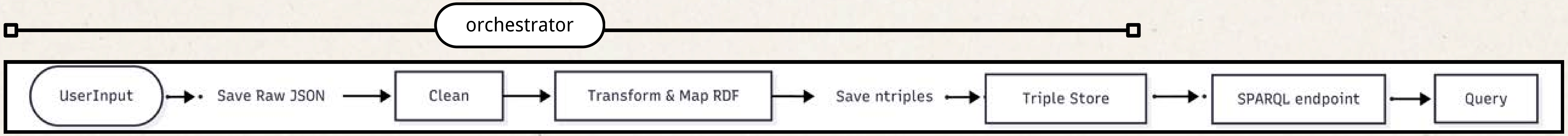
SafeVoice

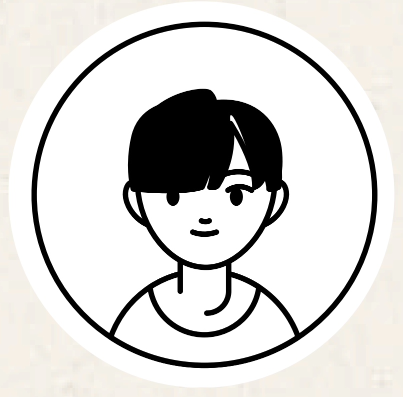
Human Trafficking Source Information App

DSIP Course Fieldlab 5

Overview

SafeVoice ensures sensitive human trafficking source data is secure, transforming it into a FAIR-compliant semantic format to be accessed trough federated queries.





Anh Trinh (s4870085)

Data Input UI

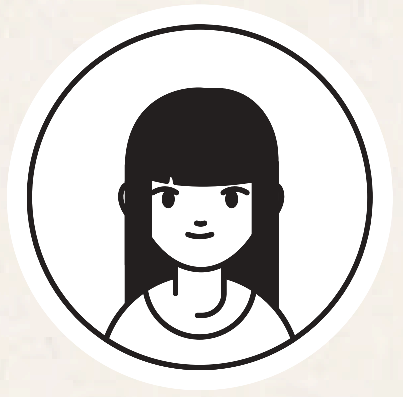
Wan Rong Shen (s240004)

Data Clean & Allegrograph ingest

Orchestrating ETL pipeline

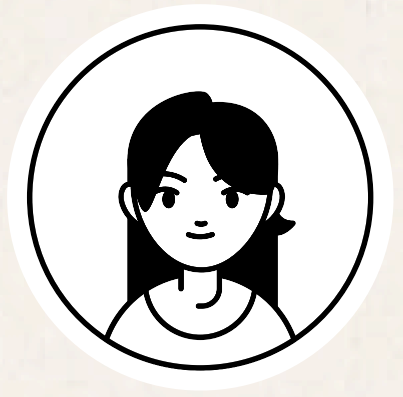
Mihir Ramani (s4841980)

CDM & RDF mapping



Nurifeiya Anwaier (s4299787)

SPARQL endpoint & UI

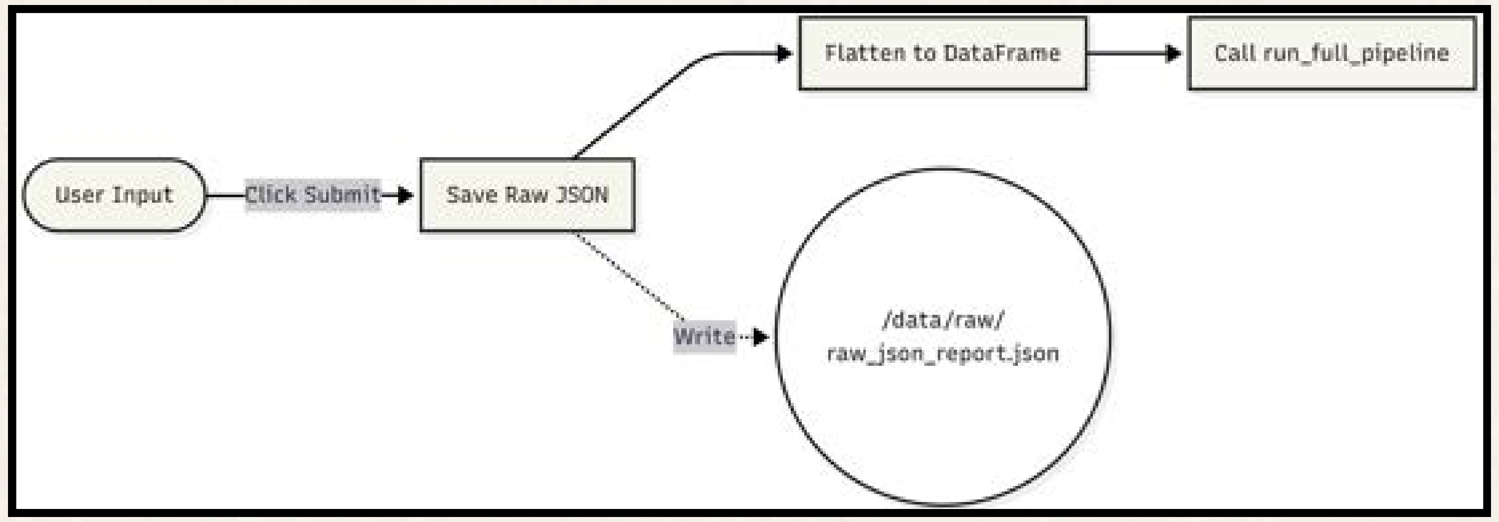


Xiong Xiatong (s4196511)

SPARQL query

Secure & Standardized Data Ingestion User Interface

- **Key Technical Features:**
  - Modern Web Stack: Built on Streamlit, render as standard HTML/JavaScript compatible with modern browsers.
  - Centralized Deployment: Can be hosted on a Linux-based server, enabling distribution via URL without local installation.
  - CDM-driven: The form logic is not hard-coded, but dynamically generated from the Common Data Model file.
  - Uniform Data Serialization: Submissions are automatically standardized into structured JSON formats, ensuring immediate readiness for analysis.



Secure & Standardized Data Ingestion User Interface

- **Challenge:**
  - Manual data entry is prone to typos, missing fields, inconsistencies.
  - Poor quality input breaks downstream pipelines and databases.
- **Solution:**
  - Controlled Vocabulary: Dropdowns powered directly by the Common Data Model to prevent non-standard terms.
  - Context-aware Validation: Dynamic forms adapt to input (certain dropdowns are available when certain options are locked in.
  - Built-in Tooltips: Official definitions from CDM appear as tooltips, helping clerks understand what to report.

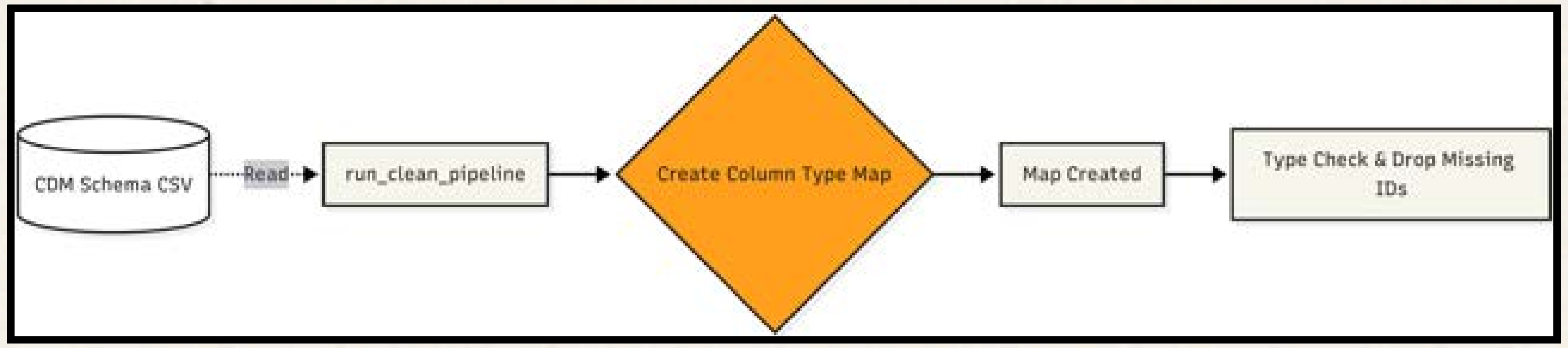
Secure & Standardized Data Ingestion User Interface

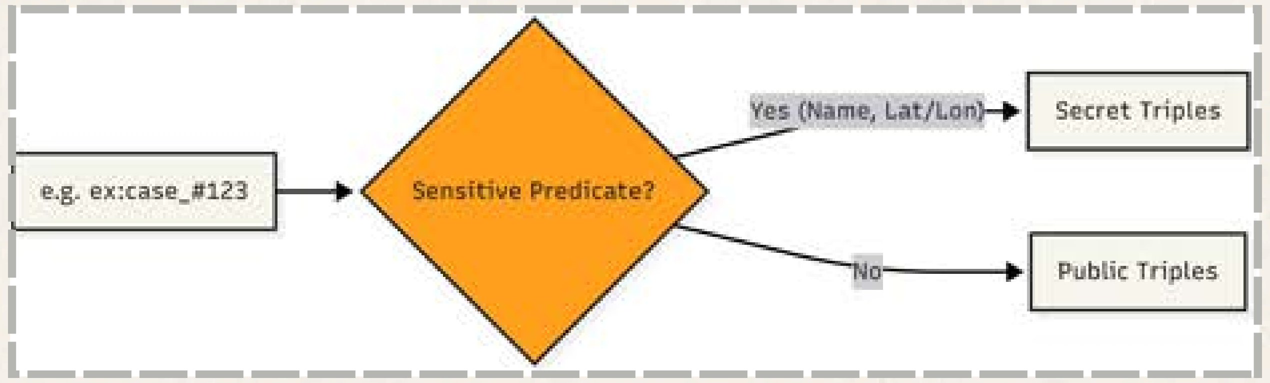
- **Challenge:**
    - Re-entering the same incident details for every victim is tedious and prone to error.
    - Sensitive human trafficking data must be strictly isolated.
- **Solution:**
    - Batch Mode: A tickbox feature to preserve incident details while resetting victim field.
    - User-based Access Control:
        - Secure Auth: Login system tags every record with specific Clerk ID and Org.
        - Immutable Logs: Clerks are able to do Read-Only on data they have input, to ensure reliable compliance to data integrity.
    -

Automate Data Cleaning Process

Once a JSON report record is created by a source, pipeline automatically triggers the data cleaning process, and returns a cleaned pandas dataframe

- **Key Technical Features:**
  - CDM-based column type mapping: cast string values to the correct data types.
  - Missing values handling: detect unknown descriptions and drop records with missing critical ID columns.
- **Challenge:**
  - De-identifying and anonymising sensitive information (e.g. latitude/longitude) would be unretrievable
- **Solution (to be implemented):**
  - Do not round or discard sensitive values in this step.
  - Instead, make them as separate secret triples, only exposing to organizations.





CDM & Ontology Layer

- We model EEPA data using the DSIP Common Data Model (CDM) extended with a Humanitarian Data Schema (HDS).

- The main class is cdm:HumanTraffickingCase, defined as a subclass of:
  hds:Incident
  hds:Event

- Supporting entities include:
 Victim, Trafficker, Perpetrator, Location, Organisation, Publication, Situation

- These are aligned with:
  Schema.org, FOAF, PROV-O, and NCI ontologies

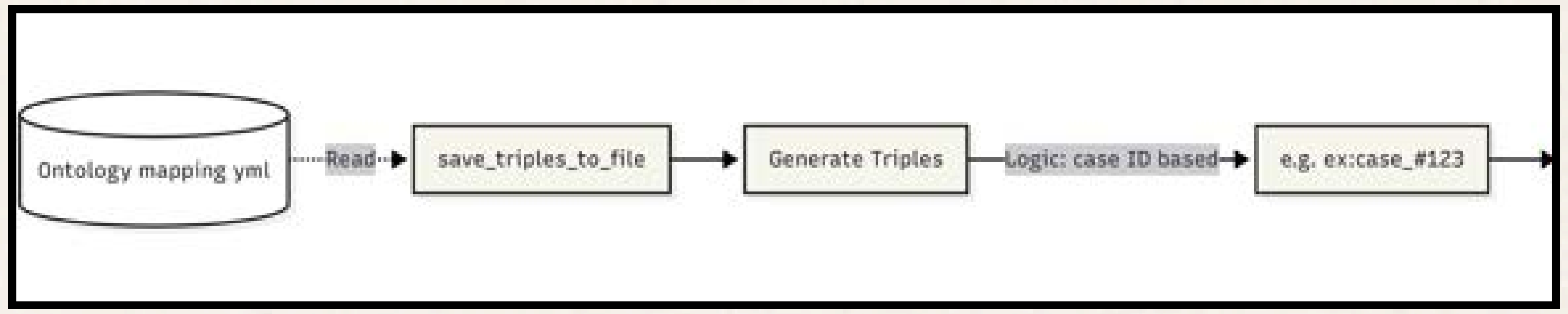
- This step gives the EEPA data a formal semantic structure instead of flat CSV rows.

Automate Mapping and Conversion Process

- The pandas-dataframe-to-RDF transformation is defined in a YARRRML mapping file.

- The mapping specifies:

1. Subject URI templates
2. Predicate-object rules
3. CDM and Schema.org properties



- A custom Python script:

1. Loads the YARRRML file
2. Validates that all mapped columns exist
3. Replaces $(ColumnName) placeholders
4. Expands prefixes (cdm:, schema: → full IRIs)
5. Each row is converted into multiple RDF triples in N-Triples format.

Generated RDF Output

```
<.../case_123> rdf:type cdm:HumanTraffickingCase .
<.../case_123> cdm:destination "Country" .
<.../case_123> schema:fromLocation <.../departure_place_45> .
<.../victim_place_45> schema:latitude "12.345"^^xsd:float .
```

- One EEPA record becomes:
  a. A semantic trafficking case
  b. Linked to locations and victim places
- The output forms a linked knowledge graph
- This RDF can be directly used for:
  a. SPARQL queries
  b. Graph analysis
  c. Knowledge integration

RDF mapping logic

Our RDF mapping logic converts data into a structured, FAIR-compliant knowledge graph based on the given CDM.

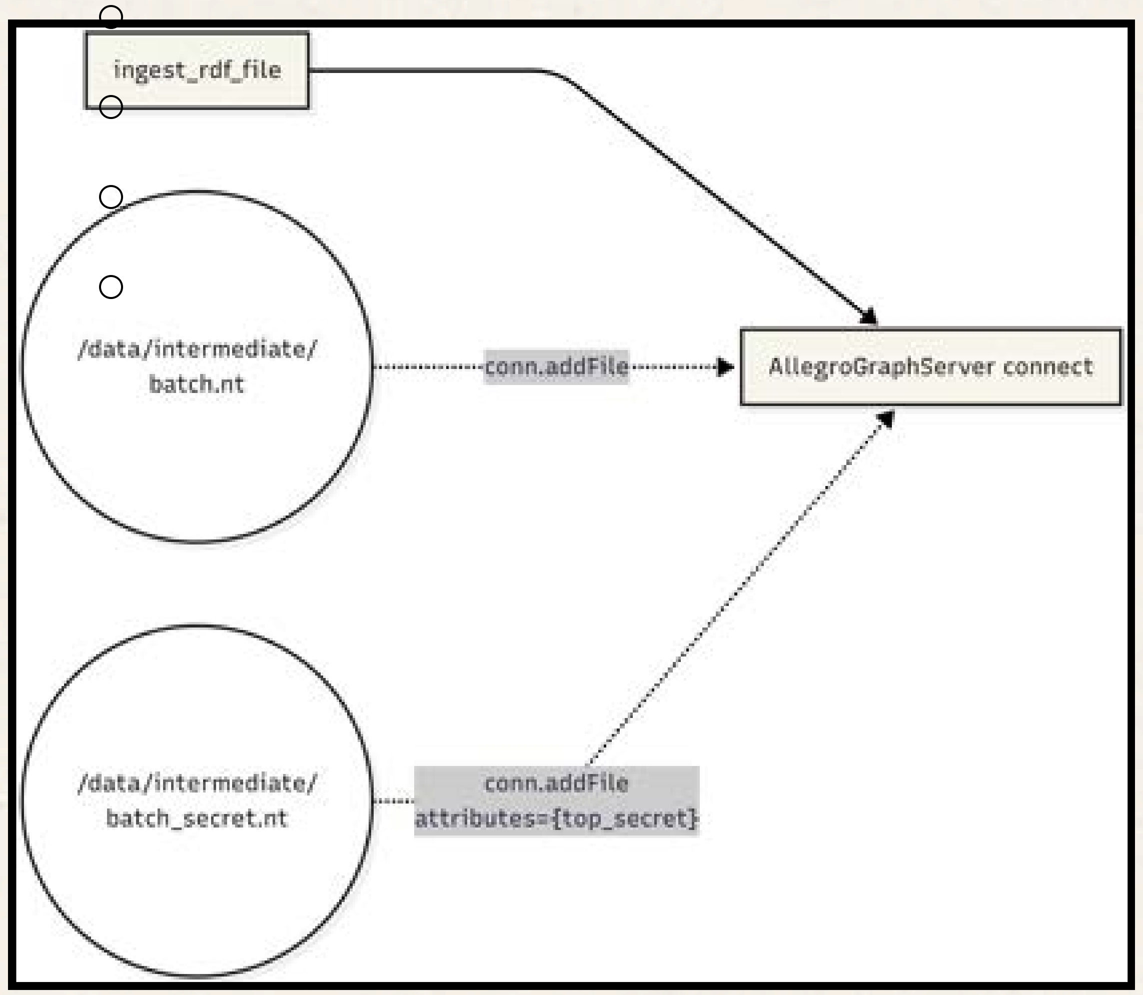
- We rely on standard prefixes (mainly schema:) and introduce two domain-specific prefixes:
    - **ex:**: Our Base Instance URI for generating globally unique IDs (e.g., ex:case_R123).
    - **cdm:**: Our custom CDM Ontology namespace, used for our domain-specific classes and relationships (e.g., cdm:HumanTraffickingCase).

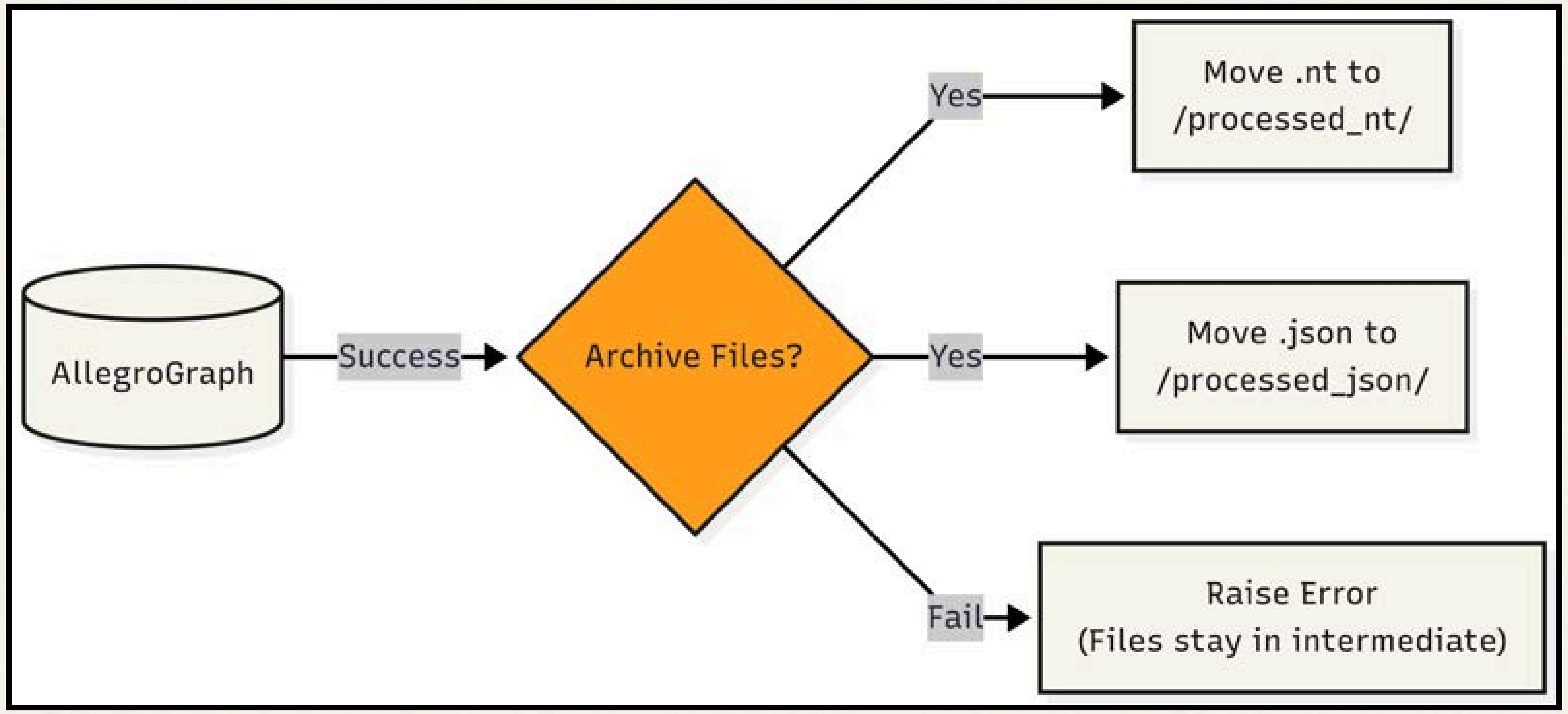
- Data Linking:
    - **Case serves as the central node for the report**
        - Linked to other peripheral nodes (e.g., Victim, Trafficker, Publication...)
    - Victim and Trafficker subjects:
        - **schema:Person**: Ensures they can be recognized as human and further ensures interoperability
        - **cdm:Victim** / **cdm:Trafficker**: Our domain-specific tag, allowing queries to specifically target the role of the individual
        - Attributes: Standard properties from Schema.org are used (e.g. schema:Age)

Data Ingestion to triplestore (Allegrograph)

Once a .nt file is created contiaining RDF triples, it's automatically ingesting to allegrograph with addFile

- **Key Technical Features:**
  - <u>Data file staging:</u> files are separated into /raw, /intermediate, and /processed stages. A file only moves to processed after it is successfully written to AllegroGraph, preventing JSON/NT files from being re-ingested or reprocessed in case of server connection failures.


ingest_rdf_file
/data/intermediate/ batch.nt
conn.addFile
AllegroGraphServer connect
/data/intermediate/ batch_secret.nt
conn.addFile attributes={top_secret}


AllegroGraph
Success
Archive Files?
Yes
Move .nt to /processed_nt/
Yes
Move .json to /processed_json/
Fail
Raise Error (Files stay in intermediate)

SPARQL Endpoint Integration

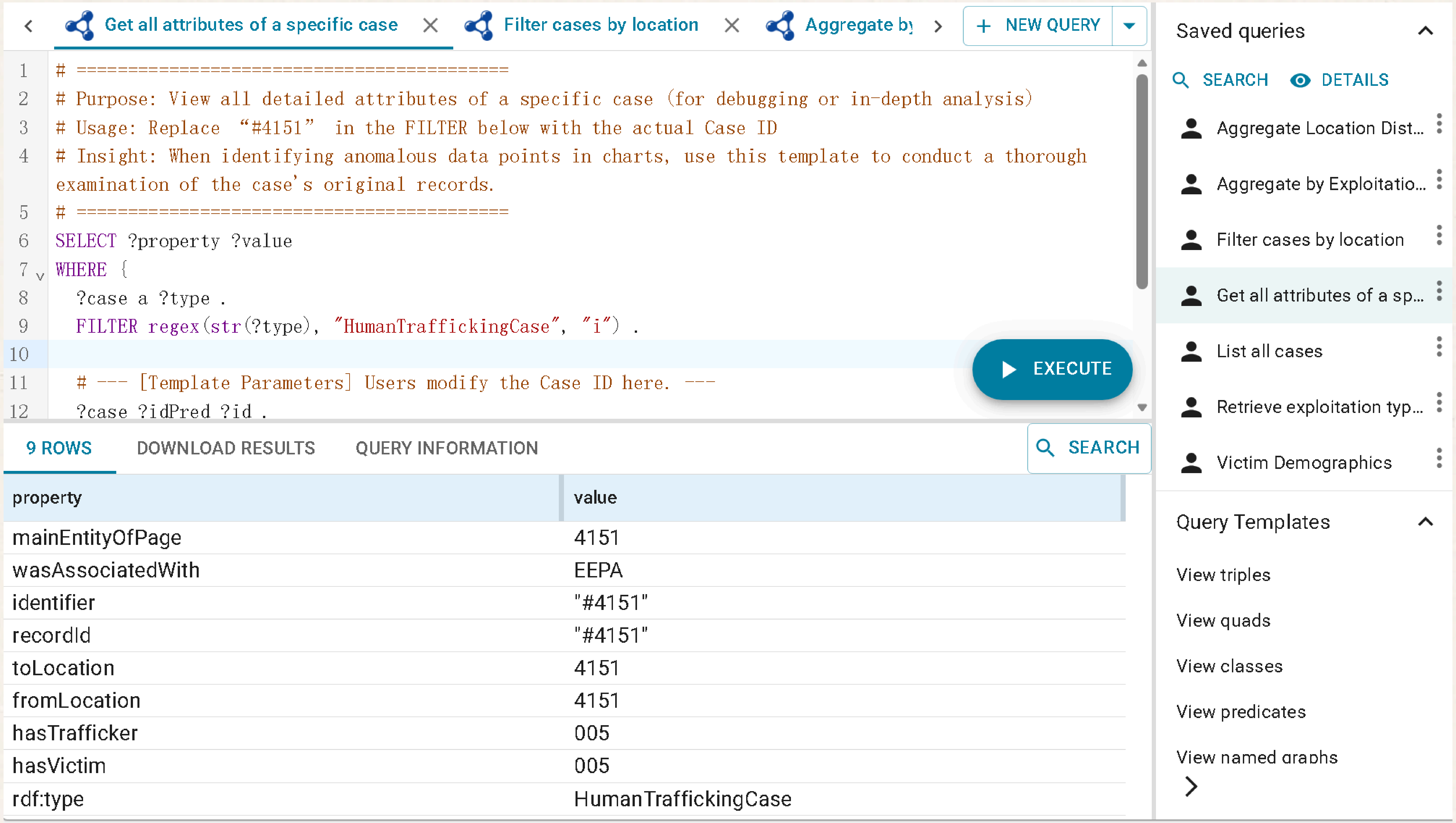
•Created a SPARQL endpoint for the RDF graph stored in AllegroGraph.

•Endpoint URL:
https://ag1950eewddzjs9z.allegrograph.cloud/repositories/trafficking_repo

• Validated that the endpoint returns correct SPARQL query results.

•Provided access credentials for FieldLab 7

Query Tamplates

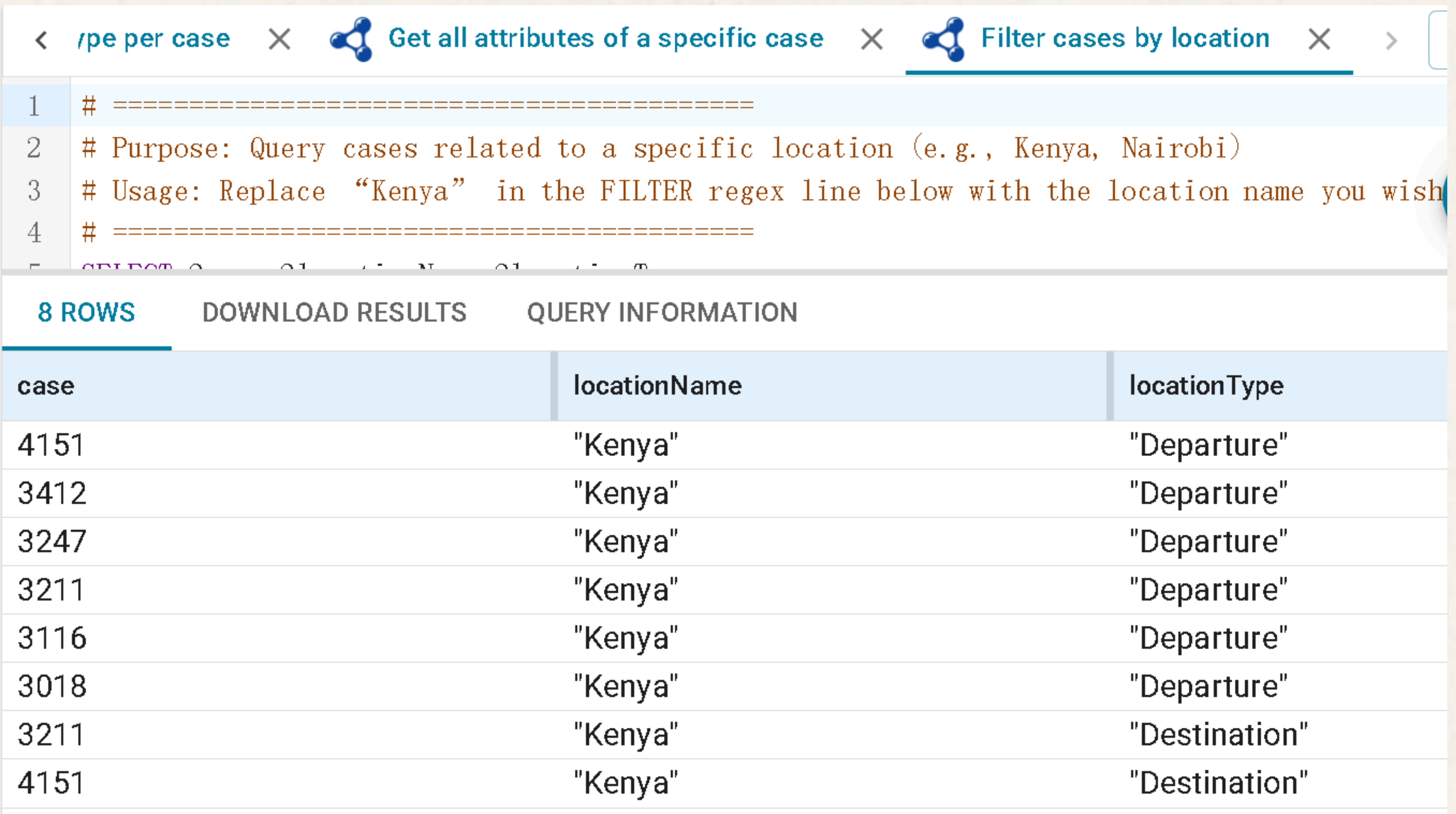


```
# ========================================
# Purpose: View all detailed attributes of a specific case (for debugging or in-depth analysis)
# Usage: Replace "#4151" in the FILTER below with the actual Case ID
# Insight: When identifying anomalous data points in charts, use this template to conduct a thorough
examination of the case's original records.
# ========================================
SELECT ?property ?value
WHERE {
  ?case a ?type .
  FILTER regex(str(?type), "HumanTraffickingCase", "i") .

  # --- [Template Parameters] Users modify the Case ID here. ---
  ?case ?idPred ?id .
```

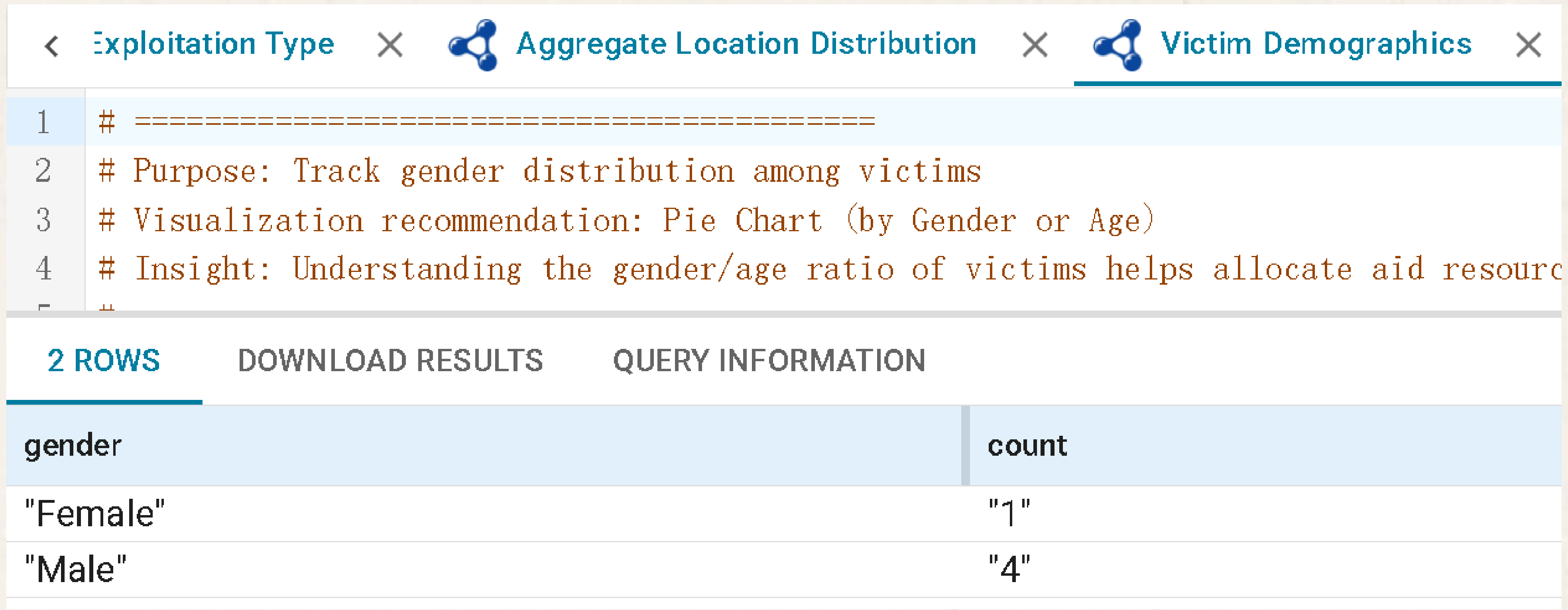
property	value
mainEntityOfPage	4151
wasAssociatedWith	EEPA
identifier	"#4151"
recordId	"#4151"
toLocation	4151
fromLocation	4151
hasTrafficker	005
hasVictim	005
rdf:type	HumanTraffickingCase

Query Results Examples



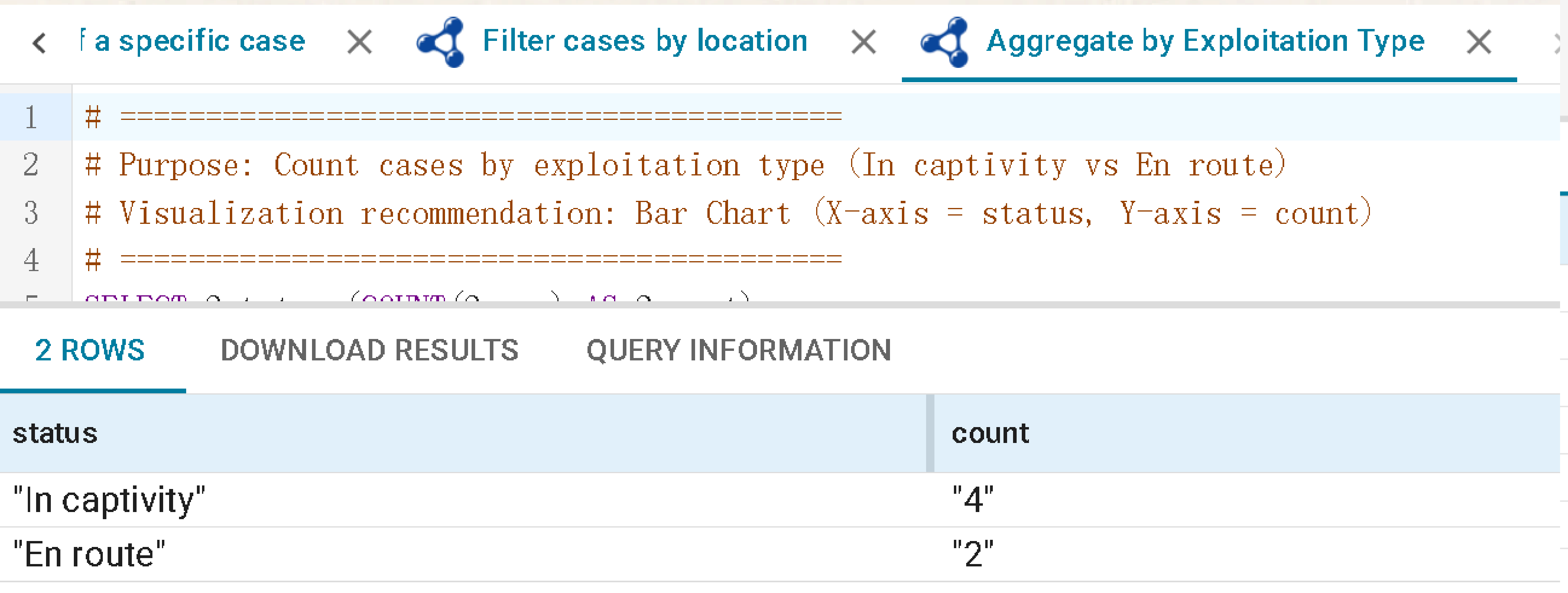
```
# ========================================
# Purpose: Query cases related to a specific location (e.g., Kenya, Nairobi)
# Usage: Replace "Kenya" in the FILTER regex line below with the location name you wish
# ========================================
```

case	locationName	locationType
4151	"Kenya"	"Departure"
3412	"Kenya"	"Departure"
3247	"Kenya"	"Departure"
3211	"Kenya"	"Departure"
3116	"Kenya"	"Departure"
3018	"Kenya"	"Departure"
3211	"Kenya"	"Destination"
4151	"Kenya"	"Destination"



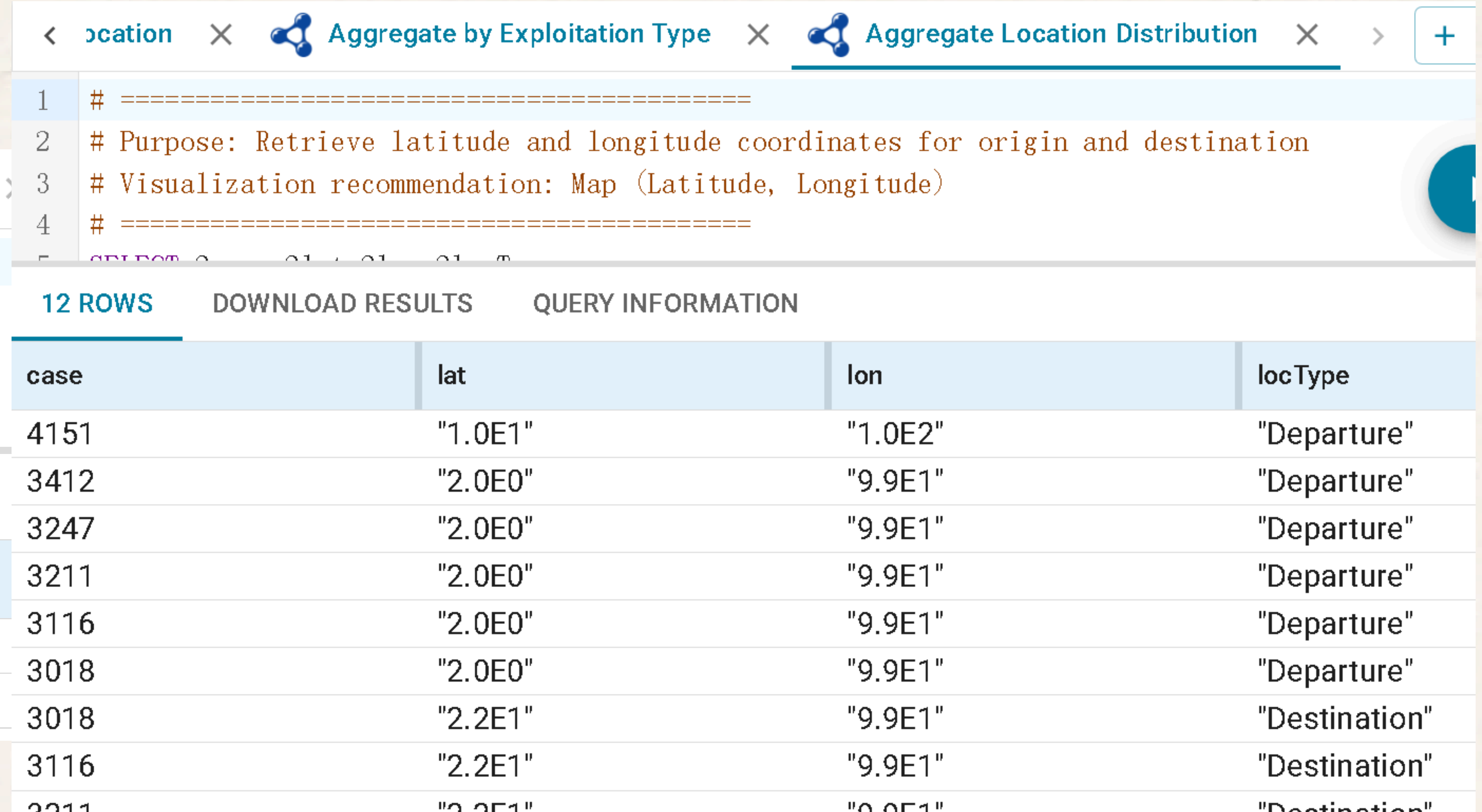
```
# ========================================
# Purpose: Track gender distribution among victims
# Visualization recommendation: Pie Chart (by Gender or Age)
# Insight: Understanding the gender/age ratio of victims helps allocate aid resourc
```

gender	count
"Female"	"1"
"Male"	"4"



```
# ========================================
# Purpose: Count cases by exploitation type (In captivity vs En route)
# Visualization recommendation: Bar Chart (X-axis = status, Y-axis = count)
# ========================================
```

status	count
"In captivity"	"4"
"En route"	"2"


ocation
Aggregate by Exploitation Type
Aggregate Location Distribution
12 ROWS
DOWNLOAD RESULTS
QUERY INFORMATION

```
# ========================================
# Purpose: Retrieve latitude and longitude coordinates for origin and destination
# Visualization recommendation: Map (Latitude, Longitude)
# ========================================
```

case	lat	lon	locType
4151	"1.0E1"	"1.0E2"	"Departure"
3412	"2.0E0"	"9.9E1"	"Departure"
3247	"2.0E0"	"9.9E1"	"Departure"
3211	"2.0E0"	"9.9E1"	"Departure"
3116	"2.0E0"	"9.9E1"	"Departure"
3018	"2.0E0"	"9.9E1"	"Departure"
3018	"2.2E1"	"9.9E1"	"Destination"
3116	"2.2E1"	"9.9E1"	"Destination"
3211	"2.2E1"	"9.9E1"	"Destination"

Future Challenges

Ontology & Interoperability

- More data sources lead to increased inconsistency across CDM and mappings， which requires continuous validation and update cycles

Access Control & Ethics

- Current manual permission system does not scale

Metadata

- We still need to create DCAT metadata, the semantic metadata model for the FAIR Data Point (FDP)

Data Containers

- Performance concerns for cross-repository access. Because more repositories lead to higher complexity in the system

User-friendly Inplementation

- More intuitive interfaces required for Non-technical users
- A user-friendly guide explaining how our pipeline works and how to use it.

Thank you!

WE WOULD LIKE TO EXPRESS OUR SINCERE GRATITUDE TO EVERYONE WHO SUPPORTED THIS PROJECT, INCLUDING:

Mirjam Liam Philip Joelle Daniel Onesmus Kyle

All FieldLab mentors and team members