# KNOWLEDGE REPRESENTATION AND EXTRACTION

## Domain-related Ontologies:

Philosophy:

* Basic Formal Ontology <https://basic-formal-ontology.org/>
* DOLCE <http://www.loa.istc.cnr.it/wp-content/uploads/2021/11/DOLCE_FOUST_2022.pdf>
* PhilO <http://philosurfical.open.ac.uk/ontology/doc/>
* <https://www.researchgate.net/publication/220607689_Foundations_of_an_ontology_of_philosophy#fullTextFileContent>
* SUMO (Suggested Upper Merged Ontology) (interoperability)
* Formal Ontology for Information Systems (FOIS)

Resource: <https://philindex.org/>

Biology:

* Gene Ontology (GO)
* Plant Ontology (PO)
* Chemical Entities of Biological Interest (ChEBI)
* Phenotypic Quality Ontology (PATO)

Ecology:

* Environment Ontology (ENVO)
* Plant Environment Ontology (PECO)
* Trait Ontology (TO)
* Ecological Metadata Language (EML)
* Ocean Biogeographic Information System (OBIS) Schema

Literature:

* <https://drops.dagstuhl.de/opus/volltexte/2013/4149/pdf/p076-damiano.pdf>
* <https://drops.dagstuhl.de/opus/volltexte/2016/6711/pdf/OASIcs-CMN-2016-10.pdf>

## Theoretical string figures

* Gilles Deleuze and Felix Guattari 🡪 concept of rhizomes
* Karen Bard 🡪 posthuman approach and entanglements between humans and nonhumans
* Lynn Margulis 🡪 symbiotic relationships in the evolution of life on Earth
* Microbiology
* Indigenous scholars and activists

Braiding Sweetgrass" by Robin Wall Kimmerer

The work of Vandana Shiva

The Idle No More movement

* Latour 🡪 actor-network theory hybrid networks (rejection of dualism human/nature), (companion species), symmetric anthropology (sympoiesis), importance of storytelling and narrative in shaping the understanding of the world: construction of facts (tentacular thinking) **!=** scale H-micro (individual relationships) VS L-macro, political orientation H-feminist and anticolonialist perspectives for more equitable and sustainable future VS L-neutral or apolitical approach (?), language H-playful and poetic, neologisms VS L- dense and technical

## Tools

* **Book NLP:** to annotate entire books with linguistic categories (parts of speech, dependencies, named entities, WordNet supersenses), using CoreNLP . Also provides character name clustering, quotations and their speakers, some coreference resolution
* **word2vec** is a word embedding (vectorial representation) of words: given a corpus and some parameters, it builds an embedding model of corpus words; the model can be queried for the best associations or analogies between words

## Competency Questions

1. What are good practices to make kin?
2. What are the stories of sympoiesis told in the book?
3. What are the events that take place in the timescape Anthropocene?
4. What are the stories that take place in the timescape Chthulucene?
5. What ecosystems are mentioned as endangered?
6. What are the neologisms Donna Haraway mints and what is their meaning?
7. From what perspective is the story of x narrated?
8. In what context did the sympoiesis between x, y, z? What was the consequence of that? Why was that necessary?
9. What are the troubles of the Chthulucene mentioned by Donna Haraway?
10. What are different types of SF?
11. Who is the most quoted person in the book?
12. What are the three most relevant theoretic domains in Haraway’s narration?
13. What is the book written by Anna Tsing that informed Haraway’s thinking
14. What are examples of Science-Art worldings for staying with the trouble?
15. Who practices tentacular thinking?
16. Who are the agents of story x?

## Expected answers

## To do:

1. Build website for presentation
2. Parse text Knowledge Extraction with **BookNLP**
3. Trovare entities, cooccurrence of entities in events
4. Events recognition
5. Speaker recognition
6. Parse with **FRED** to explore hehehe
7. Refine CQ + define Expected answers
8. Build Ontolgy in protégé
9. Create RDF to create graph + entity linking
10. Formalize CQ in SPARQL query to test the ontology and knowledge graph
11. Geolocalize events and entities in a map and connect them?

## Ontology design

Followed a model for philosophical ontologies where three main classes were identified: **Philosofers**, **Philosofical Entity**, **Field**.

**Philosophers** has been implemented as a subclass of **Thinker** which is an **Agent** in the Chthulucene Ontology. Cosìddetti “philosophers” have not more value in relation to other types of agents that we encounter in Chthulucene, 🡪 they also have developed Philosophical entities, can that be logically correct?

**Philosophical Entity** in the original model includes classes  **Concept, Theory, Argument, Method**. In Chthulucene, the lines between these categories are blurred: Methods are usually practical; Arguments are usually real or fictional examples of such practices; Practices usually include Conceptual entities to be described etc. For this reason we decided to implement only the two subclasses **Method** and **Concept.**

**Field** has been left out altogether, although we do organize knowledge about **Domain** of a **Concept**.

## Text processing

**Knowledge extraction** with **BookNLP.**

* Why Book NLP. Because it is specifically suitable and efficient with large texts.
* Event tagging is another key issue with longer documents and books. There are machine learning models that find events and you can easily cultivate a list of domain-specific events to improve a pipeline, but for BookNLP event is defined more broadly. From my experience, it is more based around key actions, rather than named events (as it is in named entity recognition). This has a tangential benefit known as triple extraction. In my opinion, it might be a bit better to view BookNLP events through this lens. Triple extraction is when we try and extract three pieces of information, such as (Actor, Action, Recipient) or (Actor, IS, Something). With these types of tuples, we can construct a knowledge tree about a corpus fairly easily. This a very challenging problem in NLP because triple extraction can be very domain-specific. BookNLP provides a great starting place for triple extraction with its events.
* Event tagging is another key issue with longer documents and books. There are machine learning models that find events and you can easily cultivate a list of domain-specific events to improve a pipeline, but for BookNLP event is defined more broadly. From my experience, it is more based around key actions, rather than named events (as it is in named entity recognition). This has a tangential benefit known as triple extraction. In my opinion, it might be a bit better to view BookNLP events through this lens. Triple extraction is when we try and extract three pieces of information, such as (Actor, Action, Recipient) or (Actor, IS, Something). With these types of tuples, we can construct a knowledge tree about a corpus fairly easily. This a very challenging problem in NLP because triple extraction can be very domain-specific. BookNLP provides a great starting place for triple extraction with its events.

1. Cleaning the text, taking out all notes and introductory parts, as well as chapter beginning quotes and page numbers + other noise digits

\n\s?[A-Z?a-z?]\s[a-z?\sa-z?A-Z?\sa-z?A-Z?\s]+\n\n

\s[0-9]{1}\.[0-9]{1}.\s

**Automated Ontology extraction:**   
In Ahmad and Gillam (2005) [2], a statistical approach is used to identify domainspecific keywords for nuclear physics, by calculating the difference between the relative frequencies of words in domain specific texts and their relative frequencies in a general corpus of language. Collocation analysis [112] is used to identify words that occur frequently within the same context and that are used to communicate specific meanings. This analysis is useful for discovering complex concepts within domain specific texts. The identified patterns can be used to produce collocational networks [136]. Is-a relationships are defined between words and their collocations and hierarchies are produced from these networks.