PIPELINE FOR A SUSTAINABLE ACTIVITY ONTOLOGY: REPRESENTING THE EU TAXONOMY FOR CSRD WITH AUTOMATED INFORMATION EXTRACTION

Abstract

This project aims to develop a Linked Data ontology to support large European corporations in reporting their sustainability initiatives, mitigating challenges related to data gaps and interoperability. Utilizing a semi-automatic pipeline, the ontology is constructed based on the EU Taxonomy, offering a standardized framework for reporting. Leveraging the ChatGPT API, we enhance the ontology's comprehensibility. The resulting ontology streamlines reporting procedures and enables machine-readable publication of sustainability data, contributing to enhanced transparency and accountability within corporate sustainability practices.

Introduction

Starting in 2024, all large corporations in Europe will be required to report on the sustainability of their business activities under the CSRD and the EU taxonomy of sustainable activities. However, many companies are unprepared due to insufficient data from suppliers and data interoperability issues.Our primary goal is to develop a Linked Data ontology to help large European corporations report on sustainability, addressing data gaps and interoperability issues. Our semi-automatic pipeline creates a standardized ontology based on the EU Taxonomy, facilitating seamless reporting. Leveraging the ChatGPT API, we enhance ontology comprehensibility. This ontology streamlines reporting and fosters machine-readable sustainability data publication.

Background

Linked Data is a method of organizing and connecting information on the Web using standardized formats and identifiers, enabling seamless navigation and discovery of related data.The EU Taxonomy is a regulatory framework established by the European Union to define environmentally sustainable economic activities, providing clarity and standardization for investors and businesses seeking to align with green finance objectives. It categorizes activities based on their contribution to six key environmental objectives, guiding investment towards sustainable and climate-resilient projects.

Methodology

Ontology construction

The flowchart on the right displays the end-to-end approach. This pipeline comprises a complete process which enables the construction of an ontology representing the Taxonomy and involves only fractional manual steps (shown in dark blue in the figure). For detailed steps, please refer to the figure.

Evaluation

Our approach is evaluated in terms of three aspects. Firstly, we assess the quality of the free text extraction through precision and recall, as well as Flesch Reading Ease (FRE) scores. For the other two aspects, we conduct a study with peer participants to evaluate the quality of our pipeline and ontology. Participants apply our pipeline to construct the ontology as well as partake in a controlled test where they answer questions related to sustainability compliance using either our ontology or the original format of the Taxonomy provided on the EU website. A follow-up questionnaire and measuring the time needed to complete these tasks allows us to report on the (re)usability of our approach.

Results

We present the Linked Data schema displayed in the bottom-right figure as the result of our pipeline. The quality of this ontology, as well as the pipeline itself, is assessed through a peer survey which is still ongoing - as is the calculation of the precision and recall of text extraction since it requires a consensus of ground truth based on several observants, which is a time-consuming process. We can report that the mean FRE score of the bodies of free text in our ontology as compared to the original representation increased from 21.15 to 33.13 (significant difference with p=0.03), which shows an improvement in readibility.

Discussion

The results above confirm that sectioning large bodies of free text into smaller pieces of information aids comprehension. Furthermore, it is shown that automation can ease the process of representing the Taxonomy in a more interoperable and accessible data format. At the same time, completing the peer survey will be necessary to assess the general usability of our method. Furthermore, the scope of our schema is limited by the time constraints of our project and could be expanded in the future.

THE PIPELINE

AUTOMATIC MANUAL

Input
Taxonomy xlsx file published and maintained by the EU

Create ontology schema
Manually create ontology schema in Turtle RDF Language

(easy to update)

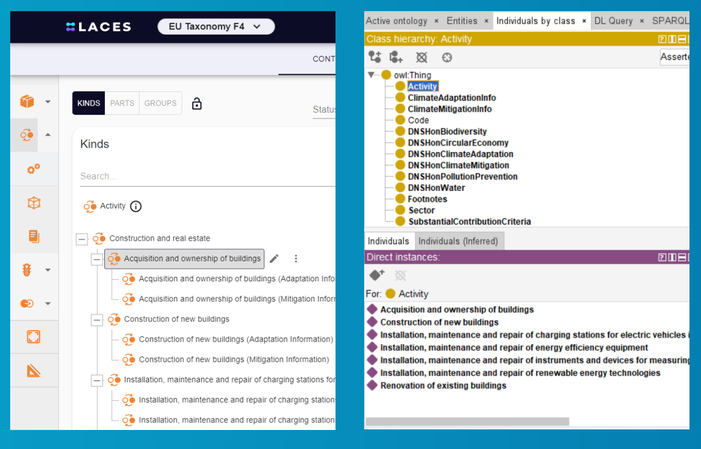
Convert data from EU into individuals of the ontology
Take rows of activities in xslx file and automatically create instances based on the schema.

(Python script that generates Turtle file (.ttl))

Extract information from free text
Using ChatGPT-4 extract valuable information from the vast free-text in the Taxonomy.

Combine into final Ontology
Add extracted concepts back to their original nodes and merge with ontology schema

Output
Final ontology file in Turtle Language



INPUT DATA

First, we download the EU taxonomy on sustainable activities from their website. In total, there are 13 different sectors within the taxonomy. Per sector, the substantial contribution criteria and do no significant harm criteria (DNSH) are described for each actitivty. This information is displayed for climate adaptation and climate mitigation. A header of the dataset is shown below:

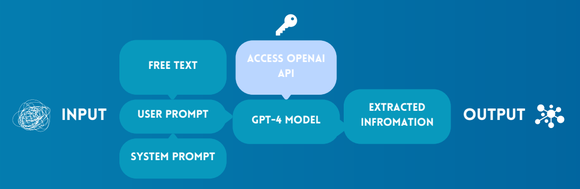
COLUMN	NACE	SECTOR	ACTIVITY NUMBER	ACTIVITY	CONTRIBUTION TYPE	DESCRIPTION	...	DNSH ON BIODIVERSITY	FOOTNOTES
DATA	A2	Forestry	1.1	afforestation	-	Establishment of forest through planting ...	...	in areas designated by the national authority ...	(1) Establishment of forest through ...

.XLS TO .TTL

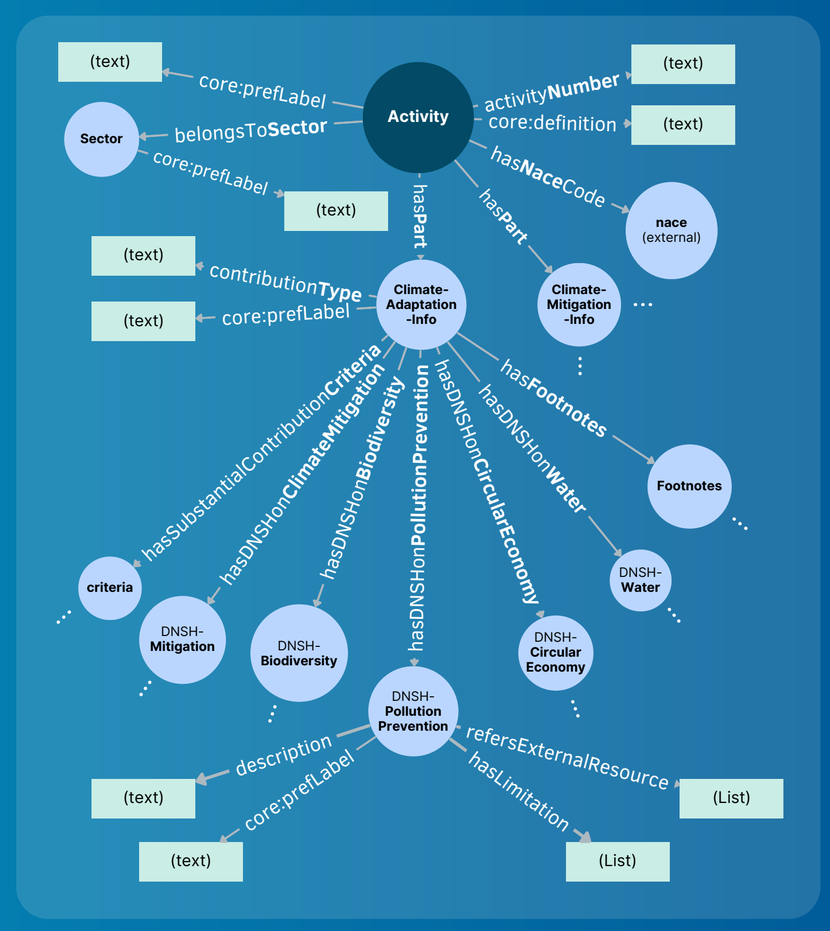
Using Python, the XLS file is automatically parsed and instances are created for all activities for both climate adaptation and climate mitigation goals.

TEXT EXTRACTION

The ChatGPT 4 API is used to extract nuanced information from free-text descriptions within the original xls file. This function schematically is shown below.



ONTOLOGY SCHEMA



Future work

Our future work encompasses several key areas: We plan to leverage data from diverse sectors to construct a comprehensive architecture through Linked Data. This will enable us to integrate various data sources for more comprehensive information processing and analysis. Additionally, we will implement ontology-based platforms to synchronize data updates from the EU Taxonomy website, ensuring that we always analyze and make decisions using the latest information. To further enhance model performance and adaptability, we will employ hyperparameter optimization techniques, fine-tuning parameters to ensure superior performance across various datasets and tasks. Furthermore, we aim to train Large Language Models (LLMs) by leveraging extensive datasets and integrating user feedback for optimization. These efforts aim to propel our technical and analytical capabilities forward to better meet evolving needs and challenges

UvA Data Systems Project 2024: Semmtech - Group F4
Xiaotian Chen, Florian de Graaf, Zsófia Katona, Dimitrios Papadopoulos