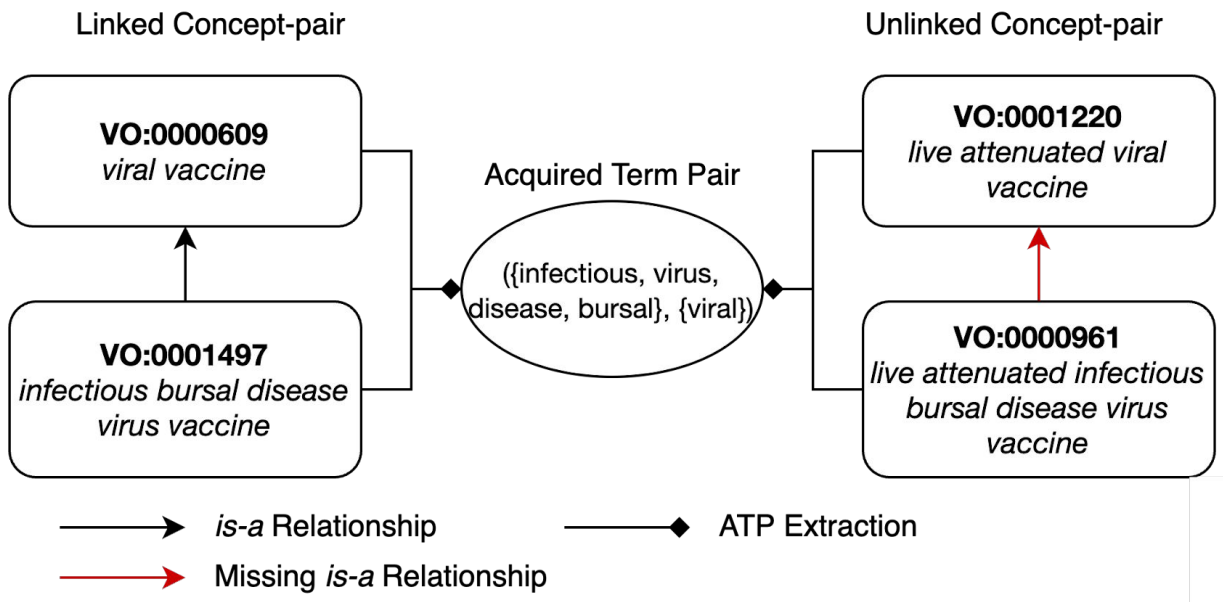
An ATP is a representation of the unique lexical features of each concept in a term pair

- ATP(C1, C2) = (F(C1) − F(C2), F(C2) − F(C1))



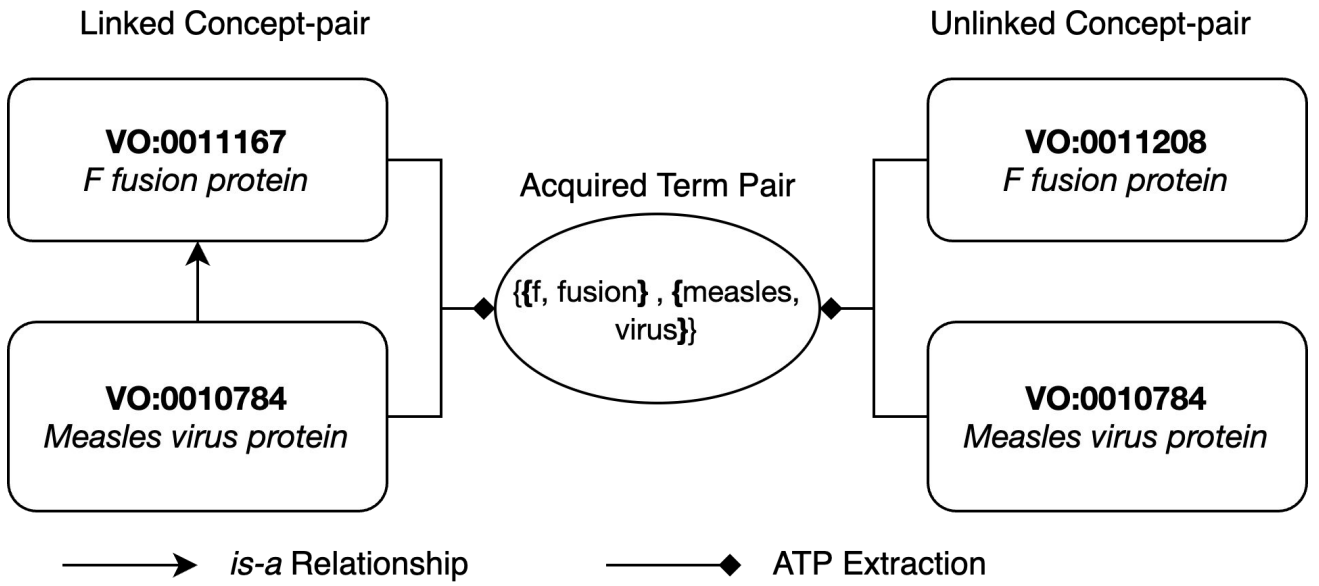
Discovery of Potentially Missing *is-a* Relations

- When a linked concept-pair and a unliked concept-pair generates the same ATP: we suggest a missing *is-a* between the unlinked concept-pair



Post Processing of Results

- Sets of concept pairs containing identical concepts with different identifiers were filtered out



Results

Total Concepts	6,883
Linked concept-pairs	62,538
Unlinked concept-pairs	17,301,802

- 232 potential missing is-a relations were identified in 1.1.192 version of VO

Evaluation

- Evaluation sample: 70 potential missing is-a relations
- 65 / 70 valid missing *is-a* relations
- Overall Precision: 92.86%

Examples of Valid Missing *is-a* Relations

Descendant	**Ancestor**
inactivated acellular pertussis vaccine (VO:0003196)	inactivated vaccine (VO:0000315)
COVID-19 recombinant vector vaccine (VO:0005199)	recombinant viral vector vaccine (VO:0005331)
Acellular Pertussis Vaccine (VO:0003389)	acellular vaccine (VO:0000756)
Human papillomavirus protein (VO:0010786)	human protein (VO:0000516)

False Positives

Descendant	Ancestor
smallpox vaccine (VO:0004613)	Smallpox virus vaccine (VO:0000651)
Brucella canis (NCBITaxon:36855)	Canis (NCBITaxon:9611)
Corynebacterium pseudotuberculosis (NCBITaxon:1719)	Corynebacterium diphtheriae (NCBITaxon:1717)
Varicella-Zoster Virus Vaccine Live (Oka-Merck) strain 29800 UNT/ML (VO:0003279)	Varicella-Zoster Virus Vaccine Live (Oka-Merck) strain Injection (VO:0003274)
toxoid vaccine (VO:0000561)	toxoid (VO:0001252)

Conclusions and Future Directions

- We applied a method to identify missing hierarchical relations in VO based on unique lexical characteristics of it's concepts
- Limited to concept pairs with common words
- Addition of lexical metadata (synonyms) and ancestral lexical features
- Performing additional normalizing strategies (lemmatization and synonym identification)

Acknowledgement

- This work was supported by
  - National Science Foundation through grant 2047001
  - National Institutes of Health National Library of Medicine through grant R02LM013335
  - National Institute of Allergy and Infectious Diseases through grant UH2AI132931
  - National Institute of Neurological Disorders and Stroke through grant R01NS116287

Thank you !