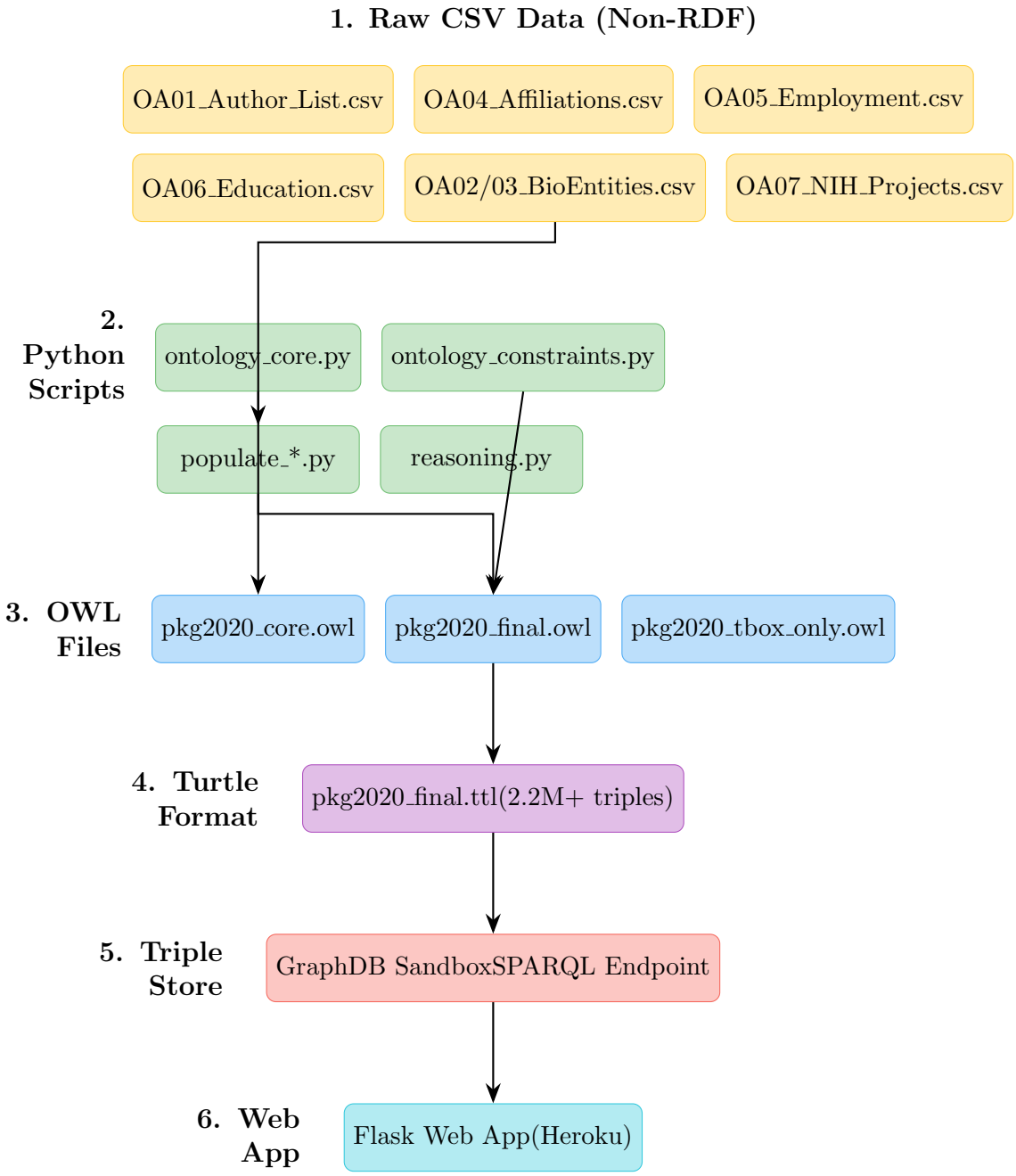
PKG2020 Knowledge Graph Pipeline

KRR Project - Data Conversion Workflow

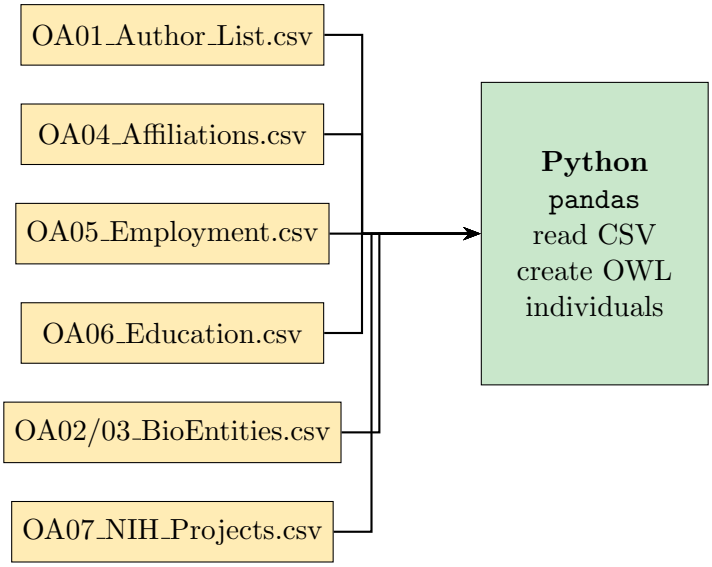
Overview

This document shows the complete pipeline for converting raw CSV data into a semantic knowledge graph with SPARQL endpoint and web visualization.

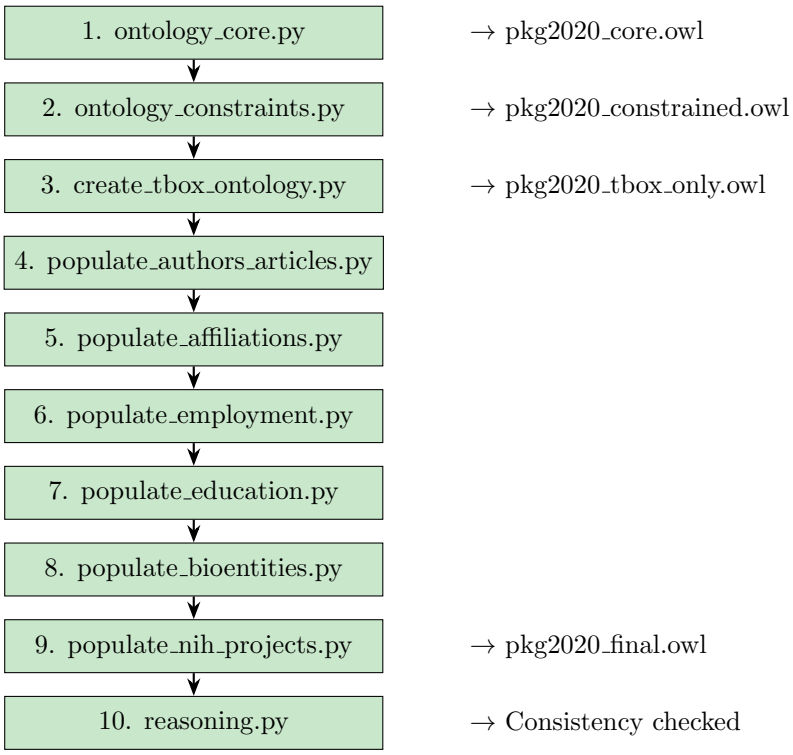

1. Raw CSV Data (Non-RDF)
OA01_Author_List.csv
OA04_Affiliations.csv
OA05_Employment.csv
OA06_Education.csv
OA02/03_BioEntities.csv
OA07_NIH_Projects.csv
2. Python Scripts
ontology_core.py
ontology_constraints.py
populate_*.py
reasoning.py
3. OWL Files
pkg2020_core.owl
pkg2020_final.owl
pkg2020_tbox_only.owl
4. Turtle Format
pkg2020_final.ttl(2.2M+ triples)
5. Triple Store
GraphDB SandboxSPARQL Endpoint
6. Web App
Flask Web App(Heroku)

Pipeline Steps - Detailed

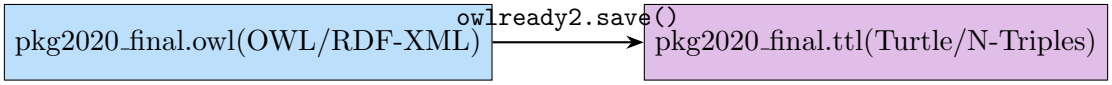
Step 1: Raw CSV Data → Python Processing



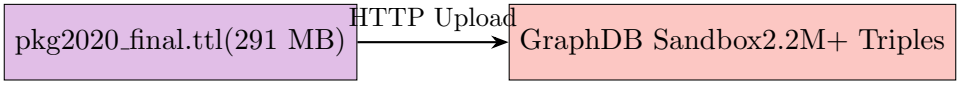
Step 2: Python Scripts Execution Order



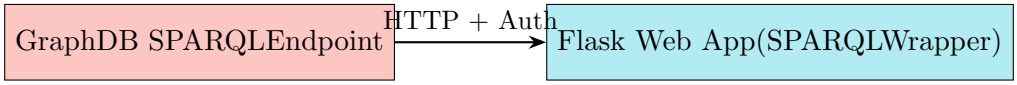
Step 3: OWL to Turtle Conversion



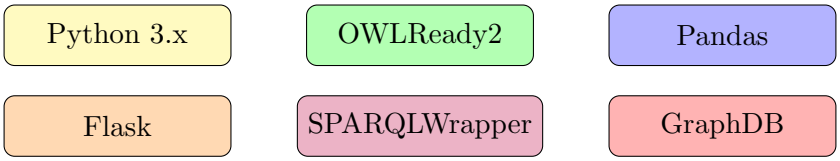
Step 4: Upload to GraphDB



Step 5: SPARQL Endpoint Access



Output Files Summary

File	Purpose	Size
pkg2020_tbox_only.owl	T-Box (schema only)	14 KB
pkg2020_hand_annotated.owl	10+ hand-annotated	19 KB
pkg2020_with_swrl.owl	SWRL rules	27 KB
pkg2020_final.owl	Complete ontology	119 MB
pkg2020_final.ttl	Turtle for GraphDB	291 MB

Technology Stack