Building an Analysis-Ready Data Resource Using a Multi-Stage Pipeline

Grace Donovan
MSDE/MSDS Practicum 2
Regis University
August 21, 2025

Overview

Introduction

Background, Objectives, and Importance

Project overview

Retrieve, Cleanup, and Enhance

Conclusion

Limitations and Future Improvements

COMMUNITY for DATA INTEGRATION

What: Collaborative network of researchers

Why:

- Forum for data practitioners to come together, share their knowledge, and learn from one another
- Advance understanding of earth systems through use of data, information, tools, and techniques

How:

- **Collaboration Areas:** Groups formed around common interests that help address challenges and identify solutions that enable data integration efforts.
- **Annual CDI Request for Proposals:** Seed funds awarded each year for projects that focus on data integration.
- **Biannual Workshops:** In-person workshop where participants present their work, propose collaborative paths to solve data challenges, and share knowledge with network of peers.

Images sourced from the CDI Website: https://www.usgs.gov/centers/community-for-data-integration-cdi

An Emoji Evolution of Project Objectives

🎯 ❌ 🔄 🧩 📊 ➡️ 🗄️ 📈 🔍

Updated Objective

Harmonize, integrate, and enrich CDI project data into a centralized data hub, transforming it into accessible, analysis ready data!

How C... project innovatio... ...ed collaborati... ...ks, and amp... within ... comm...

❌

✅

Initial Objective

Understand how CDI projects have promoted innovation, strengthened collaborative networks, and amplified impact within the scientific community.

...ate, and ... data into ... hub, ...s ready

🔍

Why is Data Integration Important?

"Increasingly, data scientists must first sort through heterogeneous, incongruent, and fragmented datasets before any analyses can be conducted. Such problems with data availability are often exacerbated in emergency situations where real time analyses are often stymied by unevenly documented or unclean data."
(Cheng, C. et al, 2024)

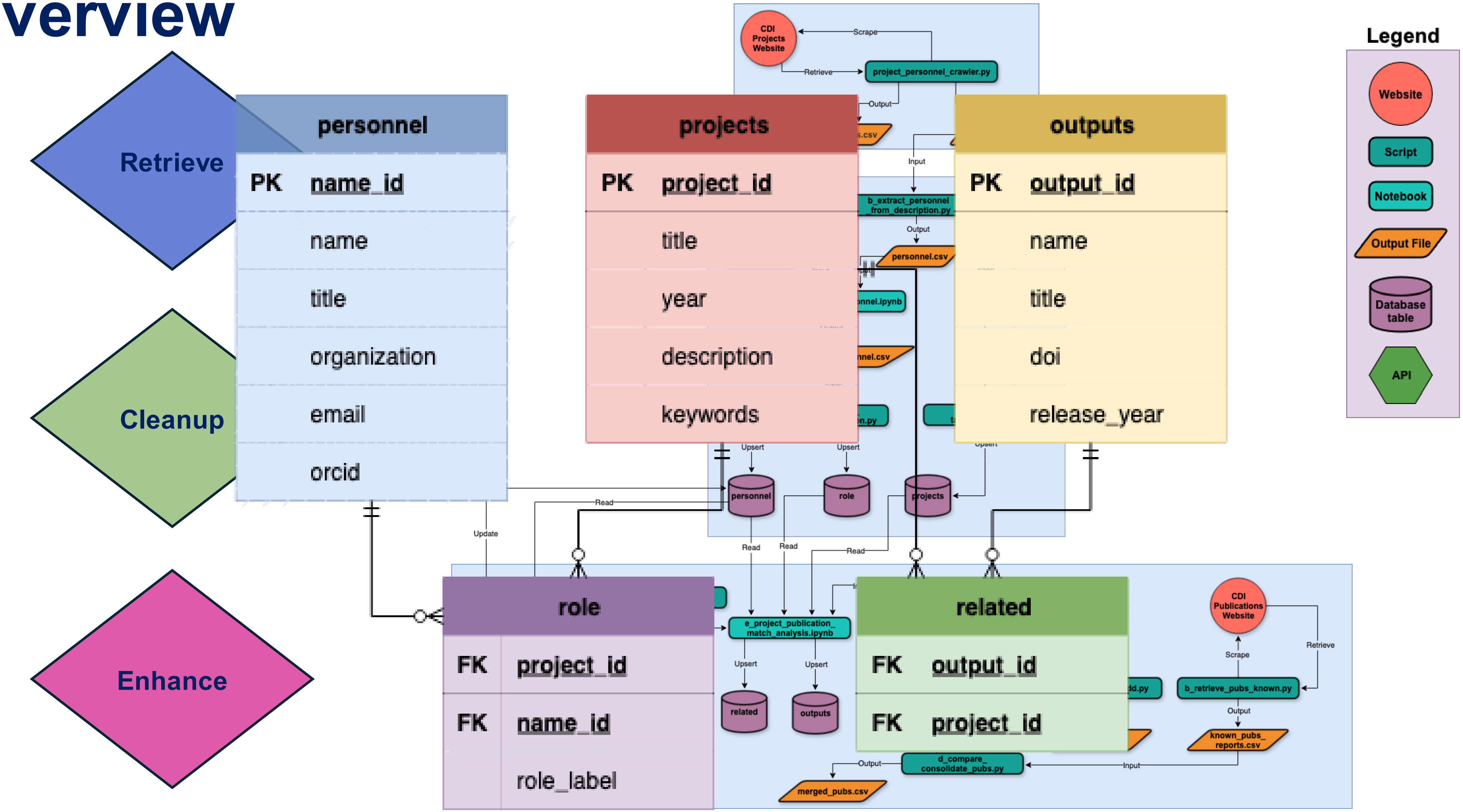
"One of the great challenges of data-driven research is that data rarely come in a form that is immediately ready for analysis."
(Huber, R. et al, 2021).

"Integrating structured, semi-structured, and unstructured data is crucial for insights and decisions. However, complexities like format mismatches, data quality issues, and interoperability hinder efficient data utilization"
(Mirza, F., 2021)

"Data integration solutions leveraging data-driven techniques improve data-delivery capabilities and facilitate data analytics."
(Liu, D. and Yoon, V., 2024)

"Integration of existing information systems is becoming more and more indispensable in order to dynamically meet business and customer needs while leveraging long-term investments in existing IT infrastructure."
(Ziegler, P., and Dittrich, K., 2007)

Overview



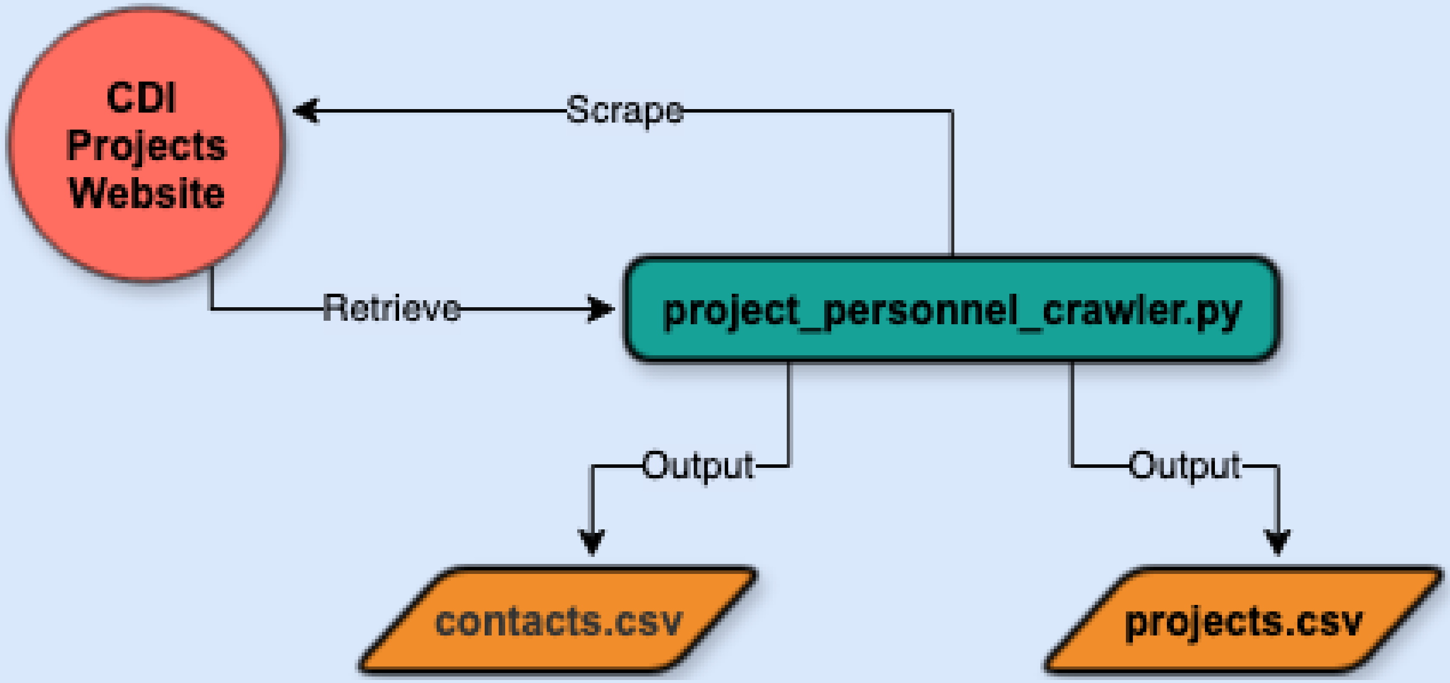
personnel	
PK	name_id
	name
	title
	organization
	email
	orcid

projects	
PK	project_id
	title
	year
	description
	keywords

outputs	
PK	output_id
	name
	title
	doi
	release_year

role	
FK	project_id
FK	name_id
	role_label

related	
FK	output_id
FK	project_id



Retrieve

- Scrape full-text content from CDI project web pages
- Extract title, year, description, and contacts
- Export structured project and project contacts (personnel) data

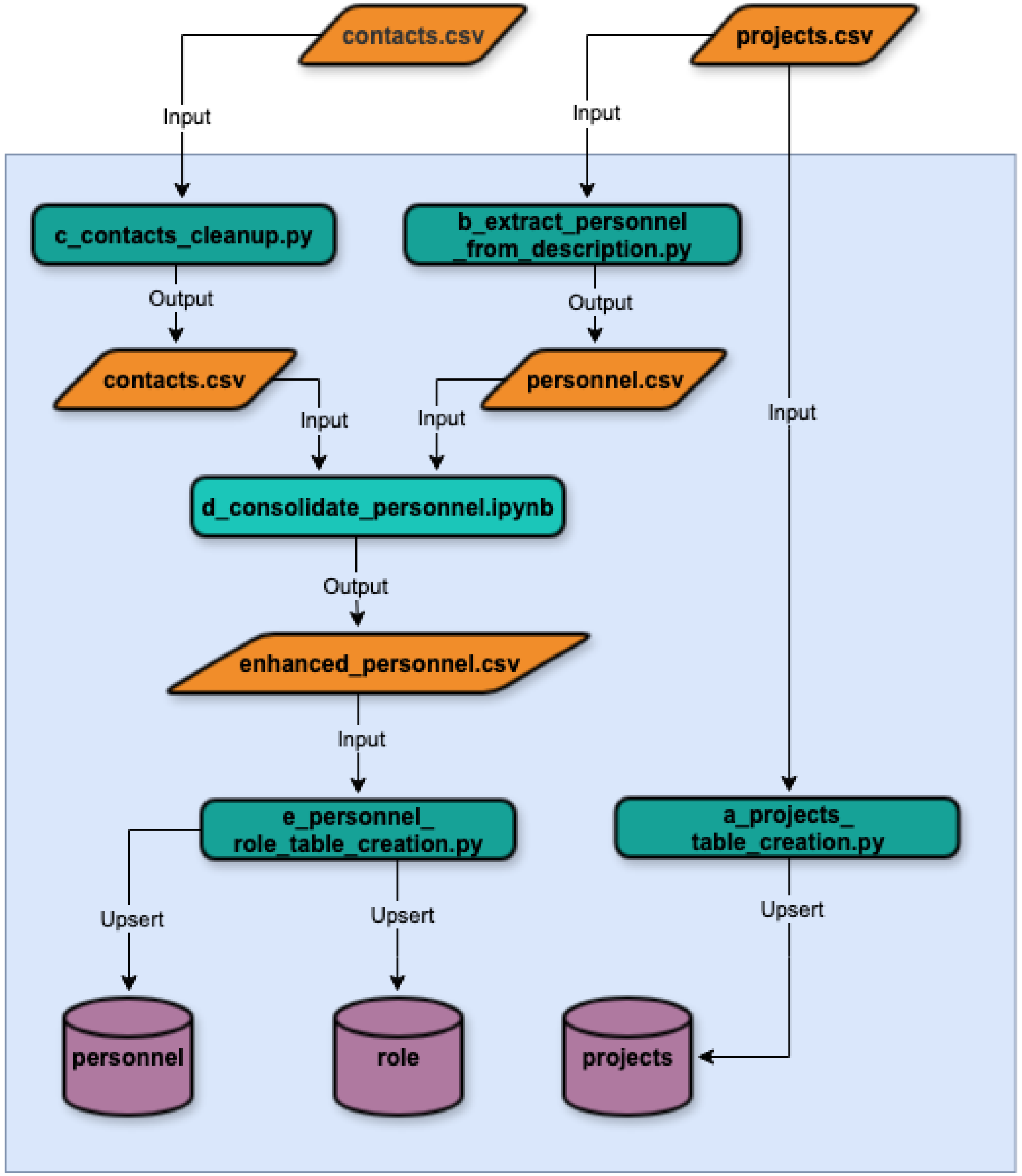
Cleanup

Projects: clean descriptions, assign unique identifiers, extract top-5 keywords, upsert into projects table

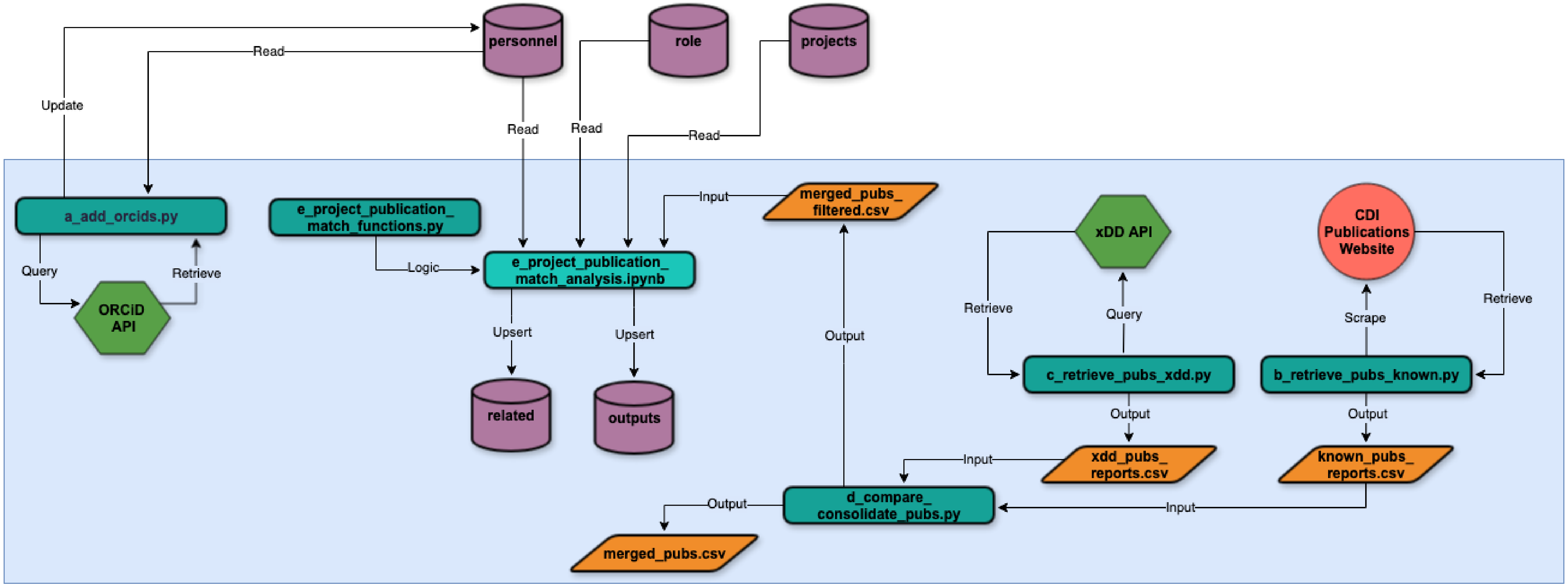
Personnel: regex-parse and clean names, tag lead PI versus collaborators, and standardize entries

Consolidation: merge personnel information from contact sections and descriptions, fuzzy-match name entries, select representative personnel entries, and remove duplicates

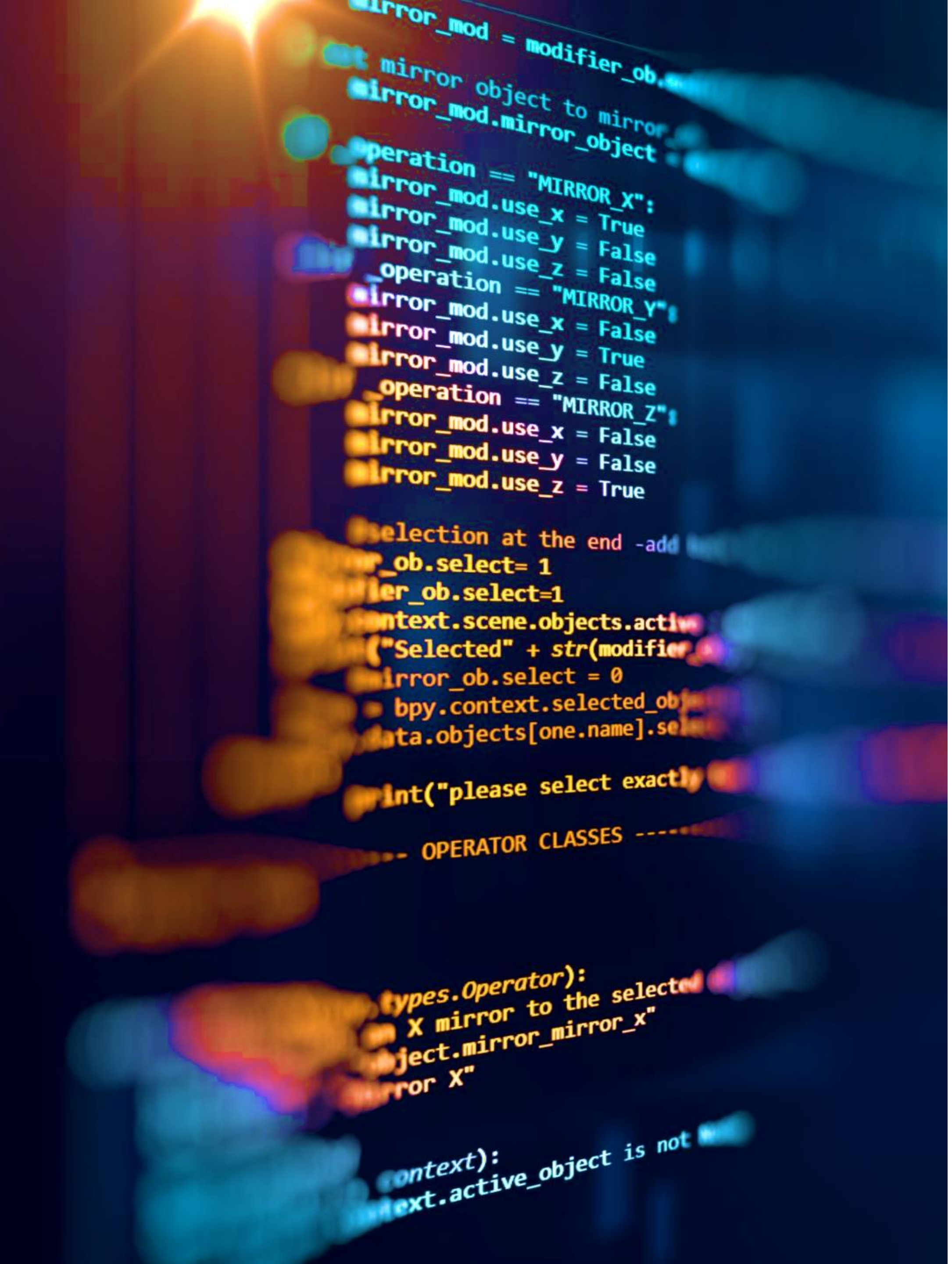
Final Load: assign each personnel entry a unique identifier, upsert into personnel and role tables in database



Enhance



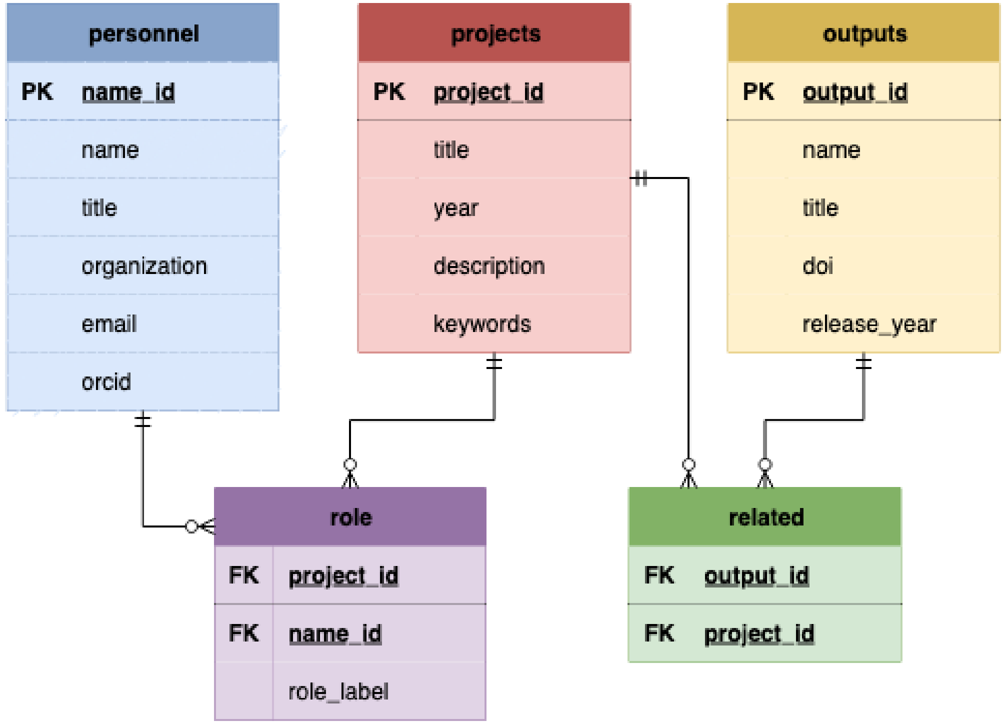
- ORCID enrichment via ORCID API
- Known publications retrieval from CDI
- xDD API publication extraction
- Publications consolidation and filtering
- Project-publication matching and upsert



Limitations

- Missing and incomplete publicly accessible CDI project data
- Project personnel may not have ORCIDs or have different names associated with their ORCIDs
- xDD is not a comprehensive corpus of all published works by CDI project personnel
- Project personnel may not be acknowledging the CDI or have cited the CDI inconsistent ways in their outputs
- Heuristic matching errors
- Fixed database schema may require refactoring if new data types are included in the future

Data Integration



personnel	
PK	name_id
	name
	title
	organization
	email
	orcid

projects	
PK	project_id
	title
	year
	description
	keywords

outputs	
PK	output_id
	name
	title
	doi
	release_year

role	
FK	project_id
FK	name_id
	role_label

related	
FK	output_id
FK	project_id

Streamlit

🗄 = 📈🔍

References

- Cheng, C., Messerschmidt, L., Bravo, I. et al. A General Primer for Data Harmonization. *Sci Data* 11, 152 (2024). https://doi.org/10.1038/s41597-024-02956-3.
- Community for Data Integration (CDI). (2025). What We Do, https://www.usgs.gov/centers/community-for-data-integration-cdi/what-we-do.
- Community for Data Integration (CDI). (2025). Participate, https://www.usgs.gov/centers/community-for-data-integration-cdi/participate.
- Huber, R., D'Onofio, C, Devaraju, A., et al. (2021 Integrating data and analysis technologies within leading environmental research infrastructures: Challenges and approaches, *Ecological Informatics*, https://doi.org/10.1016/j.ecoinf.2021.101245.
- Liu, D., and Yoon, V. (2024). Developing a goal-driven data integration framework for effective data analytics, *Decision Support Systems*, https://doi.org/10.1016/j.dss.2024.114197.
- Mirza, F. (2021). Integrating with Various Data Sources and Formats, Including Structured, Semi-Structured, and Unstructured Data, Journal of Scientific and Engineering Research, 8(2):263-268, https://jsaer.com/download/vol-8-iss-2-2021/JSAER2021-8-2-263-268.pdf.
- Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18/
- Ziegler, P., Dittrich, K.R. (2007). Data Integration — Problems, Approaches, and Perspectives. In: Krogstie, J., Opdahl, A.L., Brinkkemper, S. (eds) Conceptual Modelling in Information Systems Engineering. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-72677-7_3,

Thanks!