Studying Ontologies Using Mathematical Techniques

Keith Allen

kallen20@oswego.edu

Oswego State University
Department of Mathematics

October 14, 2023

OSWEGO
STATE UNIVERSITY OF NEW YORK

Outline

OSWEGO
STATE UNIVERSITY OF NEW YORK

What is an Ontology?

General Ontology:
Ontology is focused on finding a taxonomy of entities and their relations. These classifications and relations are universal and aim to represent reality in a consistent repository.

Upper Level

- Ontologies concerned with the most general classification.
- Classes such as Object or Process.

Domain Level

- Built to classify a specific domain i.e. Beet Ontology and Epilepsy Ontology
- Classes such as Condiment in the Food Ontology

OSWEGO
STATE UNIVERSITY OF NEW YORK

Importance of Ontology

General Ontology

- Increases reusability and sharing of information.
- Semantic labeling allows for easier computing over, and better insights from, large data sets.

Top level ontology

- Siloed domain ontologies
- BFO (Basic Formal Ontology): Top level ontology that was made an ISO standard in 2021 Used in 400+ ontologies and by 100+ organizations.

OSWEGO
STATE UNIVERSITY OF NEW YORK

Issues with Building Ontologies

Problem: Creating BFO compliant ontologies is a tedious and error prone process.

Solution: Have domain experts work with an automated system that streamlines the ontology building process by using a dialogue system and analogical reasoning.

OSWEGO
STATE UNIVERSITY OF NEW YORK

The DBOL

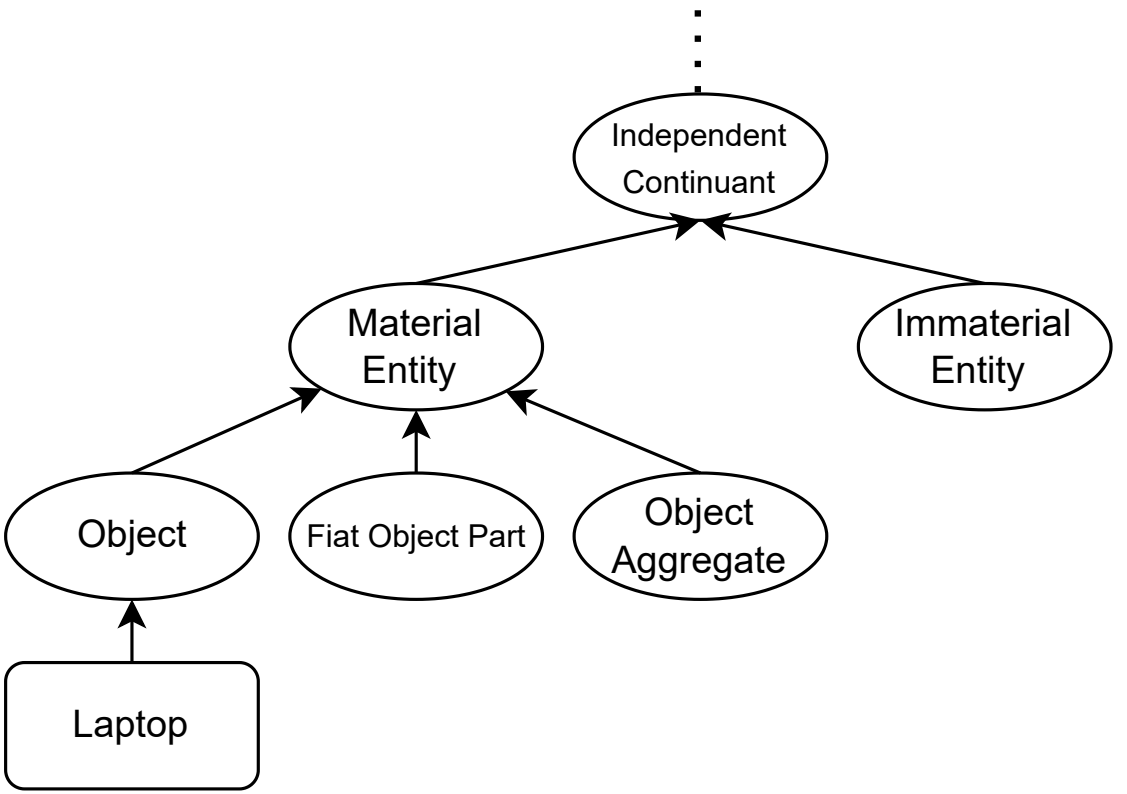
Current Work: Focus on a set of beginner rules that can correctly classify common terms.

At the end of the summer we had interns assign terms from a list of 1000 common English words.

What does it mean to assign something incorrectly?

Incorrect Assignment Example

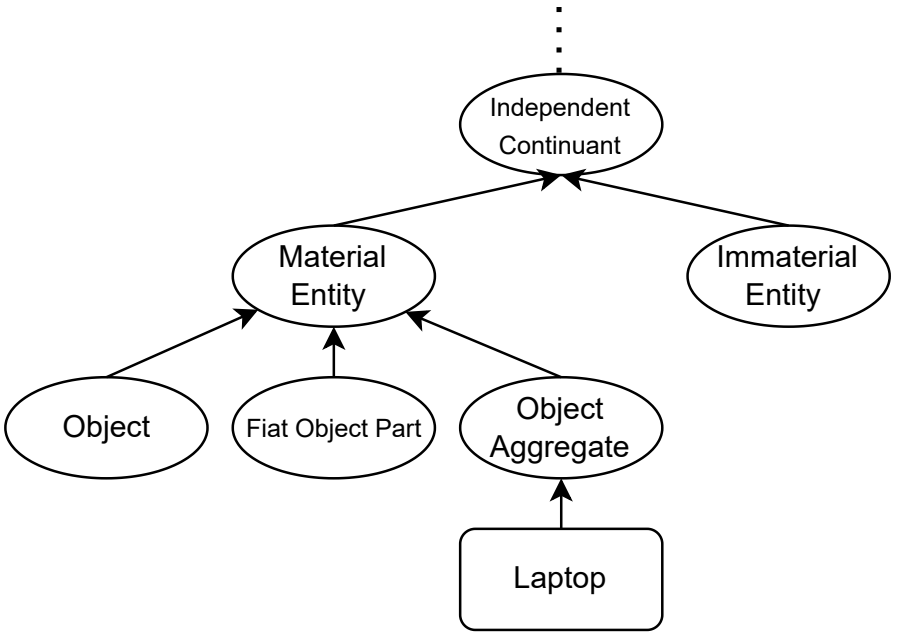
Correct assignment of the term `Laptop` within a section of the BFO hierarchy.



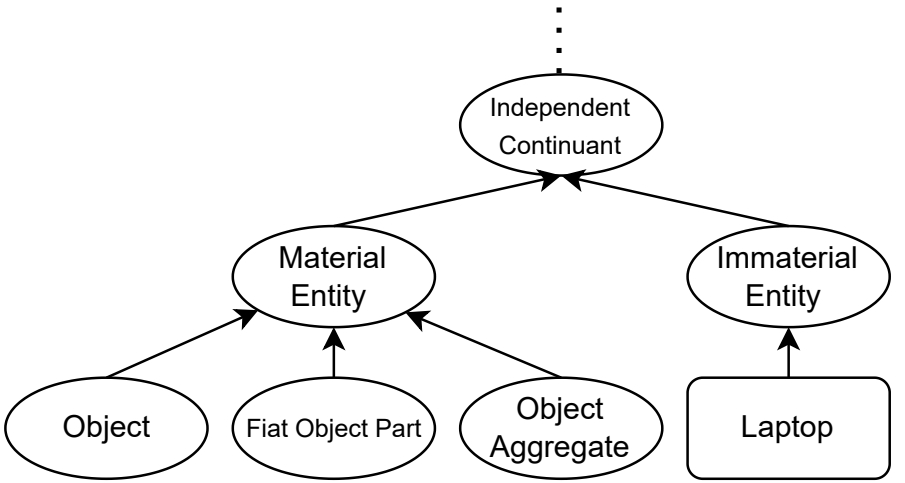
Incorrect Assignment Example

cont.

Close



Far



OSWEGO
STATE UNIVERSITY OF NEW YORK

Constraints

When looking at our examples we have

- Same terms
- Different relations between terms
- Same type of relation

But, if we think about the underlying graph we have

- Same vertices
- Different edges
- Same weight of all edges

OSWEGO
STATE UNIVERSITY OF NEW YORK

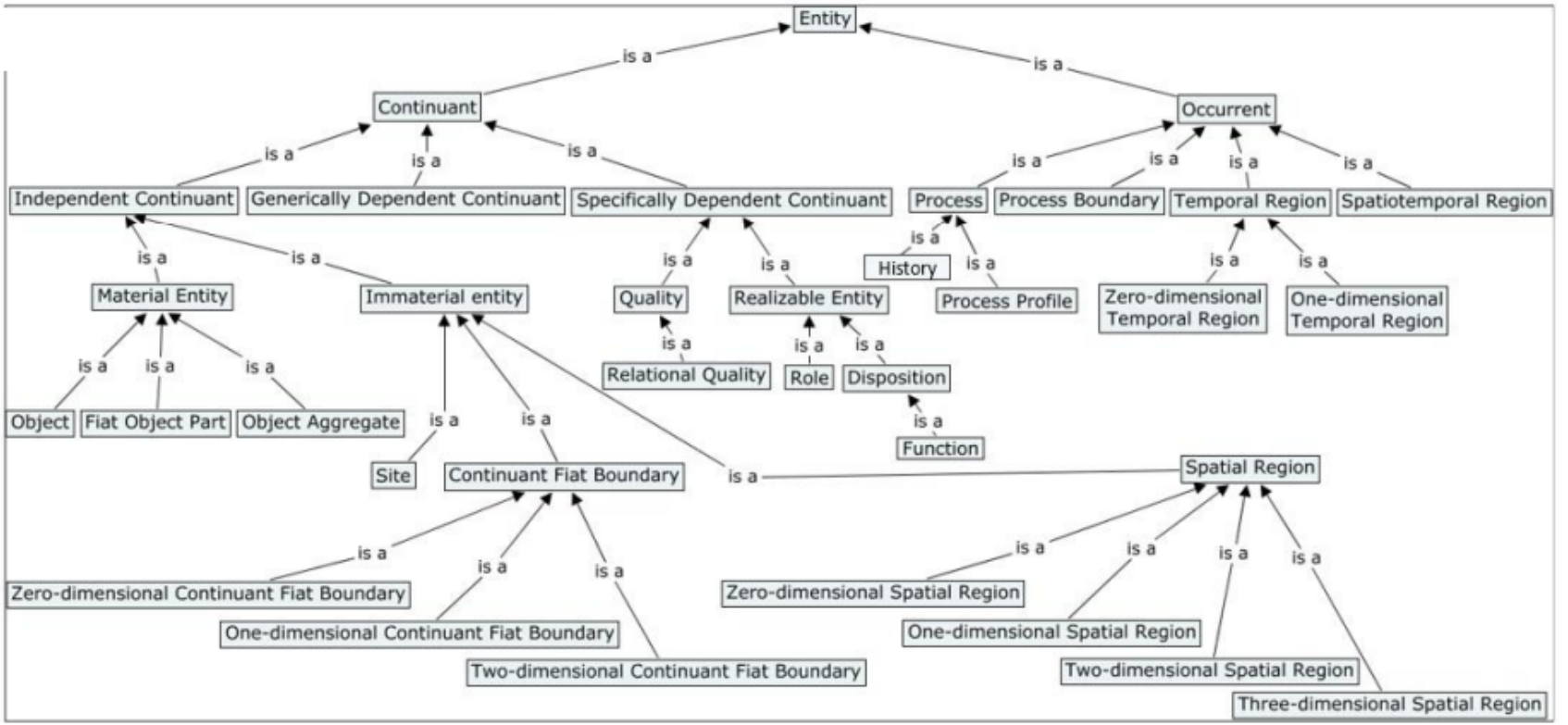
What is a Directed Acyclic Graph?

(DAG) Directed Graph defn: A directed graph D is a non-empty finite set $V(D)$ of **vertices** and a finite set $E(D)$ of ordered pairs of distinct vertices called **edges**. Edges have a tail and a head, meaning that the edge (u, v) is different from the edge (v, u).

Directed Acyclic Graph defn: A directed graph without a cycle.

Cycle defn: A cycle is a non-trivial path from vertex v to v.

Acyclic Upper Level Ontology



BFO 2.0 Hierarchy from BFO 2.0 Specification and User Manual

OSWEGO
STATE UNIVERSITY OF NEW YORK

Guarantees at Other Levels

- BFO appears to be a tree, but domain ontologies are not necessarily trees.
- Want to extend solutions used to measure top level ontologies to domain level ontologies.
- BFO book states that ontologies are directed rooted graphs.

Intro to Katz

Created by Leo Katz in 1953 to measure the influence of one actor within a graph.

Katz similarity measure captures the number, and lengths, of paths between the same vertices in different graphs.

The current implementation is based off of the work of Nayak et al. in which they compared DAG abstractions of knowledge graphs.

Why use Katz?

Pros:

- Accounts for the importance of the direct parent child relation within a graph.
- Considers least common ancestor.
- Factor in for the asymmetric relations and multiple paths

Cons:

- Computationally expensive since it requires every vertex to be compared to every vertex.

OSWEGO
STATE UNIVERSITY OF NEW YORK

Katz Similarity

Katz Similarity:

$$KS(u, v) = \sum_{l} t\alpha^{l}$$

where t is the number of paths length l from u to v and $0 < \alpha < 1$. Note: α is a tunable variable typically set to 0.05.

Ex) Two vertices u and v that have one path length 1 and three paths length 2 would have a Katz similarity equal to $\alpha + 3\alpha^2$.

OSWEGO
STATE UNIVERSITY OF NEW YORK

Katz Similarity Vector

The Katz Similarity Vector (KSV): A graph can be represented by its KSV where the n^{th} element of the vector is the Katz Similarity of the n^{th} pair of vertices in $V \times V$.

Katz Graph Similarity

Katz Graph Similarity: Given two Katz similarity vectors KSV_1 and KSV_2 derived from graphs G_1 and G_2, with the same vertex set. Then,

$$KGS(G_1, G_2) = \frac{2}{1 + \exp(\gamma ||KSV_1(:, i) - KSV_2(:, i)||_p)}$$

where $||.||_p$ is the L_p-norm of the i^{th} vector differences and $\gamma > 0$ is another tunable parameter.

Thus we know $0 < KGS(G_1, G_2) \leq 1$ where 1 indicates that G_1 and G_2 are identical graphs.

Katz Similarity Speed Up

We can harness properties of DAGs to speed up the process.

1. Create topological ordering of the vertices.
2. Start computing from the lowest ordering and after each level "remove" these vertices and edges from the graph.

$$KS'(u,v) = \alpha \times \left(\sum_{p \in parents(v)} KS'(u,p) \right) + \alpha \times \Upsilon(u \to v)$$

where $\Upsilon(u \to v) = 1$ when there is an edge from u to v and 0 otherwise.

Implementing in Java

Nayak et al. implemented these ideas in C. However this project is built in Java and this measurement is part of a larger set of evaluation programs.

Instead of removing nodes from the graph we can instead store previously computed Katz similarities in a HashMap, making retrieval quick and easy.

$KS'(u, v)$ in Java

```
public double calculateKS(Vertex u, Vertex v, double alpha){
    String key = u.toString() + v.toString();
    // base case
    if(u.toString().equals(v.toString())) { return 0; }
    // we have already calculated this value and can just grab it
    else if(katzSimilarities.containsKey(key)) { return katzSimilarities.get(key); }
    // otherwise
    else {
        double parentsSum = 0;
        // get the sum of the values of all the parents
        for ( Vertex p : parents.get(v) ) {
            parentsSum =+ calculateKS(u, p, alpha);
        }
        // calculate the final katz similarity and insert in the map
        double katzSim = alpha * parentsSum;
        katzSimilarities.put(key, katzSim);
        return katzSim;
    }
}
```

Approximation of the KGS

Nayak et al. continue on defining an approximation for the KGS on graphs so large that even the quicker version of calculating the KSV is too time intensive.

Since domain ontologies can contain thousands of terms this is a great place to focus next.

Review

- Ontologies are highly structured bodies of knowledge used to semantically label data.
- Ontologies can be abstracted into DAGs
- Katz similarity can incorporate the crucial features of an ontology and can be implemented in a way that is computationally feasible even over large graphs.

OSWEGO
STATE UNIVERSITY OF NEW YORK

Acknowledgements

Ontology

- Dr. Shane Babcock, Niagara University
- Dr. Roman Ilin, AFRL
- Dr. Daniel Schlegel, SUNY Oswego

Similarity Metrics

- Dr. Elizabeth Wilcox, SUNY Oswego

OSWEGO
STATE UNIVERSITY OF NEW YORK

References


Guruprasad Nayak, Deepak Ajwani (2001)
Automated assessment of knowledge hierarchy evolution: comparing directed acyclic graphs
Information Retrieval Journal 22(3-4) 256-284.


Jørgen Bang-Jensen, Gregory Z. Gutin (2001)
sDigraphs: Theory, Algorithms and Applications
Springer-Verlag London Second Edition.


Leo Katz (1953)
A New Status Index Derived from Sociometric Analysis
Pysometrika 18(1) 39-43.


Robert Arp, Barry Smith, Andrew D. Spear (2015)
Building Ontologies with Basic Formal Ontology
MIT Press Cambridge, Massachusetts.

OSWEGO
STATE UNIVERSITY OF NEW YORK

Studying Ontologies Using Mathematical Techniques

Keith Allen

kallen20@oswego.edu

Oswego State University
Department of Mathematics

October 14, 2023

OSWEGO
STATE UNIVERSITY OF NEW YORK