Retaining and Recovering Context for ML:

Taxonomies and KGs

Presented @ Connected Data London, April 2021

By Ashleigh faith

Agenda

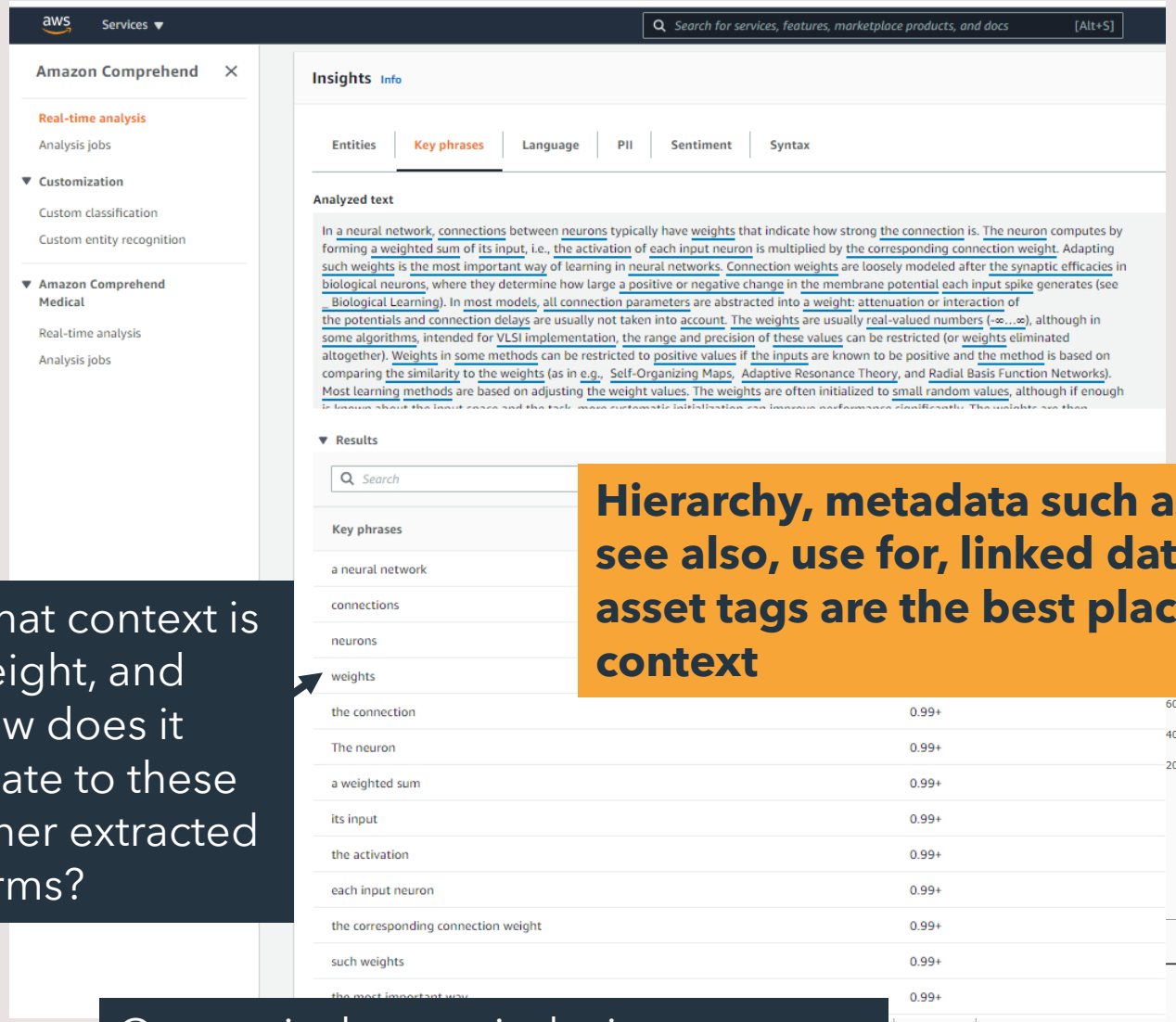
- Why context matters in ML, and why taxonomies and KGs are useful for context
- Where this fits within the architecture/pipeline
- External data, where to find it, when to use it in your own data, how to determine its quality- *Hands-on working segment*

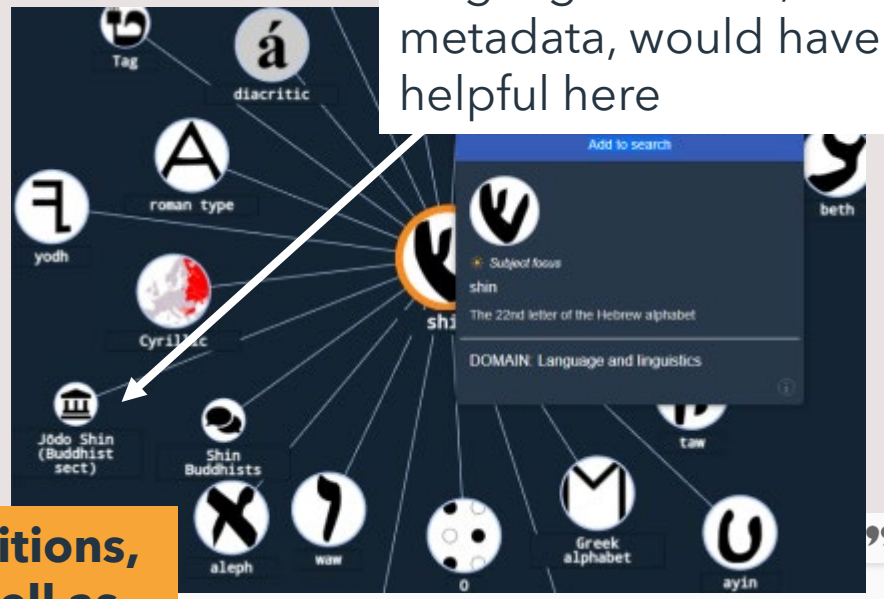
10-15 min. break

- Internal data, what if it has poor or little context, how to marry it with external data and add context back- *Hands-on working segment*
- How to pitch a project to your manager/c-suite
- Open questions

Why does context matter?

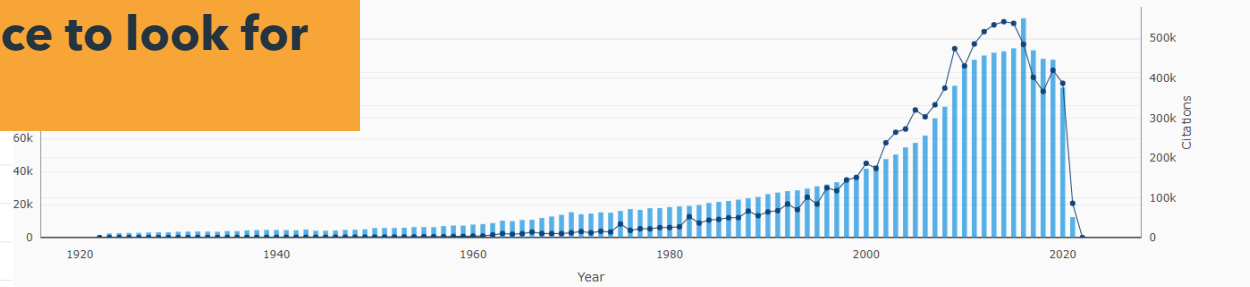
Religious context vs written language context, or more metadata, would have been helpful here



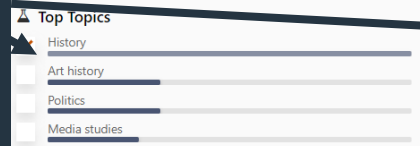


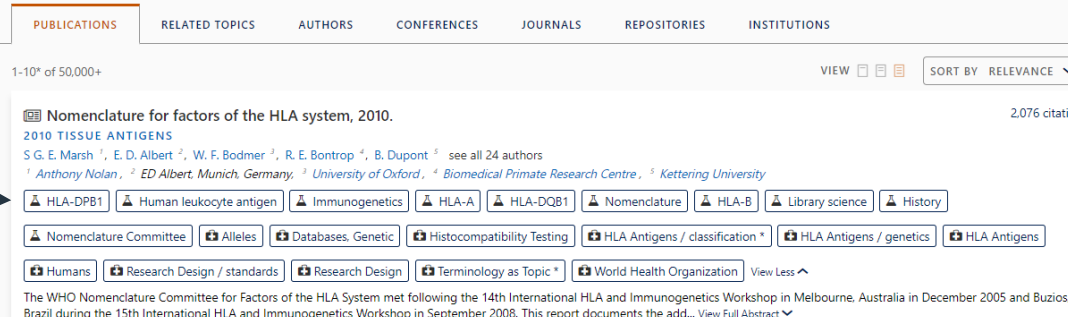
Hierarchy, metadata such as definitions, see also, use for, linked data, as well as asset tags are the best place to look for context

What context is weight, and how does it relate to these other extracted terms?



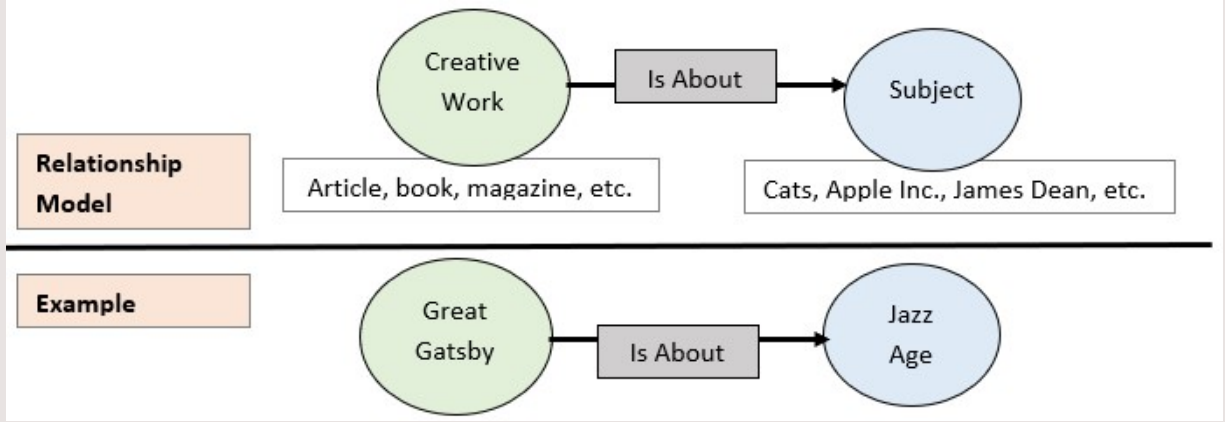
Categorical + overindexing cause this medical paper to be the #1 paper in History last year



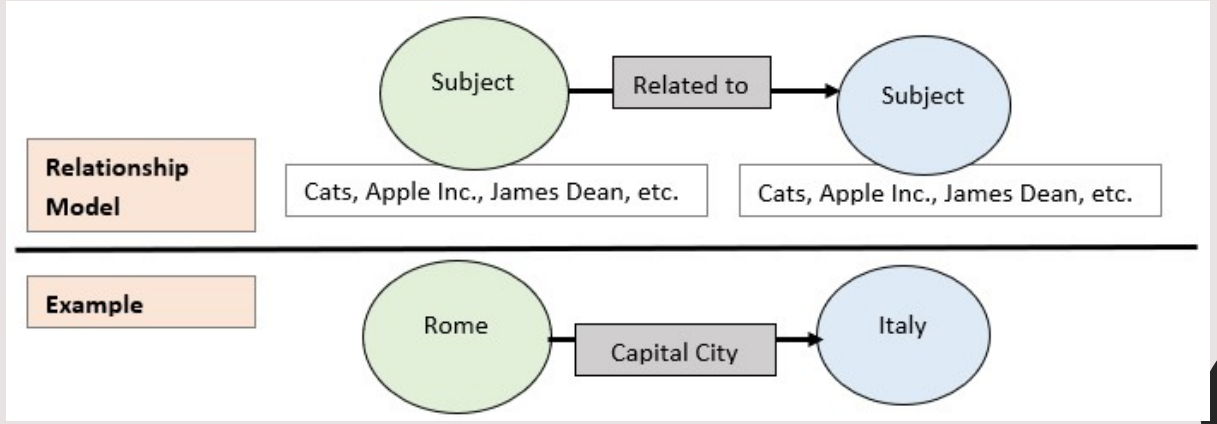


What is the difference between a taxonomy and KG?

- **Taxonomies** are often created by information or librarian professionals for tagging or organizing assets, either documents, media, products, tagging the explicit ABOUTNESS of a thing.

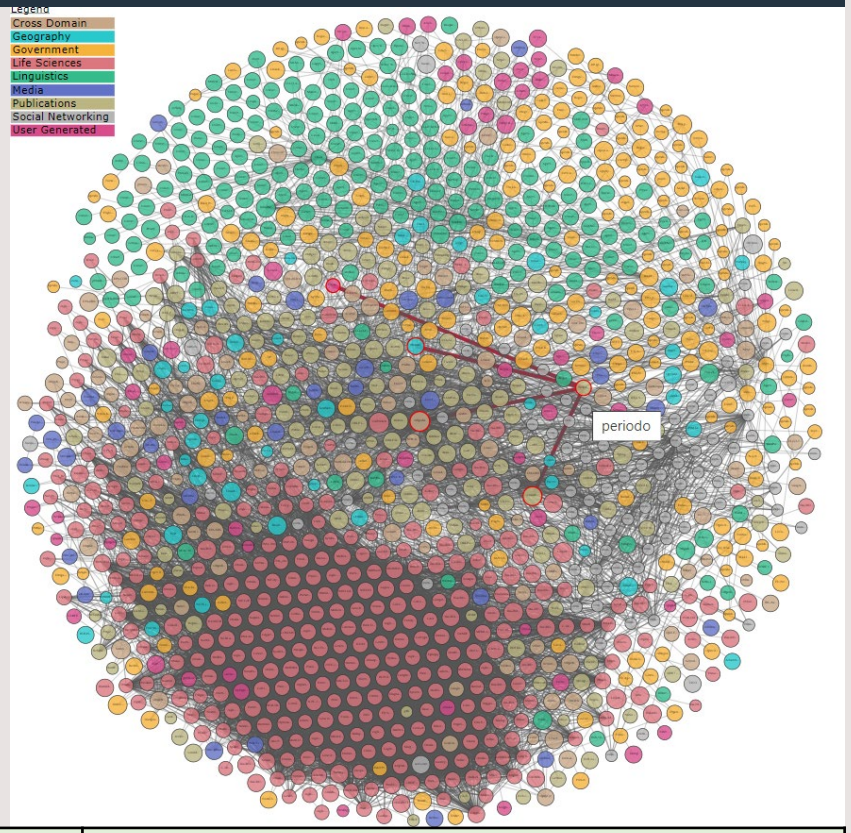


- **Knowledge graphs**, in the SEO and embedded graph use cases at least, take the asset organization a step further by attempting to capture the IMPLICIT or TACIT knowledge of people, whether through structured or unstructured means, usually through a person's world orientation, such as how one thing is related to another thing.



Why not just use them as-is?

- Most linked data sources are made for a particular use case, or are very general
- Many have errors, misalignments, or offensive content (as shown by the recent MIT study where the most popular ML data sources had up to a 3.5% error rate, with one very popular source having a 54% error rate)
- Linked data still needs merging/de-duping; as well as translation into your systems and schema
- Linked data has deferent levels of trust, knowing the source of the information and the frequency of updates and by whom, is essential
- Many linked data sources in the library and scholarly fields are not calibrated for ML- what's a 1xx vs 9xx, controlled vs keyword, linked vs mapped?!?!



Subject heading before	Common user search	Subject heading after
Dog -- food -- recipes	"Recipes for dog food"	Dog food recipes
Computer aided design in civil engineering	"CAD AND civil engineering"	Computer aided design Civil engineering
	"civil engineering 3D drawing tools"	
Stocks (Finance) -- Prices -- Databases	"trends in stock price data"	Stock price databases

Common issues to watch for, and why do they exist at all?

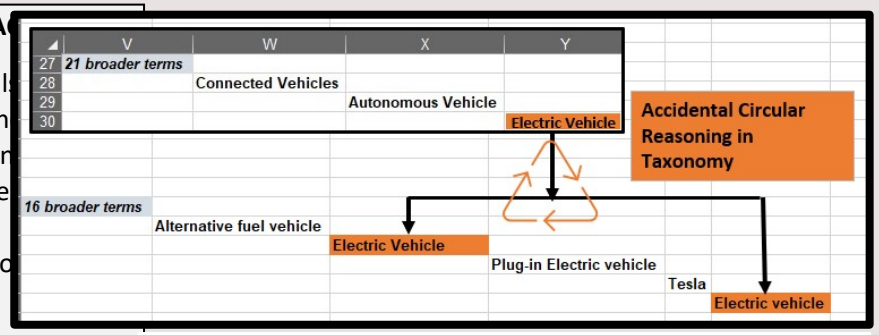
- Outdated terminology
- Fuzzy BT/NT relations
- Poor definitions/scope notes
- Bad form *(such as reverse-order named entities)*
- Overindexing
- Circular logic
- Categories vs specific tags
- Unknown/unfriendly context/schema
- Overly unique terminology

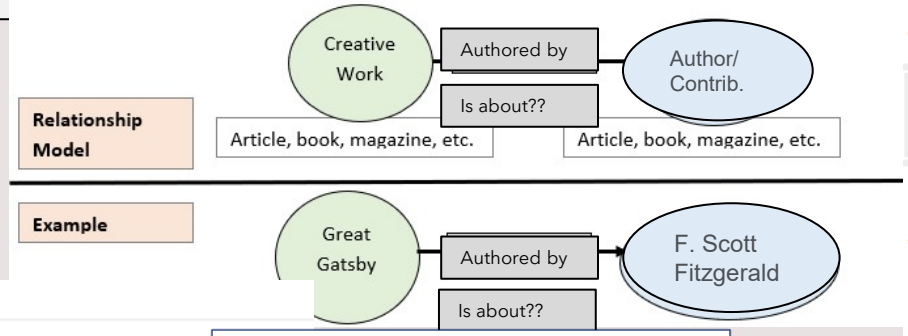
How do we test which LOD/vocabs to use and if they are introducing unwanted issues?

PLASTICS AND THE ENVIRONMENT BY SAM WELLS DATE: 03/2001 PUBLISHER A

Plastics have provided more change and technical progress than most other material decade alone. It has come at a cost, however. With over 100 billion tons of waste, an of plastics being recycled, the environmental factors associated with the introduction is dire, as well as the making of the plastic itself from fossil fuels. This book introduce field's leading researchers, covering how plastics affect the environment and how environmental factors affect plastics. The relative benefits of recycling, resource reco energy recovery are also discussed in detail.

TAGS: Plastics; Environment; Wells, Sam; Waste; Recycled Plastic; environmental factors; environmental factors – plastic; fossil fuels; plastics – environmental studies; recycling; resource recovery – plastics; energy recovery – plastics





Transportation Research Thesaurus

Term Details

TRT Keywords:

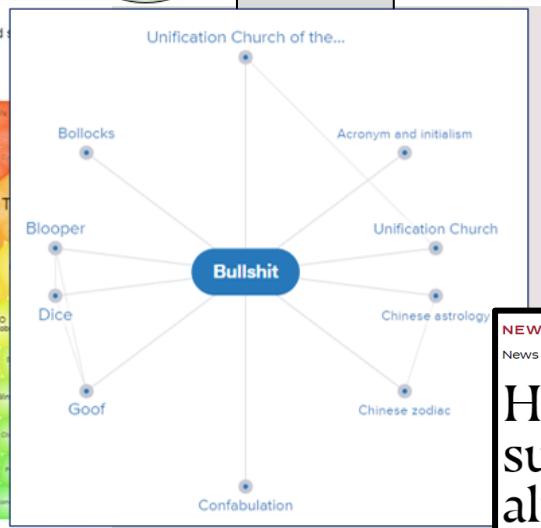
Search

Display Hierarchical | Alphabetical | KWOC | KWIC

Top Terms > Organizations > **Businesses**

Businesses (Nb)

Scope Note
Organizations offering a service rather than a product
Definition
Organizations offering a service rather than a product. (Source: TRT SN)
Use For
Service industries
Broader Term
Organizations (N)
Narrower Terms
Carriers (Nbb)
Freight transportation support businesses (Nbc)
Passenger transportation support businesses (Nbd)



NEWS +

News from Harvard schools, offices, and affiliates

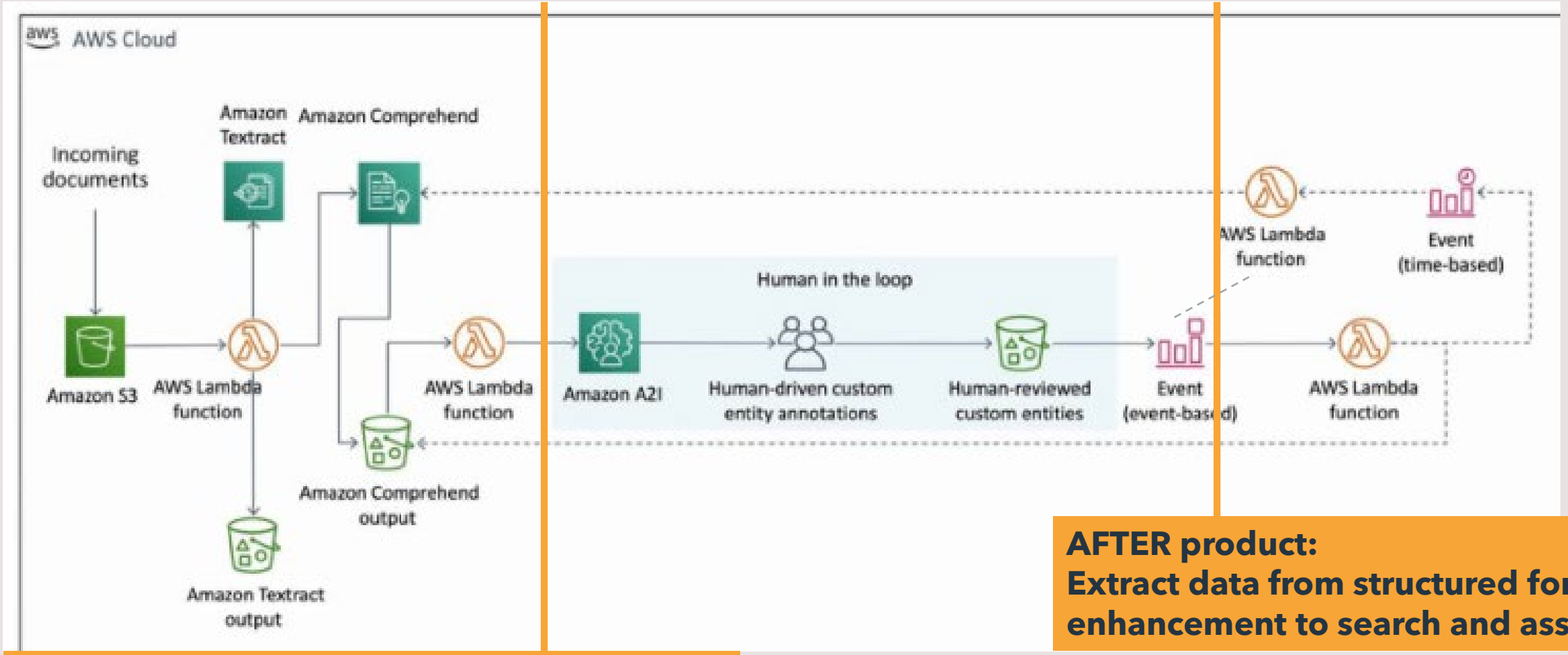
Harvard Library ends use of subject heading 'illegal alien'

What to look for- Go assess your favorite LOD/dataset

1. Provenance
2. UIDs/APIs
3. Pre- vs Post coordination
4. Definitions not scope notes, and with citation/links if possible
5. How are sensitive topics/historical data handled
6. How many ingoing, outgoing links
7. Who maintains it/data governance
8. Root source, social media, scholarly articles, gov. data..etc.
9. General/specific domain
10. Schema/framework

Try VIAF or Bioportal if you don't have a dataset in mind

Where to place the ML + knowledge models in the pipeline?



AFTER product:
Extract data from structured for ML enhancement to search and assets

BEFORE product:
Extract data from unstructured for auto-tags/auto vocab w/HITL for ML training

What to do if your OWN data is missing context?

NetworkX
Network Analysis in Python

binarytree 6.3.0

treeswift package

TreeSwift is a Python library for parsing, manipulating, and iterating over (rooted) tree structures. TreeSwift places an emphasis on speed.

Any Python Tree Data

pypi package 2.8.0 | pypi downloads 262k/month

Helps with form, and perhaps vectors, but how does this translate to open world logic and user cognitive models?

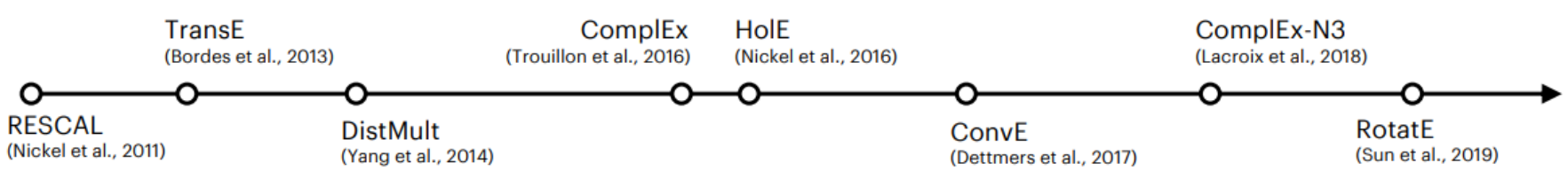
```python
from binarytree import Node

root = Node(1)
root.left = Node(2)
root.right = Node(3)
root.left.left = Node(4)
root.left.right = Node(5)

print(root)
#
#         __1
#        /   \
#       2     3
#      / \
#     4   5
#
assert root.height == 2
assert root.is_balanced is True
assert root.is_bst is False
assert root.is_complete is True
assert root.is_max_heap is False
assert root.is_min_heap is True
assert root.is_perfect is False
assert root.is_strict is True
assert root.leaf_count == 3
assert root.max_leaf_depth == 2
assert root.max_node_value == 5
assert root.min_leaf_depth == 1
assert root.min_node_value == 1
assert root.size == 5
```

(Some) KGE models in recent published literature:

CC of Accenture Labs, 2020



Step	Process	Use	Methods to try
1	Node attributes	Characteristic similarity	Conditional random field, top2vec; NLTK
2	Structural features	Connections and types of connections	Hierarchical agglomerative clustering, doc2vec
3	Node embeddings	Surrounding clusters/nodes	Deepwalk, node2vec, BERT

Retain or recovering context via taxonomies/KGs

Add:

- Use existing LOD or context rich metadata/data sources
- Link your data sources
- Types/categories
- Confidence scores and data gov.
- ML results are only as good as the search/user deems them! A/B and user testing is a MUST

Retain:

- Register your UIDs if possible, or at least document their logic
- Provenance
- Test and test again every few months to make sure the data is still sound
- Train your indexers, DAMs, KMs, and ML engineers

Resources for dealing with vocabs for context:

- WebVowl to visualize connections: http://visualdataweb.de/webvowl/#iri=http://purl.org/dc/terms/
- OntOlogy Pitfall Scanner! To check for ontology errors: http://oops.linkeddata.es/response.jsp?uri=http://purl.org/dc/terms/
- RDF triple checker for triple errors: http://graphite.ecs.soton.ac.uk/checker/?uri=http://purl.org/dc/terms/
- Linked Open Vocabulary for vocabulary/schema/framework use: https://lov.linkeddata.es/dataset/lov
- Take only the median topics
- "blast" the combo tags
- Only take MVP
- Nearest neighbor or within 3 within tags, not the full text or tags on their own
- Triangulation, can more be found/assembled from 3 or more?
- Distinguish type from part via clustering/vectoring
- Balance natural vs preferred labels

Most common mapping models in LOD:

i.e. and how to manage this data in ML flows as a pre-processing step to pipelines

Type of analysis: Term-to-term *Source* term-to-*target* term

Vocabularies	Crosswalk (M2M)	Links (n^2-n)	Hub	Links (2n)
2		2		2
3		6		6
4		12		8
5		20		10

Mapping volume comparison between crosswalk and hub structures. Modified from Binding and Tudhope (2015).

Types of term mapping:

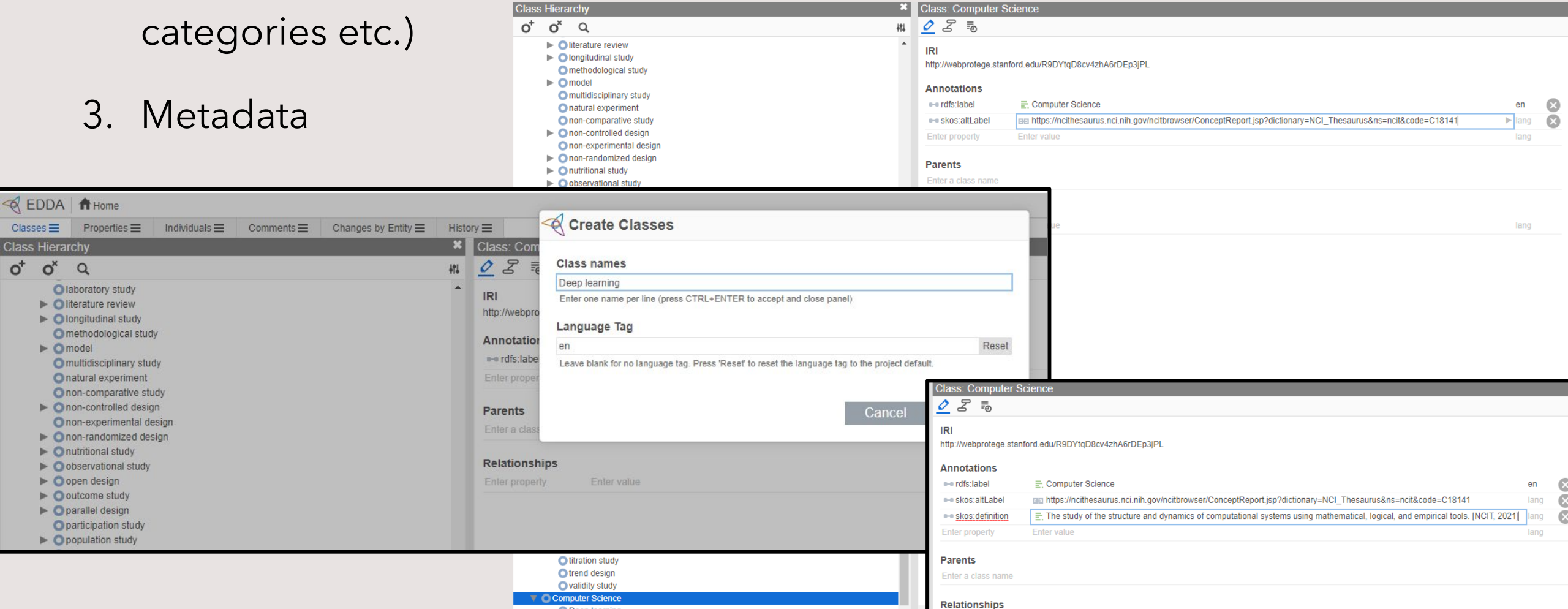
- *Partial* = syntax/semantic/contextual match
- *Exact* = matches exactly
- *Associative* = Matches as a related term
- No match = No match in target terminology

Considerations of term mapping:

- Hierarchy
- SMEs
- Automation/APIs
- Context/domain specificity

Jumping into Protege

1. Add 3 external sources from the web
2. Add 3 internal from the ML terminology project (which may need additional categories etc.)
3. Metadata



For more information, contact Ashleigh Faith

email afaith@ebsco.com

or find me on LinkedIn at /ashleighnfaith

or my Educational YouTube channel @Ashleigh Faith