# Assignment 2

Joshua Cook

Georgia Tech

CS 6460 Education Technology

## Research Log

### Background

In the past week, I feel like I've really found my footing in this course and started to develop a solid understanding of what I would like to be working on. I continue to ground the research work that I'm doing for this course in the product development I am doing for my startup, Chelle. We differentiate the core technologies of our startup as follows:

- Assets
- Knowledge
- Assessment
- Instruction

I have been working closely with my CTO on the assets component. With the research that I am doing, I very much have a focus on the knowledge component, and in particular, I am really digging in with ontological mappings and knowledge extraction.

In terms of ontological mappings, I have focused on the work of researchers like Nicolas Matentzoglu on projects such as the Simple Standard for Sharing Ontological Mappings (SSSOM), MapperGPT, and the Artificial Intelligence Ontology (AIO). In terms of knowledge extraction, I've started looking at approaches to populating knowledge bases, such as the SPIRES method. I continue to be interested in the potential of LLMs in enhancing precision in knowledge extraction, as demonstrated by the MapperGPT project.

### Papers

**Citation**: Dooley, D. M., Griffiths, E. J., Gosal, G. S., Buttigieg, P. L., Hoehndorf, R., Lange, M. C., ... & Hsiao, W. W. (2018). FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration. npj Science of Food, 2(1), 23.

**How Found**: In references of another paper

**My Summary**: Discusses the development of FoodOn, an ontology designed to standardize food terminology. Focus on standardization in the service of interoperability. Used LanguaL, a food indexing thesaurus and transforms it into an OWL-formatted ontology

**My Takeaway**:

- supports the idea that ontological mappings can have value beyond bioinformatics
- reference on challenges of curation and maintenance
- uses dated tools, likely not useful as a technical reference

**Citation**: Matentzoglu, N., Braun, I., Caron, A. R., Goutte-Gattat, D., Gyori, B. M., Harris, N. L., ... & Mungall, C. J. (2023). A simple standard for ontological mappings 2023: updates on data model, collaborations and tooling. In OM@ ISWC (pp. 73-78).

**How Found**: Google Scholar Search

**My Summary**: Discusses latest advancements on SSSOM, a standard for sharing ontological mappings. New metadata elements and advancements in toolling. Toward better sharing of ontological mappings

**My Takeaway**:

- use the sssom-py library and CLI tools
- likely a useful standard at least as a reference for data model if not the actual data model
- favors flat over nested as we do

**Citation**: Qiang, Z., Wang, W., & Taylor, K. (2025). Agent-OM: Leveraging LLM Agents for Ontology Matching. *Proceedings of the VLDB Endowment, 18*(1). https://doi.org/10.48550/arXiv.2312.00326

**How Found**: In references of another paper

**My Summary**: Introduces a framework for using LLMs to perform ontological mappings. Two agents - Retrieval Agent, Matching Agent. They use CoT prompting and various RAG strategies. High precision and recall in various OAEI tracks.

**My Takeaway**:

- excellent reference for LLM processing DAG
- hybrid database use (relational and vector)
- useful for prompting strategies
- some fine-tuning considerations

**Citation**: Jackson, R. C., Balhoff, J. P., Douglass, E., Harris, N. L., Mungall, C. J., & Overton, J. A. (2019). ROBOT: a tool for automating ontology workflows. BMC bioinformatics, 20, 1-10.

**How Found**: In references of another paper

**My Summary**: Introduces ROBOT, a tool designed to automate various tasks involved in ontology development. ROBOT provides a command-line interface and a library of high-level operations based on the OWL API. Insights related to Ontological Mapping include the use of tools like ROBOT to automate the extraction and integration of terms from multiple ontologies.

**My Takeaway**:

- template-driven ontology generation
- interacts with other tools, possible descriptive framework for processing DAGs
- designed toward being central piece of ETL pipeline

**Citation**: Guarino, N., & Giaretta, P. (1995). Ontologies and knowledge bases. Towards very large knowledge bases, 1-2.

**How Found**: In references of another paper

**My Summary**: Aims to clarify the various interpretations of the term "ontology". Propose distinct technical terms to differentiate between ontologies as conceptual frameworks and as concrete artifacts. Defining and classifying the term "ontology" to prevent confusion in the AI community.

**My Takeaway**:

- reference for our own understanding of ontological mapping
- toward a shared vocabulary … Nash (CTO) started in on this
- tool recommendations a bit dated

**Citation**:  Osumi-Sutherland, D., Courtot, M., Balhoff, J. P., & Mungall, C. (2017). Dead simple OWL design patterns. Journal of biomedical semantics, 8, 1-7.

**How Found**:  In references of another paper

**My Summary**:  Introduces Dead Simple OWL Design Patterns, a system designed to simplify the creation, documentation, and validation of design patterns in bio-ontologies. System leverages YAML-based syntax. Aims to automate and standardize the classification processes within bio-ontologies.

**My Takeaway**:

- conceptual approach; toward simplicity
- dated in technology
- look at what older technologies did (TermGenie)

**Citation**:  Bikeyev, A. (2023). Synthetic ontologies: A hypothesis. Available at SSRN 4373537.

**How Found**:  In references of another paper

**My Summary**:  Using machine-generated knowledge models to create synthetic ontologies using large language models like GPT-3. Fully automated bottom-up approach to ontology engineering, leveraging LLMs to generate hierarchies of terms and extract relationships between them. Aims to create more comprehensive and accurate knowledge graphs that can be tailored to specific use cases and updated as new data becomes available.

**My Takeaway**:

- of great interest, our approach will by necessity need to be fully automated
- incorporates human in the loop verification as we will also need to
- useful for prompting strategies

**Citation**:  Babaei Giglou, H., D’Souza, J., & Auer, S. (2023, October). LLMs4OL: Large language models for ontology learning. In International Semantic Web Conference (pp. 408-427). Cham: Springer Nature Switzerland.

**How Found**:  In references of another paper

**My Summary**:  Can LLMs effectively perform ontology learning tasks: term typing, taxonomy discovery, non-taxonomic relation extraction? Evaluations using zero-shot prompting and tested various LLMs, including BERT, GPT-3, GPT-4.

Fine-tuning significantly improves LLMs' performance, making them suitable assistants for ontology construction.

**My Takeaway**:

- zero-shot can help to prototype
- fine-tune LLMs on specific ontological learning tasks
- useful for prompting strategies toward specific tasks
- points at fewer challenges in general areas

**Citation**: Moxon, S. A., Solbrig, H., Unni, D. R., Jiao, D., Bruskiewich, R. M., Balhoff, J. P., ... & Mungall, C. J. (2021). The Linked Data Modeling Language (LinkML): A General-Purpose Data Modeling Framework Grounded in Machine-Readable Semantics. ICBO, 3073, 148-151.

**How Found**: In references of another paper

**My Summary**: Linked Data Modeling Language (LinkML). Leverages semantic web standards to produce ontology-ready data, aligning with FAIR principles. Supports a variety of data structures, from simple checklists to complex interrelated datasets. Integrates with existing tools and databases.

**My Takeaway**:

- possible at rest format for ontological encodings
- is a new system to learn; starting to hit a bit of system fatigue
- which set of tools is the correct set of tools for our purposes

**Citation**: Matentzoglu, N., Goutte-Gattat, D., Tan, S. Z. K., Balhoff, J. P., Carbon, S., Caron, A. R., ... & Osumi-Sutherland, D. (2022). Ontology Development Kit: a toolkit for building, maintaining and standardizing biomedical ontologies. Database, 2022, baac087.

**How Found**: Google Scholar Search

**My Summary**: Toolkit designed to streamline the creation, maintenance, and standardization of ontologies. Integrates a variety of tools and workflows within a Docker image. Facilitating quality control, release management, and dependency management. Allows SMEs to focus on content rather than the intricacies of engineering.

**My Takeaway**:

- framework for building entire ETL flows

- uses Docker which we are already using in our stack
- emphasis on CICD processes is great for ETL

**Citation**: Mateiu, P., & Groza, A. (2023, September). Ontology engineering with large language models. In 2023 25th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC) (pp. 226-229). IEEE.

**How Found**: Google Search

**My Summary**: Explores the use of GPT-3 for translating natural language descriptions into description logic representations. Fine-tuned a GPT-3 model using a dataset of NL-DL pairs. Reduce need for human curation.

**My Takeaway**:

- not sure if we will be looking at fine-tuning at first
- human curation will be a must
- human curation may be a feature rather than a bug

**Citation**: Jackson, R., Matentzoglu, N., Overton, J. A., Vita, R., Balhoff, J. P., Buttigieg, P. L., ... & Peters, B. (2021). OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies. Database, 2021, baab069.

**How Found**: In references of another paper

**My Summary**: Efforts to formalize and automate the evaluation of ontologies within the OBO Foundry. Objective was to enhance the quality and interoperability of ontologies by encoding principles into automated validation checks. Make data findable, accessible, interoperable, and reusable (FAIR).

**My Takeaway**:

- useful for looking at stitching various tools together (ROBOT, LinkML)
- useful as survey of what's been done
- may be a bit dated

**Citation**: Forssell, H., Kindermann, C., Lupp, D. P., Sattler, U., & Thorstensen, E. (2018). Generating ontologies from templates: A rule-based approach for capturing regularity. arXiv preprint arXiv:1809.10436.

**How Found**: In references of another paper

**My Summary**: Second-order language designed for specifying ontologies using a rule-based approach. Leverages ontology templates (OTTR) to capture recurring patterns in ontological modeling. Define the language and its semantics, explore reasoning over ontologies specified using this language.

**My Takeaway**:

- template-based approach
- a bit dated but useful in understanding approach
- toward a fully-automated system

**Citation**:  Guarino, N., Oberle, D., & Staab, S. (2009). What is an ontology?. Handbook on ontologies, 1-17.

**How Found**:  In references of another paper

**My Summary**:  Various definitions and aspects of ontologies, particularly in the context of computational systems. Definitions of ontologies and provides a detailed account of the notions of conceptualization and explicit specification. How to clearly define and formalize the concept of an ontology to make it useful for computational purposes. How conceptualizations can be mapped to ontologies using formal languages and meaning postulates.

**My Takeaway**:

- on the development of ontologies
- will there be shared ontologies across organizations? is this a useful idea
- much older paper but useful in understanding the field

**Citation**:  Jonquet, C., & Grau, N. (2024). M4. 4-Review of Semantic Artefact Catalogues and guidelines for serving FAIR semantic artefacts in EOSC.

**How Found**:  Google Scholar Search

**My Summary**:  Improve the user experience when interacting with complex biomedical ontologies. Simplify ontology visualization by highlighting key root terms, thus making the ontology more accessible to non-expert users. Ontology browsers can present a more user-friendly hierarchy.

**My Takeaway**:

- useful in thinking UI/UX
- opens up a whole new field of questioning, the presentation of ontologies to end users
- follow up on references here

## Synthesis

I have created Diagram 1 to help me begin to think about the research that I am doing, and how it informs the knowledge piece of our product development. I'm struggling a little bit with synthesizing this into a whole so this synthesis may be more of a stream of consciousness with many thoughts. If there is a deliverable that I'm working towards it is an architectural system diagram to reflect all of the things that I'm learning here. One of the things that I'm really struggling with is leveraging tools, especially SSSOM, ROBOT, and LinkML (Matentzoglu et al., 2023; Jackson et al., 2019; Moxon et al., 2021). I don't have enough knowledge in this space to understand how to architect a system using these and other tools.

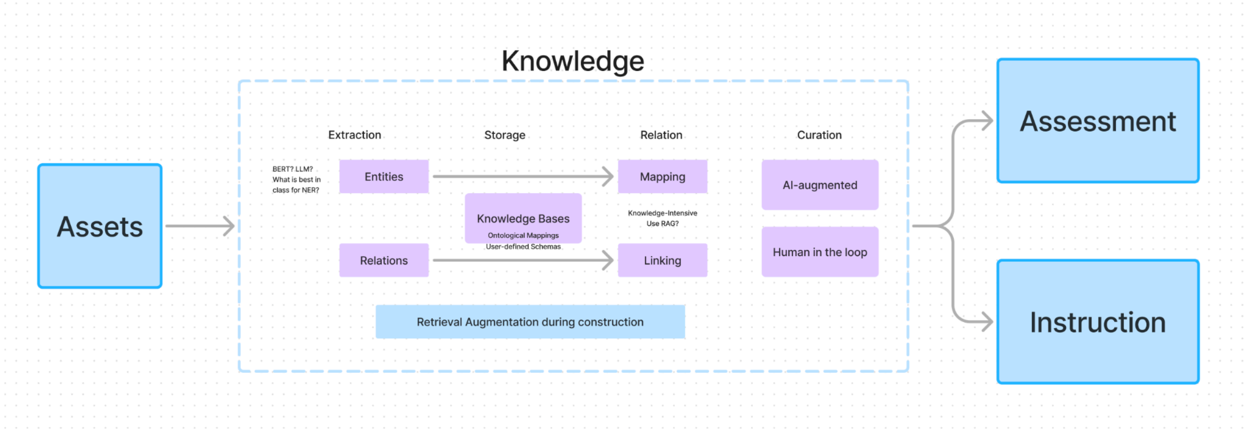

Diagram 1: Preliminary Diagram of Knowledge Architecture

In terms of entity extraction and relation, I think that I've seen some really good tools for extraction, e.g. mapperGPT mentions using LOOM for entity extraction, and then using GPT to increase the precision of the entities extracted (Matentzoglu et al., 2023). Others are using BERT and its derivative models to perform entity extraction (Babaei Giglou et al., 2023). I think that we are going to need to design not only a process for doing the entity extraction, but ways to test the validity of each different method for extraction against the others. We're then going to need to do the same thing for the relation process.

Much of what we've looked at here is in terms of building automated systems, and this is critical. We also need to think about human-in-the-loop. I am beginning to think about this in terms of strong course developer and user experiences. One of the papers presented a way to simplify an ontology for presenting to us. It's almost like the human in the loop is not critical to construction of ontology, but rather a feature of the application that a user is able to view a created ontology and make changes where they think it's appropriate. In this context, I would think of non-expert users as people who are non-experts in the construction of ontologies, but are experts in terms of the subject matter of the domain being mapped (Jonquet & Grau, 2024). The value of Chelle is in constructing ontologies, and then presenting them to subject matter experts in ways that they can interact with them.

In terms of CICD, I'm just thinking about automation and I'm thinking that most of what we're doing can be described the same way that one might describe an ETL process (Matentzoglu et al., 2022). I'm beginning to think about all of this data engineering but with knowledge. This of course draws on my experience working at Databricks. One thing Nash (our CTO) and I discussed this week was the medallion architecture that we used at Databricks, especially how it relates to document parsing and constructing a complete system. According to Claude, "Databricks' medallion architecture organizes data processing into bronze (raw), silver (cleaned/conformed), and gold (aggregated/feature-rich) layers, progressively refining data quality and usability for analytics and machine learning applications". One thing of note and I know this from my time working at Databricks, is that we think of the silver layer as the single source of truth.

After this synthesis, the Diagram 2.

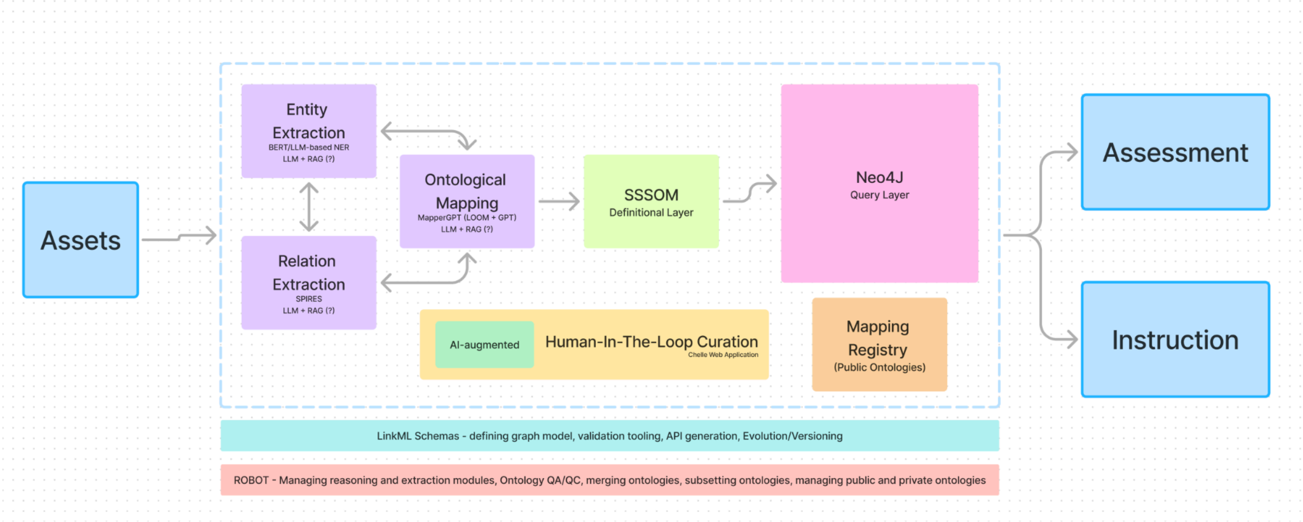

Diagram 2: Post-Synthesis Knowledge Architecture Diagram

## Reflection

Most of the papers identified this week were discovered by reading the references of papers that I had already read. For example, the SSSOM was an update to a paper. I looked out last week on the same topic. It wasn't particularly challenging. Actually finding papers just kind of fell into place. The part that was challenging was understanding what I was reading. The field of knowledge, extraction and ontological mapping is a rich and robust sub field of natural language processing. I have been working in NLP for years and never thought about this field. I feel a bit like I'm swimming.

I've prepared an architecture and feel like I could start building pieces of it. That said I very much feel like I am stabbing in the dark at the edge of my own knowledge. The biggest challenge here is fatigue from learning all of these new terms. I thought I knew a lot about NLP and artificial intelligence, but I don't have a ton of prior knowledge hooks to hang this new knowledge on. That it is all very exciting and feels like the right direction for everything that I'm working on.

In thinking about everything that I've studied, I'm left with many open questions:

1. What exactly does curation look like? Are users curating assets? Entities? Relations? Are we putting ontological mappings in front of them? All of this intersects with recent industry discussion around a lack of trust in AI. I like to think about AI augmentation as opposed to AI automation and finding ways to put the results of an AI process in front of users for approval and refinement.
2. How can retrieval augmented generation support this work? Qiang, et. al (2025) looks at Leveraging LLM agents for ontology matching and uses various RAG strategies to achieve this.
3. So many questions about treating this knowledge service as ETL on top of raw documents.
4. Fine-tuning?
5. Mapping Registry? Especially with public ontologies? There will be entities that exist across multiple organizations, and it will be useful to have a notion of public ontologies.

I think the next step is to implement a proof-of-concept, perhaps looking at the Artificial Intelligence Ontology that I read last week and either re-implementing that work or perhaps implementing a variation of that work on ontologies and the papers that I've been reading.

## Planning

Next week, I'd like to focus on the following:

1. Named entity recognition. What models are best in class? What are the best strategies? What are assessment strategies for comparing different models/strategies?
2. LOOM (Lexical OWL Ontology Matcher). I'd like to learn more about this tool.
3. Work done integrating LinkML with Neo4j and/or other data stores.
4. BoomerGPT. This is the tool I write about in week one that I think I dropped last week. I'd like to pick that up.
5. Explore the references in Qiang, et. al (2025). Sounds like they are doing some interesting things with LLMs and ontology mappings.
6. Research around knowledge extraction as ETL on documents.
7. Mapping registries.
8. Tools for human-in-the-loop curation.