A High-throughput Data Ingest Pipeline for Semantic Data Stores *…and a look at the Plant Phenology Ontology*

J

John Deck, Brian Stucky, Ramona Walls, Rodney Ewing, Melissa Genazzio, Henry W Loescher, Robert Guralnick

Science Across Virtual Institutes (SAVI)
NSF-SAVI Award Number 1321595

Powered by CYVERSE™

Input Datasets

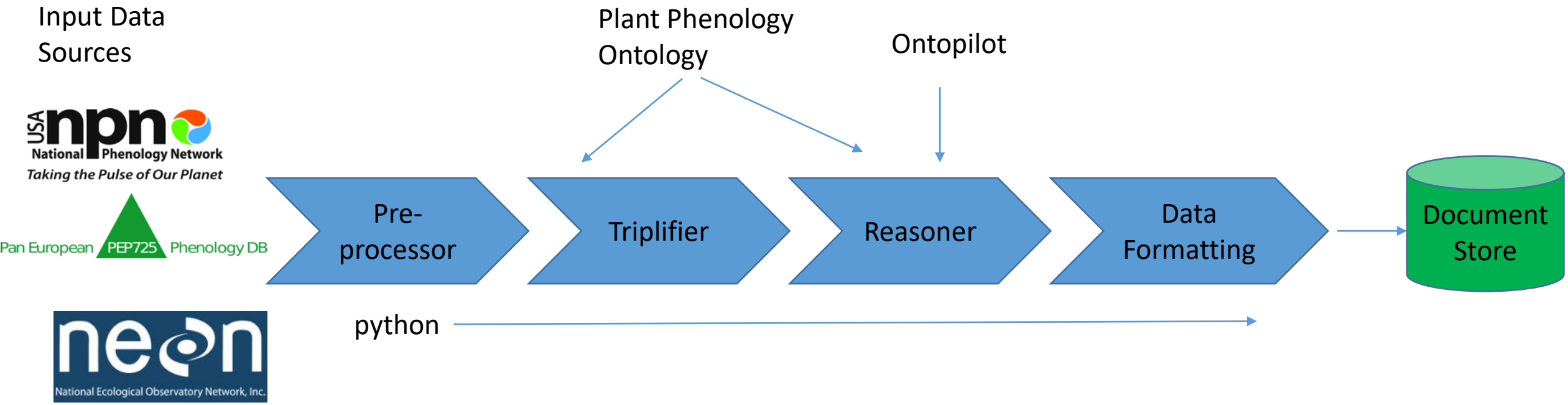
National Phenology Network
9 million observations; Presence/Absence – network of multiple US datasets

Pan European Phenology Database
10 million observations; Presence only – network of multiple European datasets

National Ecological Observatory Network
1.5 million records from just the last 3 years

Current total of 20+ million phenology observations

Data Integration Pipeline



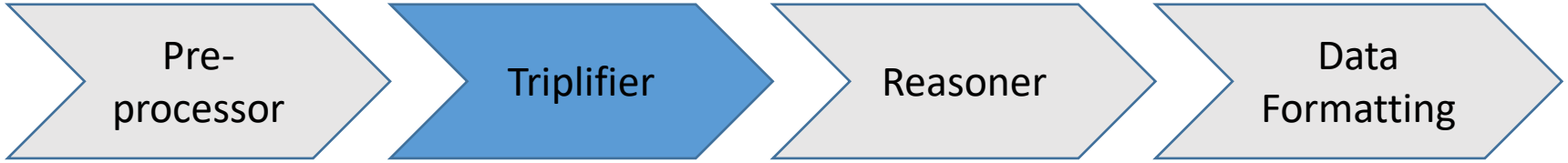
Pipeline code available at:

https://github.com/biocodellc/ppo-data-pipeline/

- Sanitizes and standardizes input data sources
- Project specific pre-processors inherit (python) abstract class

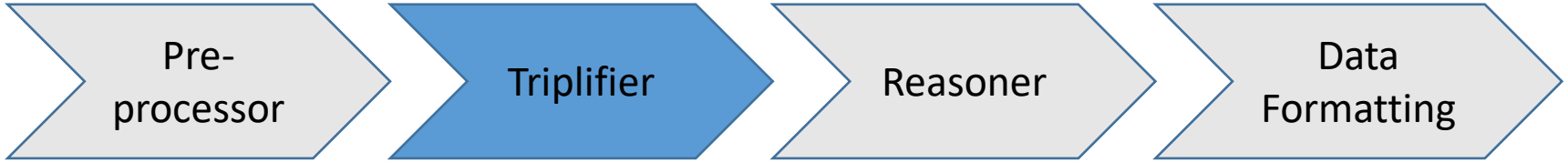
Sample pre Processor output: Stored as CSV File:

Obs ID	Lat	Long	Genus	Species	PhenophaseDescription	Day of Year	Lower count	Upper count	Year	All Stages
5943640	34.675	-120.041	Quercus	Lobata	Increasing Leaf Size	106	10	20	2013	



Triplifer Objectives

- Using incoming datasets and an ontology, generate a graph structure for all individuals, containing: plant structure, whole plant, observation process, and measurement
- Map project terms to ontology terms with delimited text files
- Converts Tabular data to RDF triples

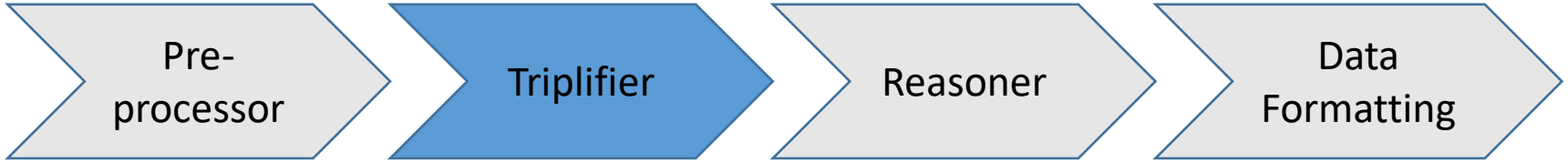


Triplifier Mapping: Plant Structure Presence Mapping File (CSV)

Executable File | 103 lines (102 sloc) | 7.59 KB

Search this file...

	USA National Phenology Network Field Name	PPO Labels
2	Flower heads (grasses/sedges)	{non-senesced flower head presence}
3	Flowers or flower buds	{non-senesced floral structure presence}
4	Open flowers (lilac)	{opened flower presence}
5	Full flowering (lilac)	{opened flower presence}
6	Open flowers (grasses/sedges)	{opened flower presence}
7	Open flowers	{opened flower presence}
8	Open pollen cones (conifers)	{open pollen cone presence}
9	Pollen cones (conifers)	{pollen cone presence}



Triplifier Observation Model:

IAO: Information Artifact Ontology
PO: Plant Ontology
PPO: Plant Phenology Ontology

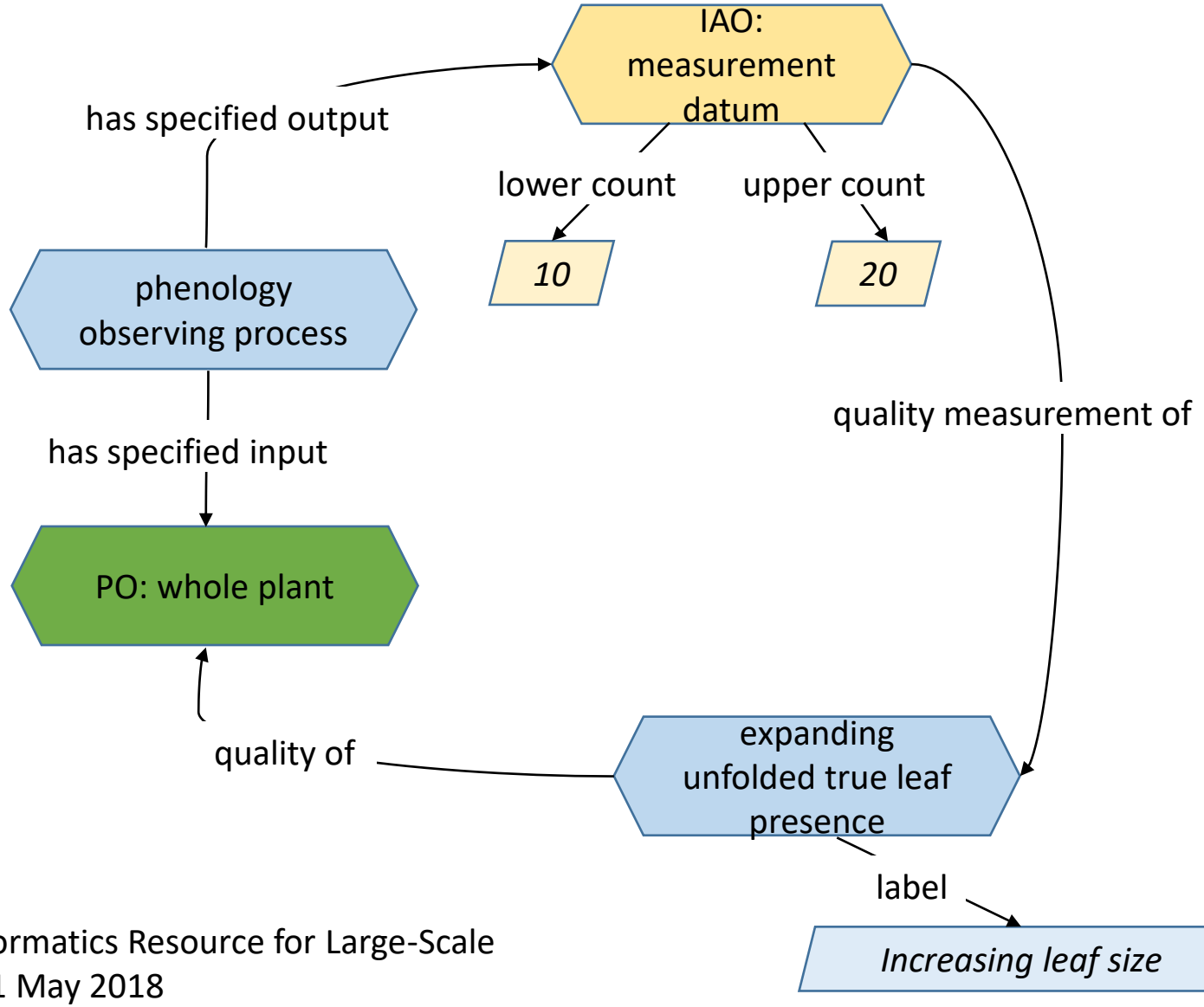
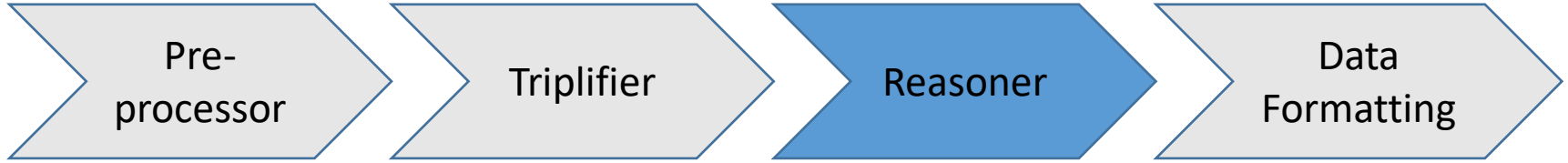
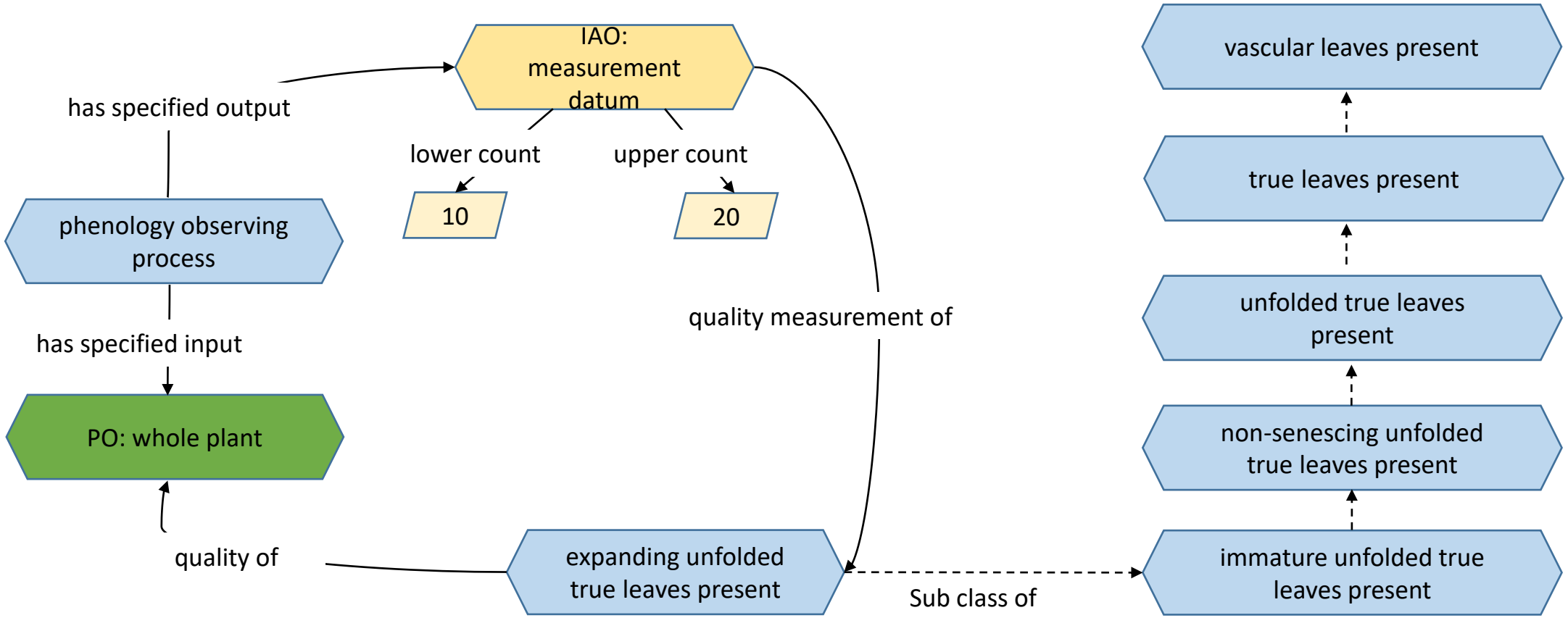


Figure from:
Stucky, et al. The Plant Phenology Ontology: A New Informatics Resource for Large-Scale Integration of Plant Phenology Data Front. Plant Sci., 01 May 2018



Reasoner: uses ontopilot software (https://github.com/stuckyb/ontopilot) uses a modified ELK reasoner to enable fast reasoning over instance data along with the specified ontology.



Starting Point (from pre-processor)

Obs ID	Lat	Long	Genus	Species	Phenophase Description	Day of Year	Lower count	Upper count	Year	All Stages
5943640	34.673	-120.068	Quercus	Lobata	Increase leaf size	106	10	20	2013	

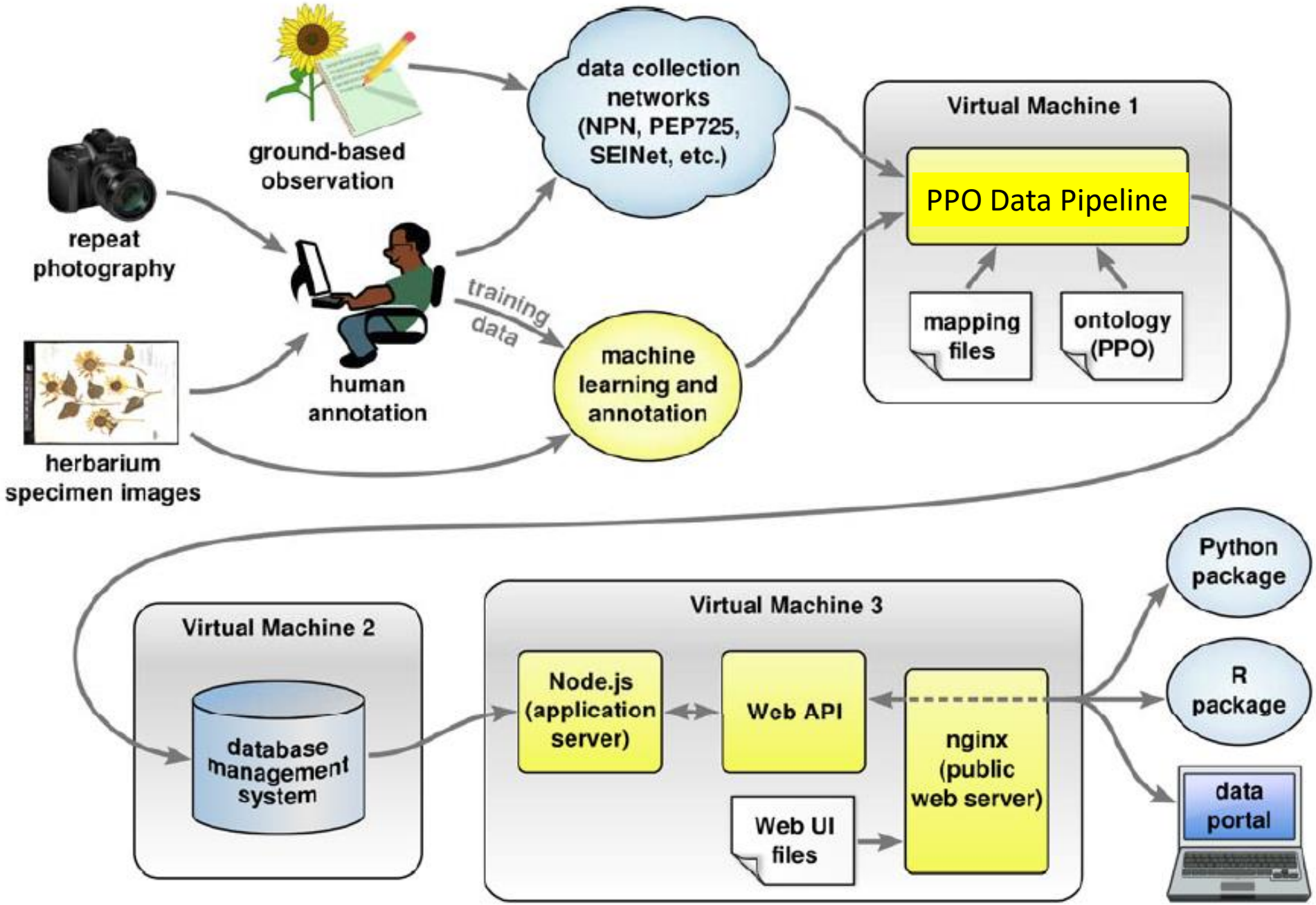
Final Result:

Obs ID	Lat	Long	Genus	Species	Phenophase Description	Day_ of_ Year	Lower count	Upper count	Year	All Stages
5943640	34.673	-120.068	Quercus	lobata	Increasing leaf size	106	10	20	2013	Expanding unfolded true leaves present; Immature unfolded true leaves present; Unfolded true leaves present; True leaves present; Vascular leaves present

Capacities

- 20 million records through pipeline in 6 hours on 8 cpu / 16Gb RAM server

- Final format can be easily loaded into a relational db or document store. We have tested on 80 million record sample running on Elasticsearch set with query results in <2 seconds.


ground-based observation
data collection networks (NPN, PEP725, SEINet, etc.)
repeat photography
human annotation
training data
machine learning and annotation
herbarium specimen images
Virtual Machine 1
PPO Data Pipeline
mapping files
ontology (PPO)
Virtual Machine 2
database management system
Virtual Machine 3
Node.js (application server)
Web API
Web UI files
nginx (public web server)
Python package
R package
data portal

Global Plant Phenology Data Portal

Map | Table

Traits flowers present

Genus Quercus

Species

Source
-- Select Source(s) --
USA National Phenology Network
Pan European Phenology Database
National Ecological Observatory Network

Year 1868 — 2018

Day of Year 1 — 365

Search | Download

Field Data on plant phenology is made accessible here based on inputs from USA-NPN, PEP725, and NEON partners. Incoming trait data is processed using the ppo-data-pipeline with annotations from the plant phenology ontology and phenological stage inferences processed using ontopilot's reasoner.

USA-NPN and NEON data is harvested every month. PEP725 provides yearly updates to their data store, typically in May. The date of last refresh is listed currently as part of the download package under "citation_and_use_polices.txt", and available online at the ppo-data-server citation file. Currently, this portal contains over 20 million records, with 10 million PEP725 records, 9.5 million USA-NPN records, and 1.5 million NEON records.

Funding support is from the USGS NSF-SAVI Award Number 1321595, the National Ecological Observatory Network, and the USGS Powell Center. *Development of this site is ongoing and currently this page is in a Beta release stage, made available here for viewing and comment. Comments can be made by email to jdeck88@gmail.com*

https://www.plantphenology.org/

rppo

An R package for accessing the PPO Data Portal

John Deck, Brian Stucky, Ramona Walls, Kjell Bolmgren, Ellen Denny, Robert Guralnick' (2018). rppo: An interface to the Plant Phenology Ontology and associated data store. R package version 1.0 https://github.com/biocodellc/rppo

This package is part of the rOpenSci project

To learn more, please visit http://ropensci.org

http://github.com/ropensci/rppo

Next Steps

- Process herbarium data (SEINET has marked up records with phenology terms… just need to be mapped)
- Machine learning and annotation of herbarium records
- Process Phenocam data (requires work with ontology to map green-ups to ontology terms)

Pipeline Development:
https://github.com/biocodellc/ppo-data-pipeline

R package:
https://github.com/ropensci/rppo

Portal:
https://www.plantphenology.org/

Plant Phenology Ontology:
https://github.com/PlantPhenoOntology/ppo

Thank You

John Deck
jdeck@berkeley.edu

Stucky, et al. The Plant Phenology Ontology: A New Informatics Resource for Large-Scale Integration of Plant Phenology Data Front. Plant Sci., 01 May 2018