Extralit Project Knowledge Handoff

Jonny T., Amelia BV., Kate B., Caitlin B.

Overview

- What is extralit, what you can do with it?
- Concepts in data extraction with extralit for the users
- Walk through of the tool
- Project maintenance and resources
- Future contact points for extralit open-source

ITN-recal-data-extraction

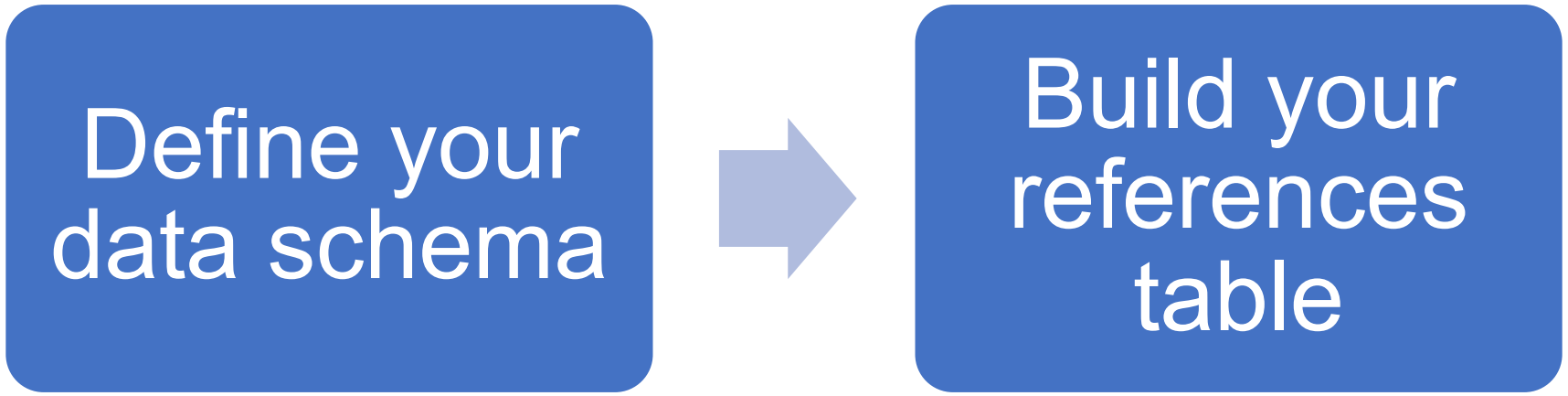
The project aims to automate and streamline the extraction of ITN efficacy and mosquito outcomes data from a large volume of malaria literature. The system involves several key components:

1. PDF Preprocessing: Detection and extraction of table structures in PDF documents.
2. Schema-driven Extraction: Using predefined schemas to accurately extract relevant data fields.
3. Human-in-the-loop: Manual data extraction steps to verify and correct automated extractions through an open-source web interface built from the Argilla project .
4. Microservices Orchestration: Managing the entire data storage, processing servers, and other services on Kubernetes.

Key Features

- Schema-driven extraction: Ensures high specificity, contextual relevance, and automated validation of the extracted data.
- Advanced PDF preprocessing: AI optical character recoginition (OCR) algorithms to detect and correct table structures within documents.
- User-friendly interface: Facilitates easy verification and correction of extracted data.
- Data flywheel: Continuous data collection of table extractions and LLM outputs to monitor performance and build datasets.

Starting a new extraction project



OBSERVATION TABLE
Observation_ref
Country
Site
Start_{Month, Year}
End_{Month, Year}
Time_elapsed
Study_type

ENTOMOLOGICAL OUTCOME TABLE
Observation_ref
ITNCondition_ref
Anoph_spp
Mosquito_age
Source
Measured_outcome
Total_mosquitos
Dead
Mortality_{rate, lower, upper}
KD_{constant, count, rate, time, lower, upper}
Blood_fed_{count, rate, lower, upper, inhibition}
Endo_{net, control}
Deter_{rate, lower, upper}
Penetrate_{N, rate, lower, upper}
Parity_{rate, lower, upper}
Spor_{pos, rate, lower, upper}

CLINICAL OUTCOME TABLE
Observation_ref
ITNCondition_ref
N_people
Age_{lower, upper}
N_pos
PR_{rate, lower, upper}
CM_{count, rate, lower, upper}
Net_{retention, count, sleep_nets, sleep_pct}

ITN TABLE
ITNCondition_ref
Net_type
Insecticide
Net_washed
Net_age
Net_holes
pHI_{category, median, lower_IQR, upper_IQR}

Schemas

reference	title	pmid	file_path
abdella2009does	Does Insecticide Treated Mosquito Nets (ITNs) ...	18958607	data/pdf/Adbella_et_al_2009_J_Community_Health...
abdulla2001impact	Impact on malaria morbidity of a programme sup...	11157527	data/pdf/Abdulla_et_al___2001___Impact_on_mala...
abdulla2005spatial	Spatial effects of the social marketing of ins...	None	data/pdf/abdulla2005spatial.pdf
abilio2015bio	Bio-efficacy of new long-lasting insecticide-t...	None	data/pdf/12936_2015_Article_885_pdf.pdf
agossa2014laboratory	Laboratory and field evaluation of the impact ...	24884502	data/pdf/Agossa_et_al_2014_Mal_J.pdf
...	...	...	...
tungu2021effectiveness	Effectiveness of a long-lasting insecticide tr...	34412651	data/pdf/s12936_021_03871_3.pdf
tungu2021field	Field evaluation of Veeralin, an alpha-cyperme...	35284898	data/pdf/1_s20_S2667114X21000248_main.pdf
verma2022laboratory	Laboratory evaluation of a new alphacypermethr...	35876911	data/pdf/LLINPhaseIVV2022.pdf
yewhalaw2022experimental	An experimental hut study evaluating the impac...	35987650	data/pdf/An_experimental_hut_study_evaluating_...
zahouli2023small	Small-scale field evaluation of PermaNet(®) Du...	36726160	data/pdf/Small_scale_field_evaluation_of_Perma...

References table

Extralit

reference	user_name	Anoph_spp	Source	Total_mosquitoes	Mortality_rate	KD_constant
sreehari2009wash	jonnytr	An. culicifacies	Lab	11	100.0	quantile
sreehari2009wash	jonnytr	An. culicifacies	Lab	11	100.0	quantile
sreehari2009wash	jonnytr	An. culicifacies	Lab	11	100.0	quantile
sreehari2009wash	jonnytr	An. culicifacies	Lab	11	87.1	quantile
sreehari2009wash	jonnytr	An. culicifacies	Lab	11	82.0	quantile
...	...	...	...	...	...	...
azizi2022implementing	ameliabv	An. gambiae	Lab	398	50	NaN
azizi2022implementing	ameliabv	An. gambiae	Lab	398	100	NaN
azizi2022implementing	ameliabv	An. gambiae	Lab	398	75	NaN

Extraction output

Short demo walkthrough

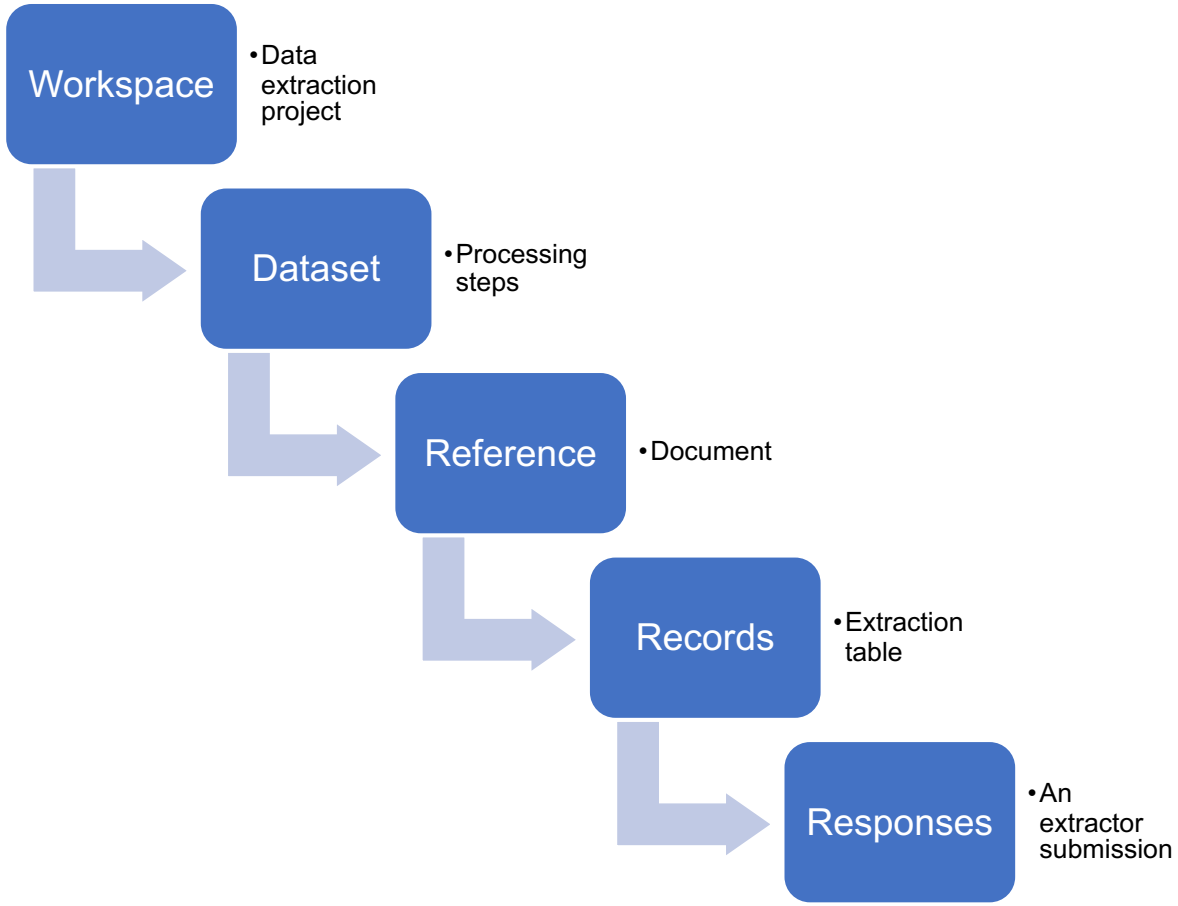
1. Text and table sections extracted from a PDF document
2. Sections uploads to Weaviate vector database, with metadata and content
3. User can select specific document sections for their LLM queries to extract into a table

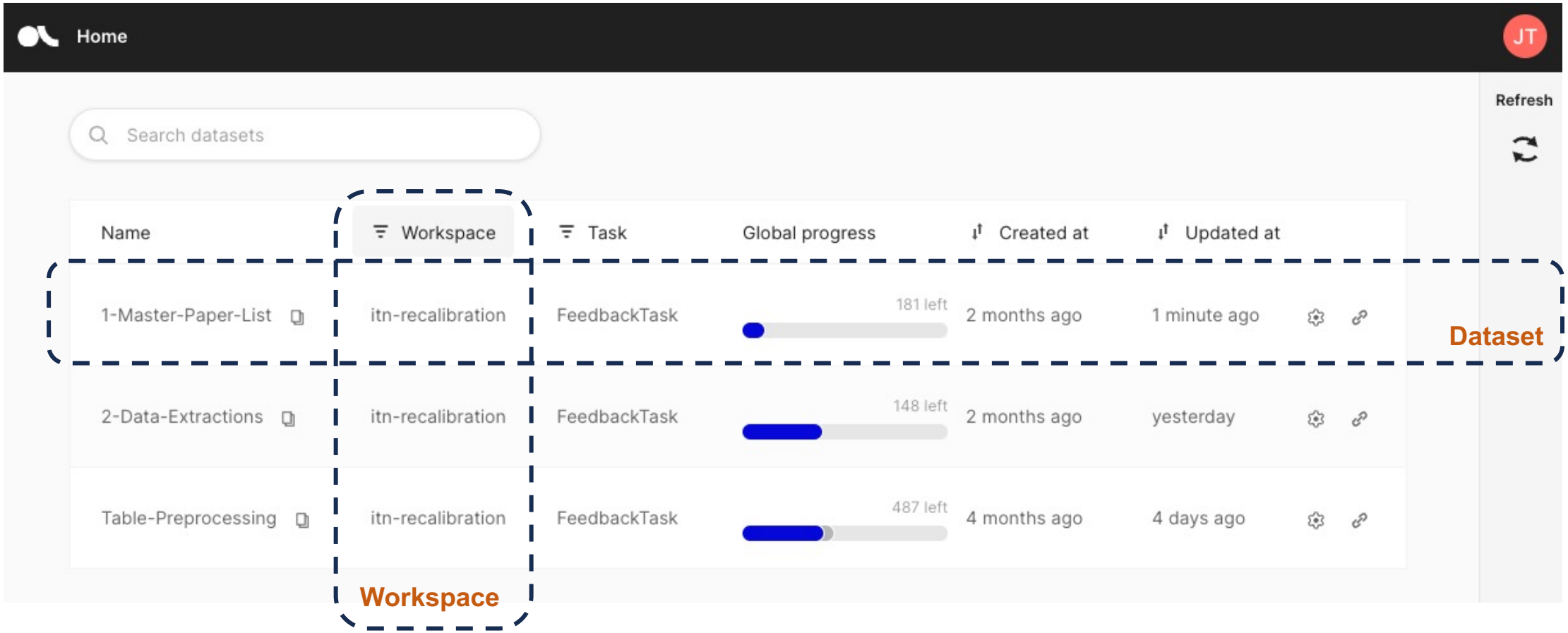
Extralit data hierarchy

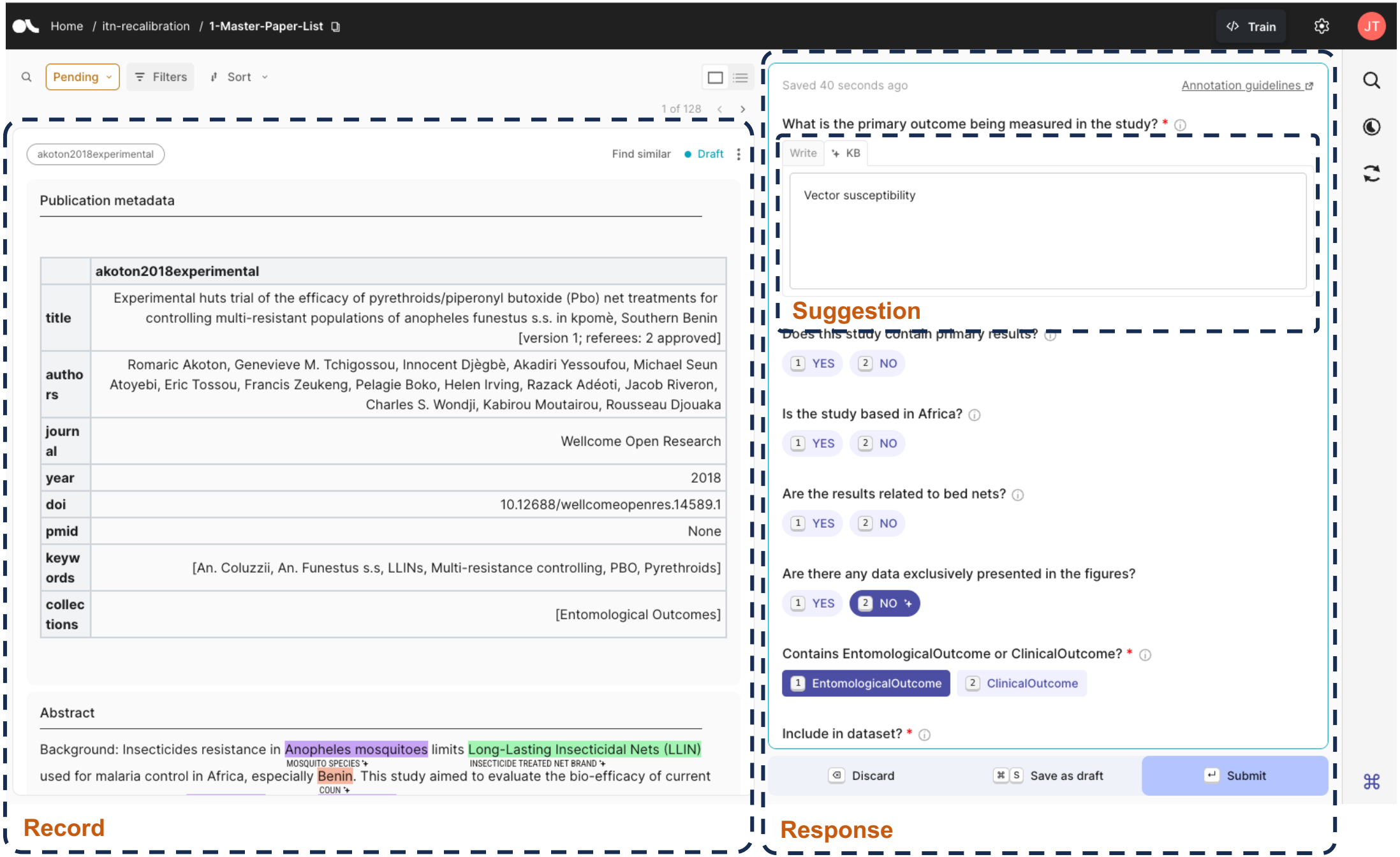
A "Workspace" organizes documents and extractions into multiple processing steps as “Datasets”.

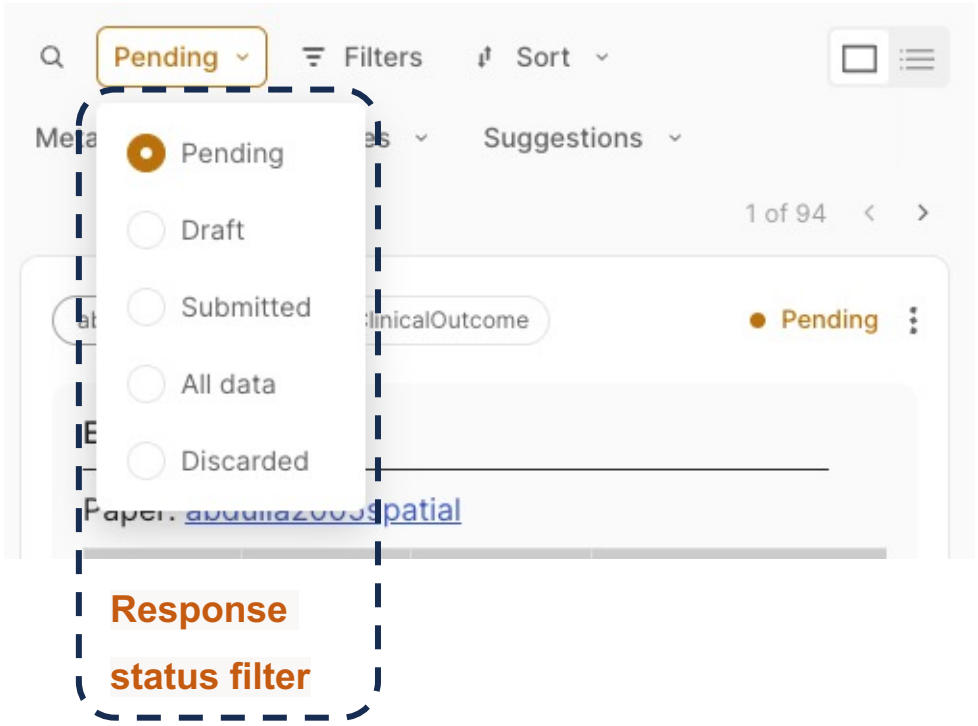
At the extraction process, each document “Reference” have multiple extraction “Records” for each schema table.

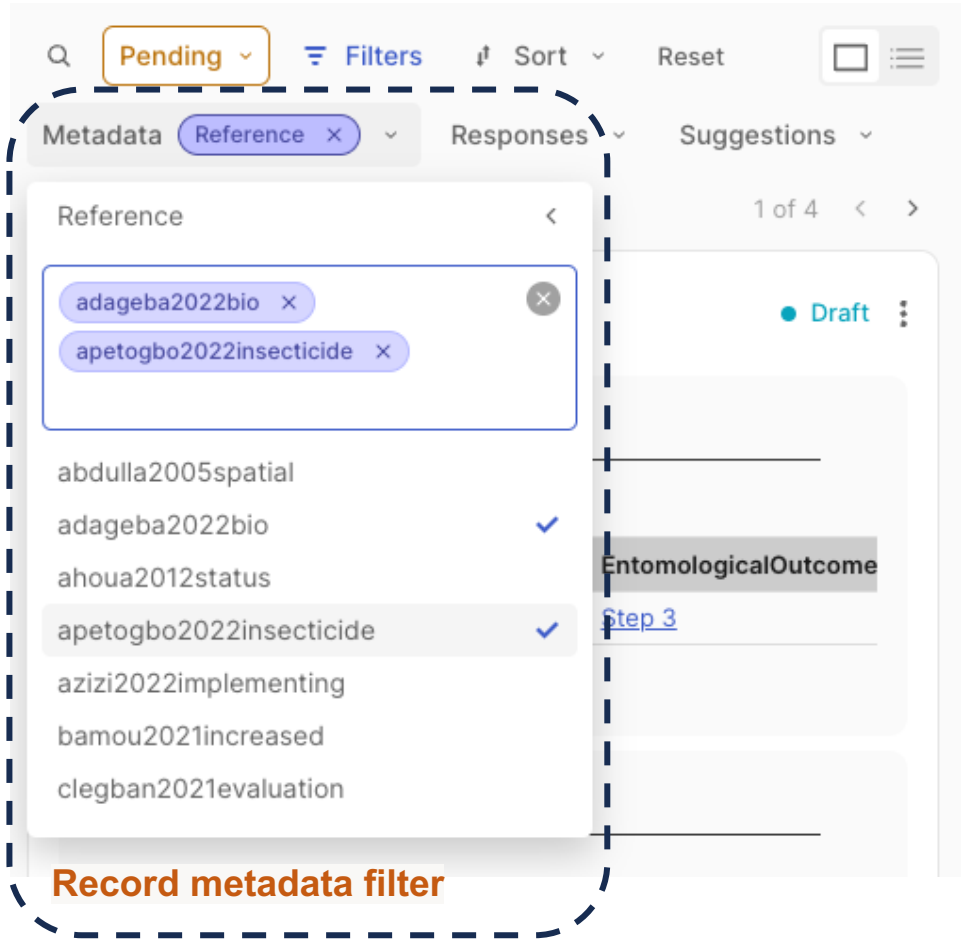
As extractors submit extraction tables, a reviewer can validate for correctness, then export data.

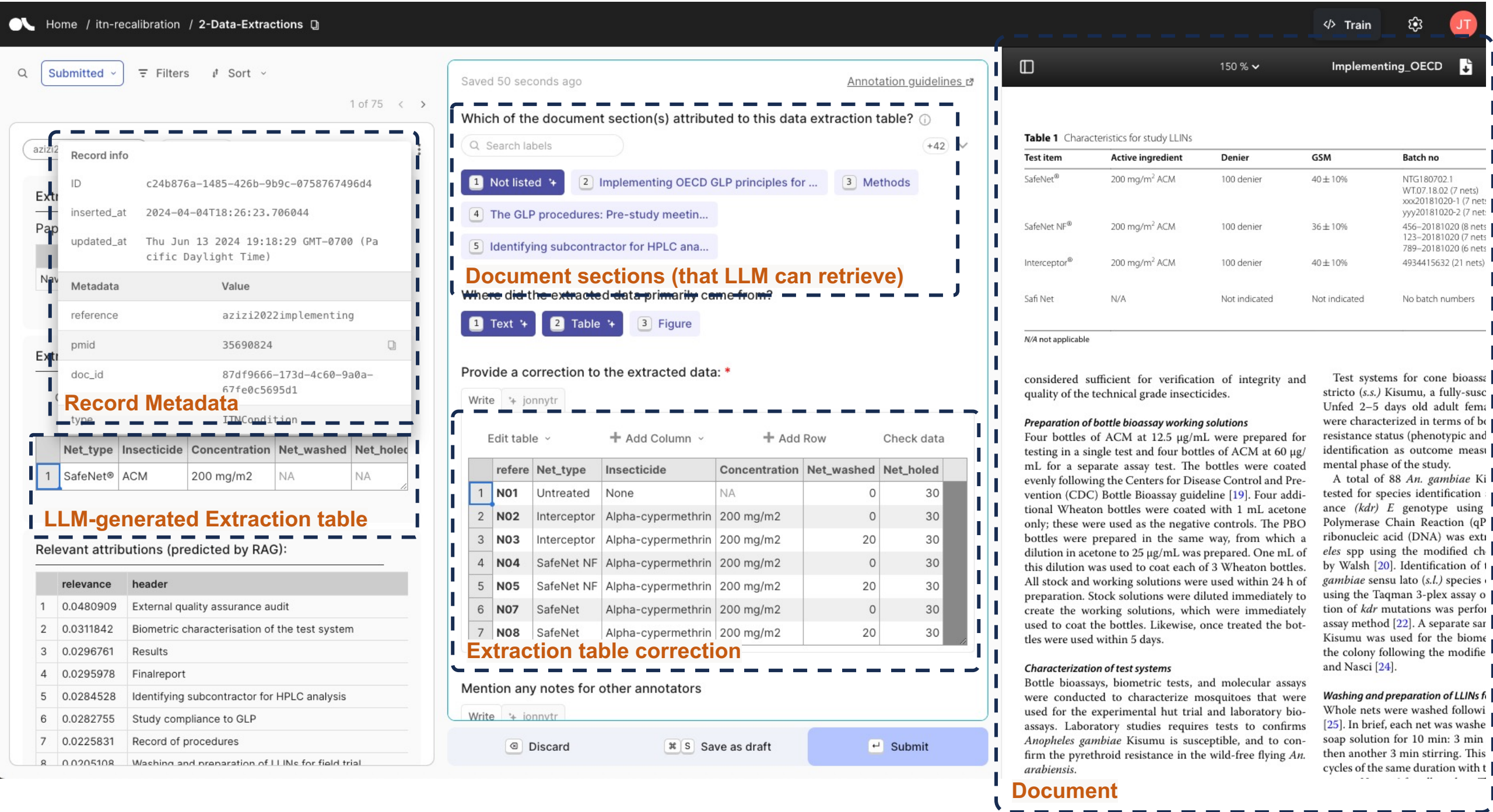



Home
JT
Refresh
Search datasets
Name
Workspace
Task
Global progress
Created at
Updated at
1-Master-Paper-List
itn-recalibration
FeedbackTask
181 left
2 months ago
1 minute ago
2-Data-Extractions
itn-recalibration
FeedbackTask
148 left
2 months ago
yesterday
Table-Preprocessing
itn-recalibration
FeedbackTask
487 left
4 months ago
4 days ago
Dataset
Workspace

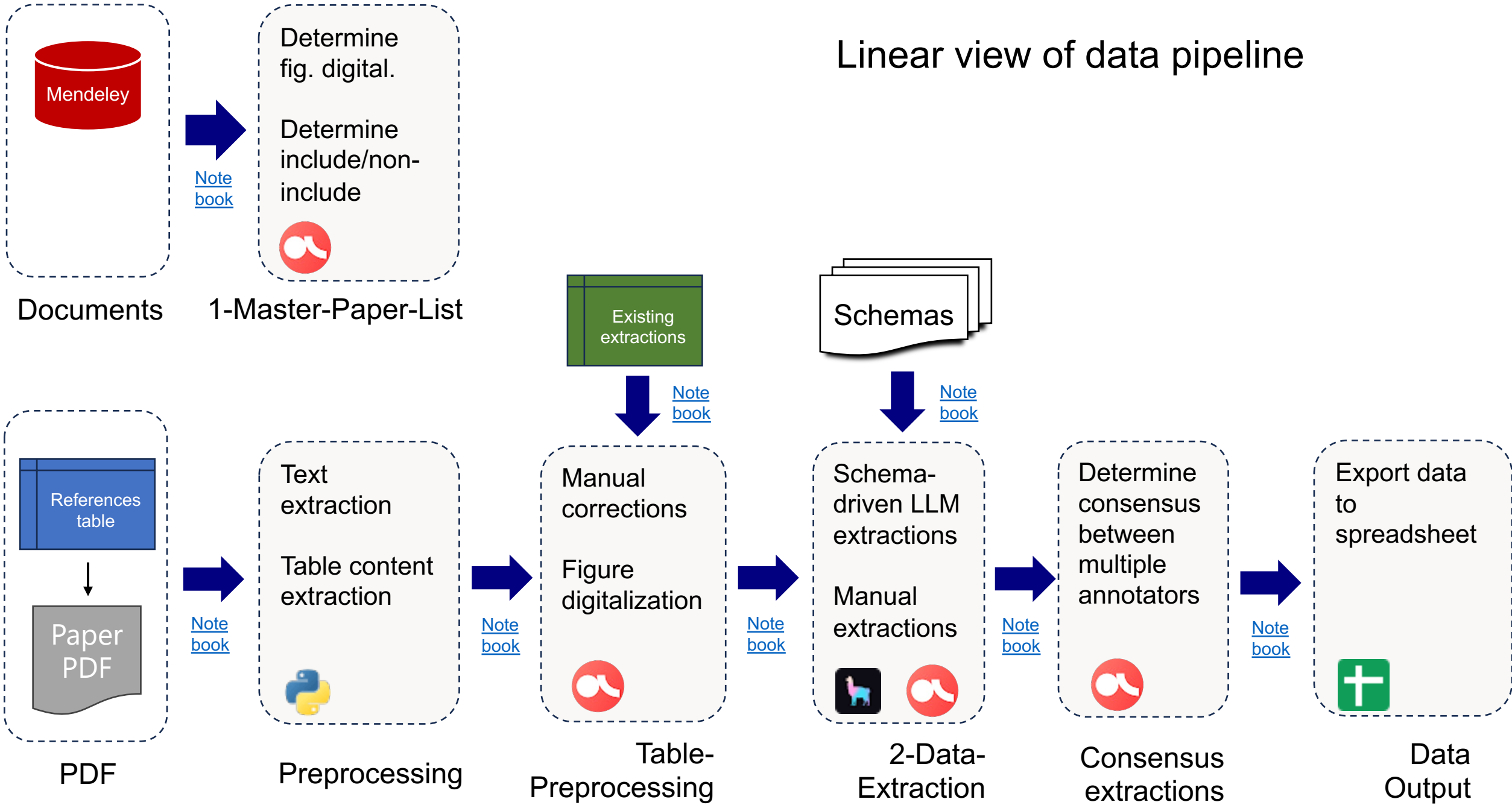

Home / itn-recalibration / 1-Master-Paper-List
Train
JT
Pending
Filters
Sort
1 of 128
akoton2018experimental
Find similar
Draft
Publication metadata
akoton2018experimental
title
Experimental huts trial of the efficacy of pyrethroids/piperonyl butoxide (Pbo) net treatments for controlling multi-resistant populations of anopheles funestus s.s. in kpomè, Southern Benin [version 1; referees: 2 approved]
authors
Romaric Akoton, Genevieve M. Tchigossou, Innocent Djègbè, Akadiri Yessoufou, Michael Seun Atoyebi, Eric Tossou, Francis Zeukeng, Pelagie Boko, Helen Irving, Razack Adéoti, Jacob Riveron, Charles S. Wondji, Kabirou Moutairou, Rousseau Djouaka
journal
Wellcome Open Research
year
2018
doi
10.12688/wellcomeopenres.14589.1
pmid
None
keywords
[An. Coluzzii, An. Funestus s.s, LLINs, Multi-resistance controlling, PBO, Pyrethroids]
collections
[Entomological Outcomes]
Abstract
Background: Insecticides resistance in Anopheles mosquitoes limits Long-Lasting Insecticidal Nets (LLIN) used for malaria control in Africa, especially Benin. This study aimed to evaluate the bio-efficacy of current
MOSQUITO SPECIES
INSECTICIDE TREATED NET BRAND
COUN
Record
Saved 40 seconds ago
Annotation guidelines
What is the primary outcome being measured in the study? *
Write
KB
Vector susceptibility
Suggestion
Does this study contain primary results?
1 YES
2 NO
Is the study based in Africa?
1 YES
2 NO
Are the results related to bed nets?
1 YES
2 NO
Are there any data exclusively presented in the figures?
1 YES
2 NO
Contains EntomologicalOutcome or ClinicalOutcome? *
1 EntomologicalOutcome
2 ClinicalOutcome
Include in dataset? *
Discard
S Save as draft
Submit
Response


Pending
Filters
Sort
Suggestions
1 of 94
Pending
Draft
Submitted
All data
Discarded
Pending
Response
status filter


Pending
Filters
Sort
Reset
Metadata
Reference
Responses
Suggestions
1 of 4
Reference
adageba2022bio
apetogbo2022insecticide
abdulla2005spatial
adageba2022bio
ahoua2012status
apetogbo2022insecticide
azizi2022implementing
bamou2021increased
clegban2021evaluation
Draft
EntomologicalOutcome
Step 3
Record metadata filter


Home / itn-recalibration / 2-Data-Extractions
Train
JT
Submitted
Filters
Sort
1 of 75
Record info
ID c24b876a-1485-426b-9b9c-0758767496d4
inserted_at 2024-04-04T18:26:23.706044
updated_at Thu Jun 13 2024 19:18:29 GMT-0700 (Pacific Daylight Time)
Metadata Value
reference azizi2022implementing
pmid 35690824
doc_id 87df9666-173d-4c60-9a0a-67fe0c5695d1
type ITNCondition
Record Metadata
Net_type Insecticide Concentration Net_washed Net_holed
1 SafeNet® ACM 200 mg/m2 NA NA
LLM-generated Extraction table
Relevant attributions (predicted by RAG):
relevance header
1 0.0480909 External quality assurance audit
2 0.0311842 Biometric characterisation of the test system
3 0.0296761 Results
4 0.0295978 Finalreport
5 0.0284528 Identifying subcontractor for HPLC analysis
6 0.0282755 Study compliance to GLP
7 0.0225831 Record of procedures
Saved 50 seconds ago
Annotation guidelines
Which of the document section(s) attributed to this data extraction table?
Search labels
+42
1 Not listed
2 Implementing OECD GLP principles for ...
3 Methods
4 The GLP procedures: Pre-study meetin...
5 Identifying subcontractor for HPLC ana...
Document sections (that LLM can retrieve)
Where did the extracted data primarily come from?
1 Text
2 Table
3 Figure
Provide a correction to the extracted data: *
Write
jonnytr
Edit table
Add Column
Add Row
Check data
refere Net_type Insecticide Concentration Net_washed Net_holed
1 N01 Untreated None NA 0 30
2 N02 Interceptor Alpha-cypermethrin 200 mg/m2 0 30
3 N03 Interceptor Alpha-cypermethrin 200 mg/m2 20 30
4 N04 SafeNet NF Alpha-cypermethrin 200 mg/m2 0 30
5 N05 SafeNet NF Alpha-cypermethrin 200 mg/m2 20 30
6 N07 SafeNet Alpha-cypermethrin 200 mg/m2 0 30
7 N08 SafeNet Alpha-cypermethrin 200 mg/m2 20 30
Extraction table correction
Mention any notes for other annotators
Write
jonnytr
Discard
Save as draft
Submit
150 %
Implementing_OECD
Table 1 Characteristics for study LLINs
Test item Active ingredient Denier GSM Batch no
SafeNet® 200 mg/m² ACM 100 denier 40 ± 10% NTG180702.1 WT.07.18.02 (7 nets) xxx20181020-1 (7 nets yyy20181020-2 (7 net
SafeNet NF® 200 mg/m² ACM 100 denier 36 ± 10% 456–20181020 (8 nets 123–20181020 (7 nets 789–20181020 (6 nets
Interceptor® 200 mg/m² ACM 100 denier 40 ± 10% 4934415632 (21 nets)
Safi Net N/A Not indicated Not indicated No batch numbers
N/A not applicable
considered sufficient for verification of integrity and quality of the technical grade insecticides.
Preparation of bottle bioassay working solutions
Four bottles of ACM at 12.5 µg/mL were prepared for testing in a single test and four bottles of ACM at 60 µg/mL for a separate assay test. The bottles were coated evenly following the Centers for Disease Control and Prevention (CDC) Bottle Bioassay guideline [19]. Four additional Wheaton bottles were coated with 1 mL acetone only; these were used as the negative controls. The PBO bottles were prepared in the same way, from which a dilution in acetone to 25 µg/mL was prepared. One mL of this dilution was used to coat each of 3 Wheaton bottles. All stock and working solutions were used within 24 h of preparation. Stock solutions were diluted immediately to create the working solutions, which were immediately used to coat the bottles. Likewise, once treated the bottles were used within 5 days.
Characterization of test systems
Bottle bioassays, biometric tests, and molecular assays were conducted to characterize mosquitoes that were used for the experimental hut trial and laboratory bioassays. Laboratory studies requires tests to confirms Anopheles gambiae Kisumu is susceptible, and to confirm the pyrethroid resistance in the wild-free flying An. arabiensis.
Test systems for cone bioassa
stricto (s.s.) Kisumu, a fully-susc
Unfed 2–5 days old adult fema
were characterized in terms of bo
resistance status (phenotypic and
identification as outcome measu
mental phase of the study.
A total of 88 An. gambiae Ki
tested for species identification
ance (kdr) E genotype using
Polymerase Chain Reaction (qP
ribonucleic acid (DNA) was extr
eles spp using the modified ch
by Walsh [20]. Identification of
gambiae sensu lato (s.l.) species
using the Taqman 3-plex assay o
tion of kdr mutations was perfor
assay method [22]. A separate sar
Kisumu was used for the biome
the colony following the modifie
and Nasci [24].
Washing and preparation of LLINs f
Whole nets were washed followi
[25]. In brief, each net was washe
soap solution for 10 min: 3 min
then another 3 min stirring. This
cycles of the same duration with t
Document

Linear view of data pipeline



A command-line interface coming soon next week

```
(base) jonnytr@BMGF-Y0F73D3YXV ITN-recal-data-extraction % extralit

 Usage: extralit [OPTIONS] COMMAND [ARGS]...

 Extralit CLI

╭─ Options ──────────────────────────────────────────────────────────────────────╮
│ --install-completion          Install completion for the current shell.        │
│ --show-completion             Show completion for the current shell, to copy   │
│                               it or customize the installation.                │
│ --help                        Show this message and exit.                      │
╰────────────────────────────────────────────────────────────────────────────────╯
╭─ Commands ─────────────────────────────────────────────────────────────────────╮
│ datasets          Commands for dataset management                              │
│ extraction        Commands for extraction data management                      │
│ info              Displays information about the Argilla client and server     │
│ login             Login to an Argilla Server                                   │
│ logout            Logout from an Argilla Server                                │
│ train             Starts the ArgillaTrainer                                    │
│ users             Holds CLI commands for user management.                      │
│ whoami            Check the current user on the Argilla Server                 │
│ workspaces        Commands for workspace management                            │
╰────────────────────────────────────────────────────────────────────────────────╯

(base) jonnytr@BMGF-Y0F73D3YXV ITN-recal-data-extraction % extralit extraction export --workspace itn-recalibration --output data/output/
```

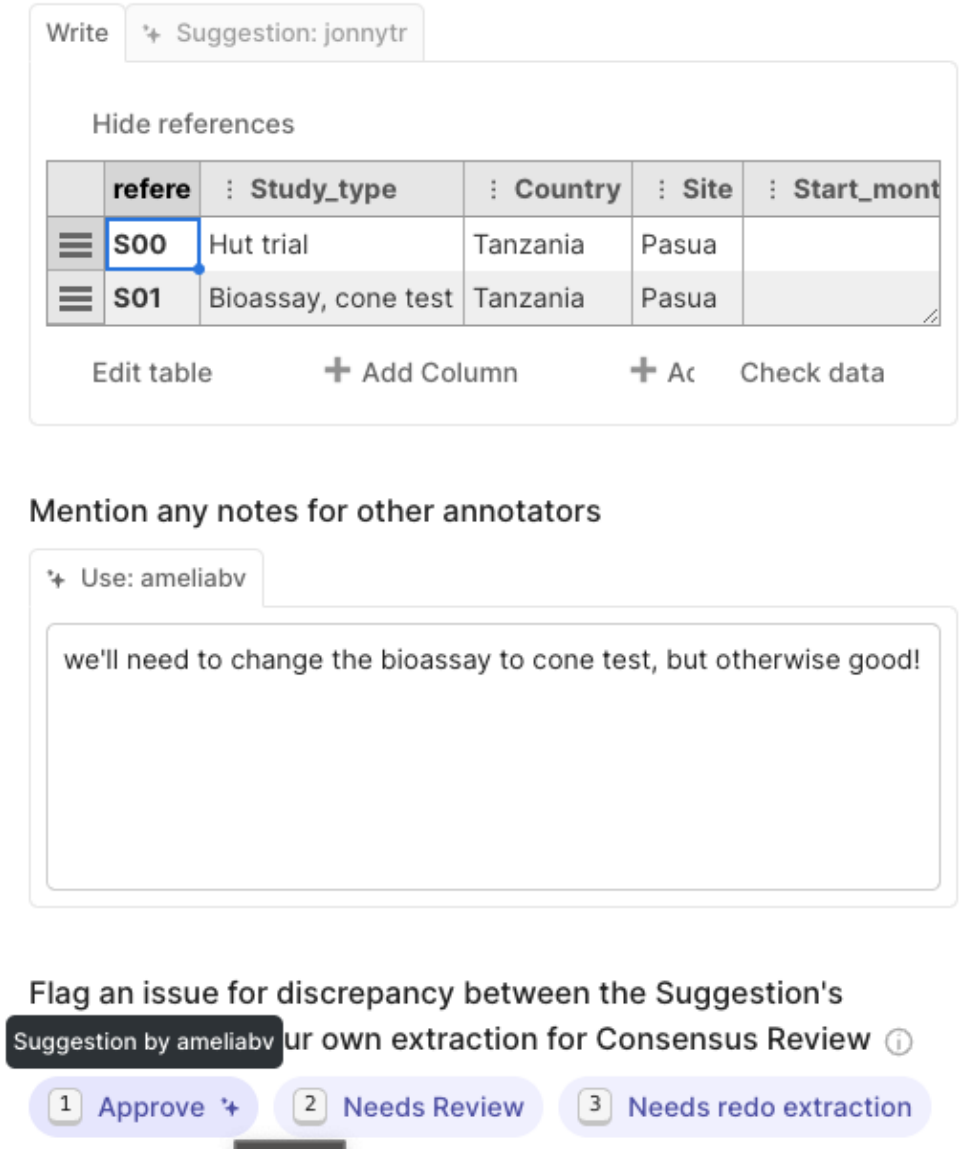
On-going feature: Reviewing & Consensus functionality

The first round of extractions from a human extractor needs review from a reviewer, which may suggest further edits.

We currently have a way to broadcast the first-round extraction, notes, and Approve/Review flags

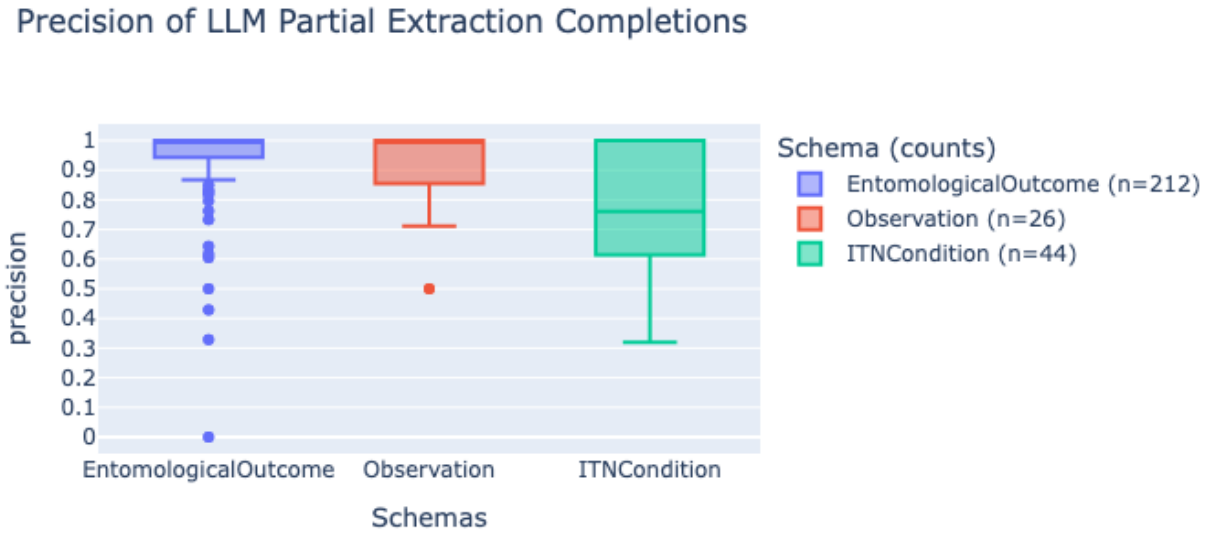
Future work needed to:

1. Formally change the record status to "Validated" to indicate there isn't further work needed and make it easier to filter out of the queue.
2. When a review is flagged and the extractor had already made the edit, need a way for the reviewer to sort which records was most recently updated.



LLM partial extraction completion feature

- **Uses**: The LLM to predict extraction values within a scope that the user defined in the extraction table. It allows the user to breakdown the extraction job into smaller, context-specified tasks and to iteratively build the table, resulting in increased overall precision and speeding up the manual extraction process by 2.5x.
- **Inputs:**
  - Selected section and table titles
  - Selected rows and columns to extract
  - Pre-filled values of the selected rows
  - Instructions (optional)
- **Model outputs**:
  - Extracted values for the selected rows and columns
- **Model type**:
  - gpt-4o (OpenAI), avg. $0.022/completion
- **Limitations**:
  - Lower accuracy when the user isn't precise about selecting the correct section and table title for each query.



Maintaining the extralit system

User roles and interactions to the system

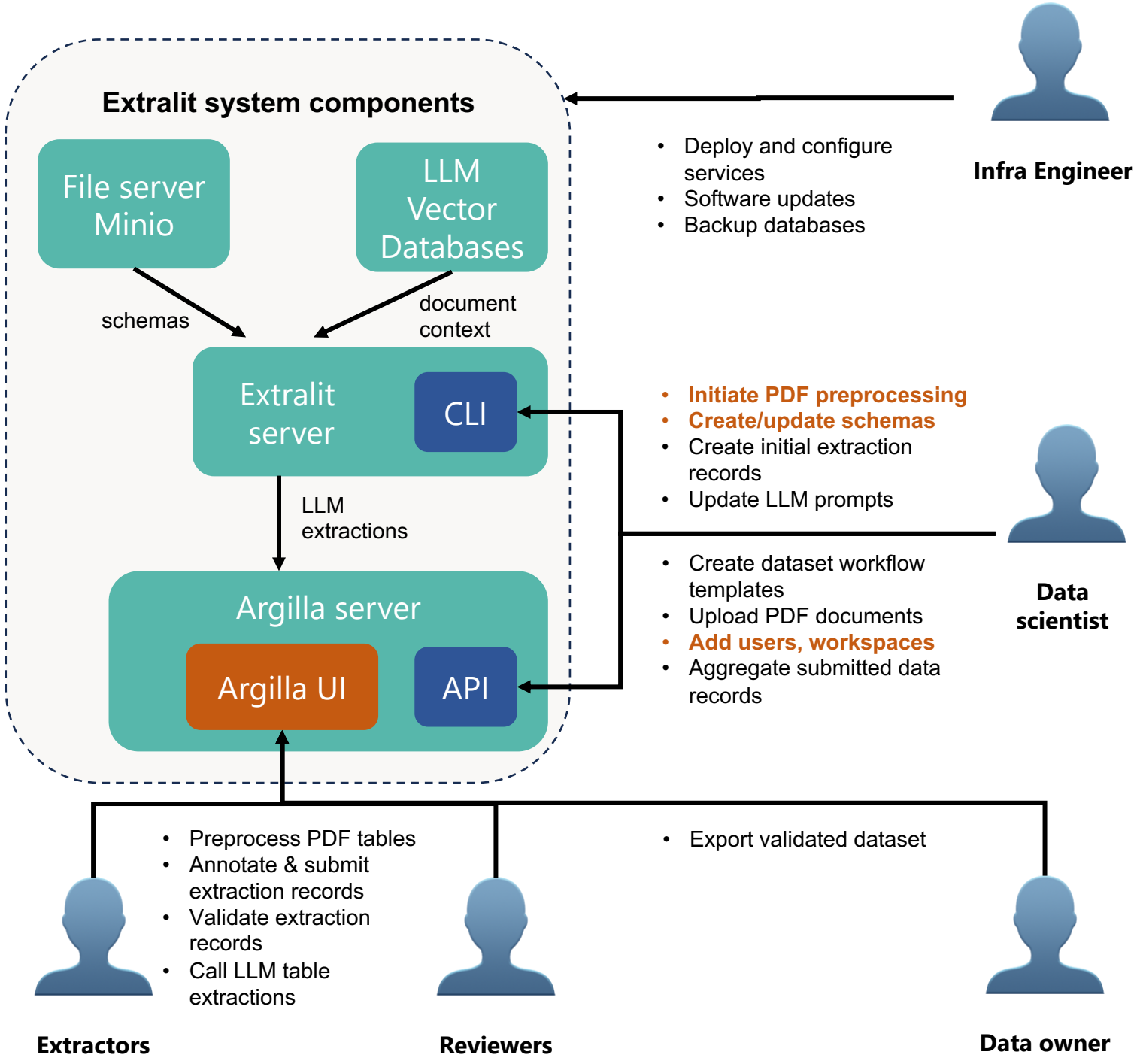
Priority work items:

1. Many essential data extraction tasks can only be accomplished by the Data scientist

2. No automation for data movements

Project repository for wiki on resources, configurations, and reproducibility

https://github.com/InstituteforDiseaseModeling/ITN-recal-data-extraction



Other system liabilities

- Two programming languages:
  - Python, Javascript/Typescript
- Two code repositories:
  - Extralit, and extralit-server
- A number of microservices
  - Webserver
  - PDF processing server
  - Postgres database
  - ElasticSearch text search
  - Weaviate vector database
  - File blob storage

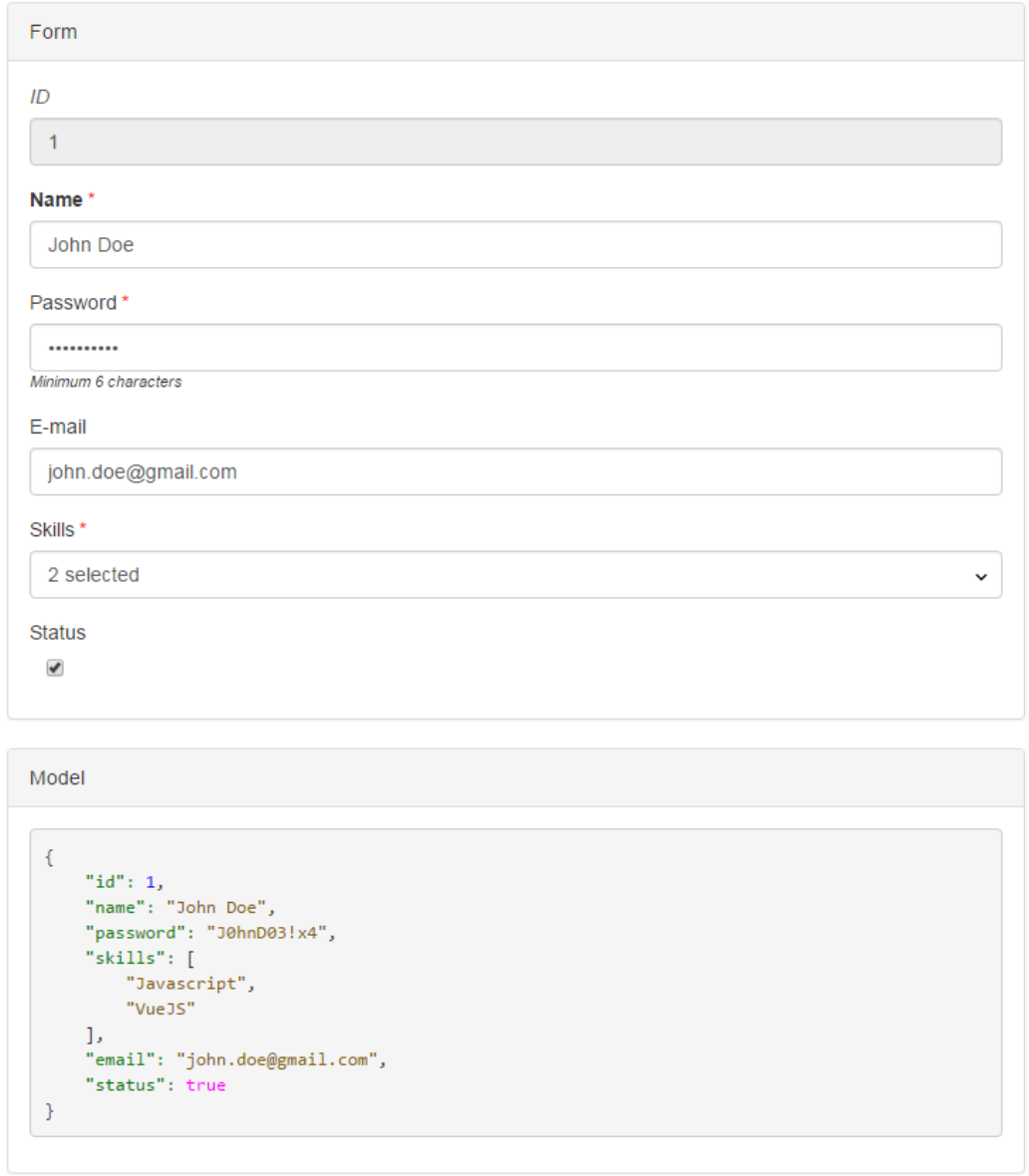
Extralit open-source

- https://github.com/extralit/extralit
- https://github.com/extralit/extralit-server
- Documentation site TDB
- Version for reproducibility: v0.2.0

- Slack: Join the extralit slack channel for bug issues, public discussions and release updates!
- Subscribe to community meetings: https://lu.ma/extralit-community-calendar

New feature in v0.2.0: Data schema editor

- The extraction schema often goes through frequent refinements, which currently is only able to edit through the code.
- We need:
    1. A non-code interface to edit schema specifications.
    2. Version tracking of the schema changes.
    3. Highlighting for when a table was last updated with an older schema version.
    4. File storage for the schema files.



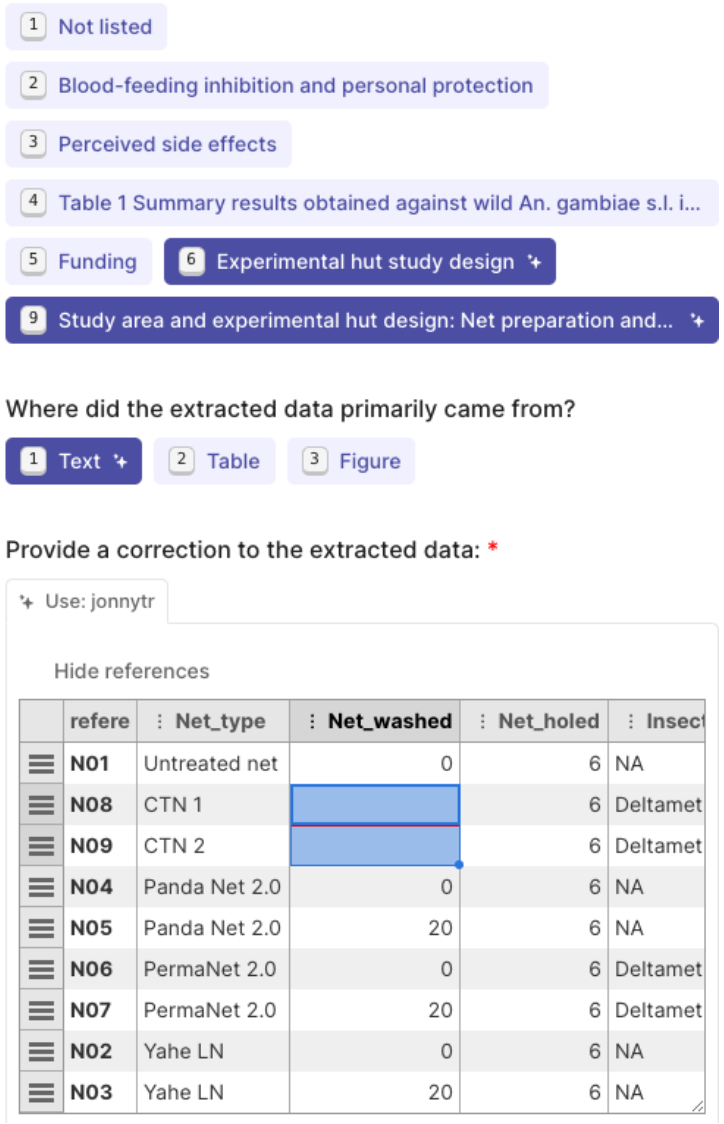
Interactive LLM autocomplete

Currently the LLM only predicts extractions at the initialization of the data records, often incorrect because:

- Retrieved context didn't include relevant tables or sections
- Missing certain Observations and ITNs that the human extractor identified

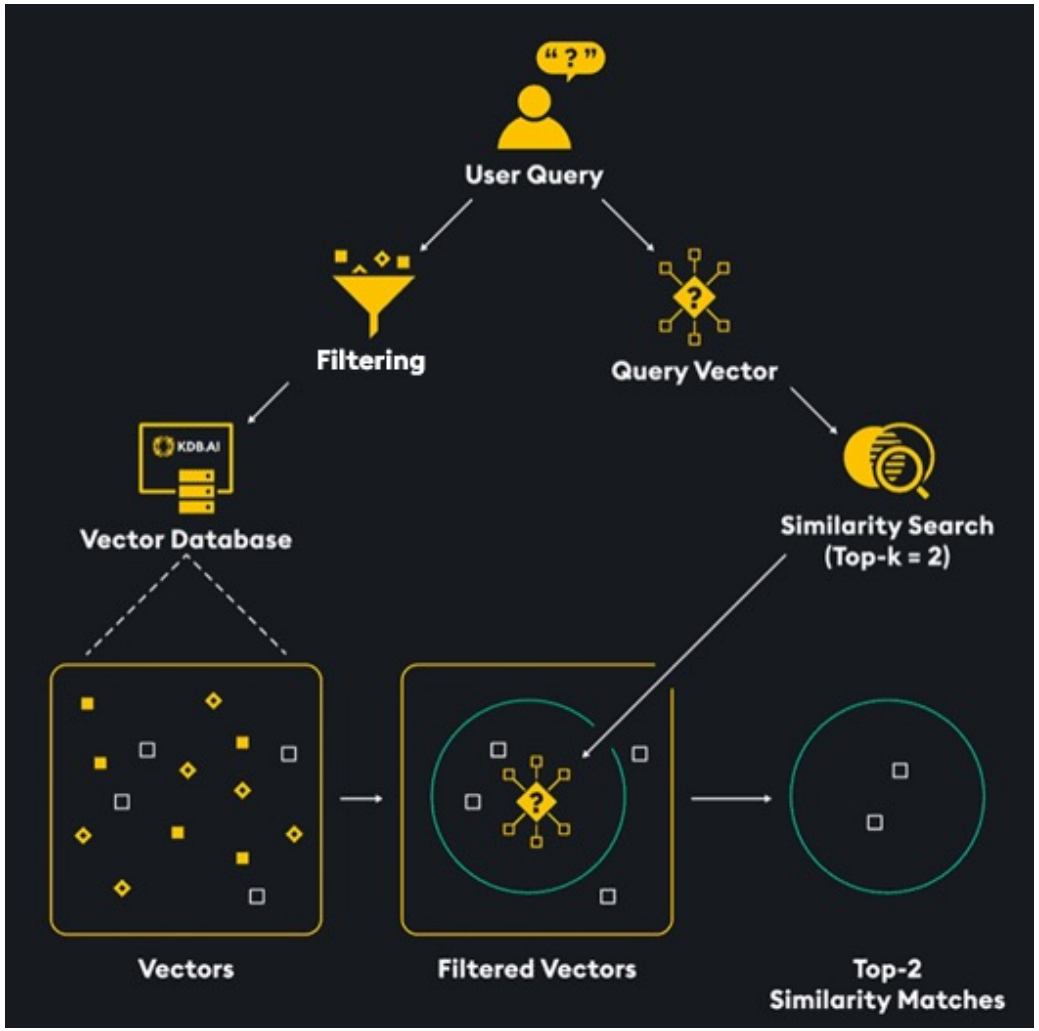
We need:

- A function to call the LLM from the UI
- The user can select the table and sections as context, and select a range of table cells for the LLM to fill



RAG with Metadata filtering

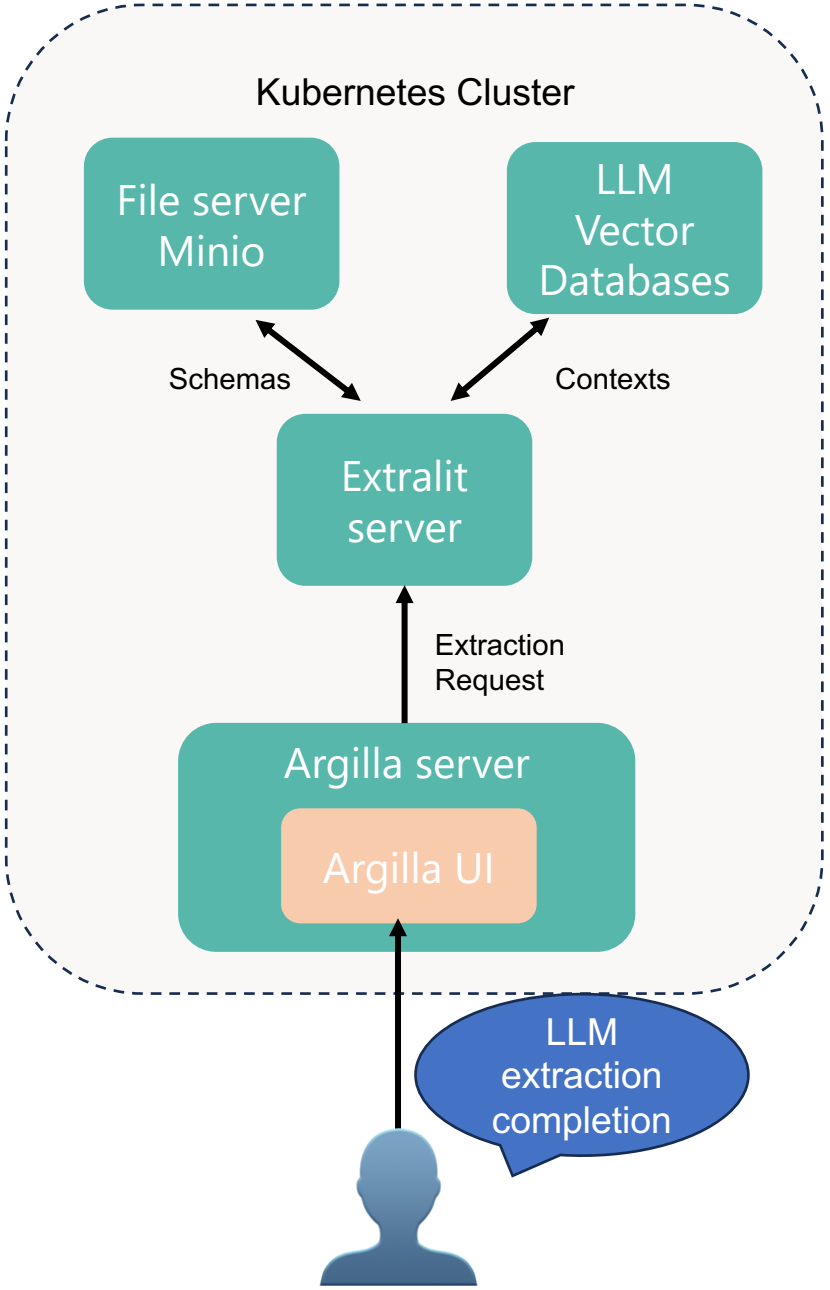
- Papers are "chunked" by section
- Each chunk have a semantic vector, with variety of metadata attached to it:
  - Header
  - Columns (for tables)
  - Relationships with other sections
- When we filter our search, we constrain the search space by reducing the number of vectors to be searched over
- This enables the user to manually select and filter relevant sections, which mitigates the retrieval issues with RAG



Extraction

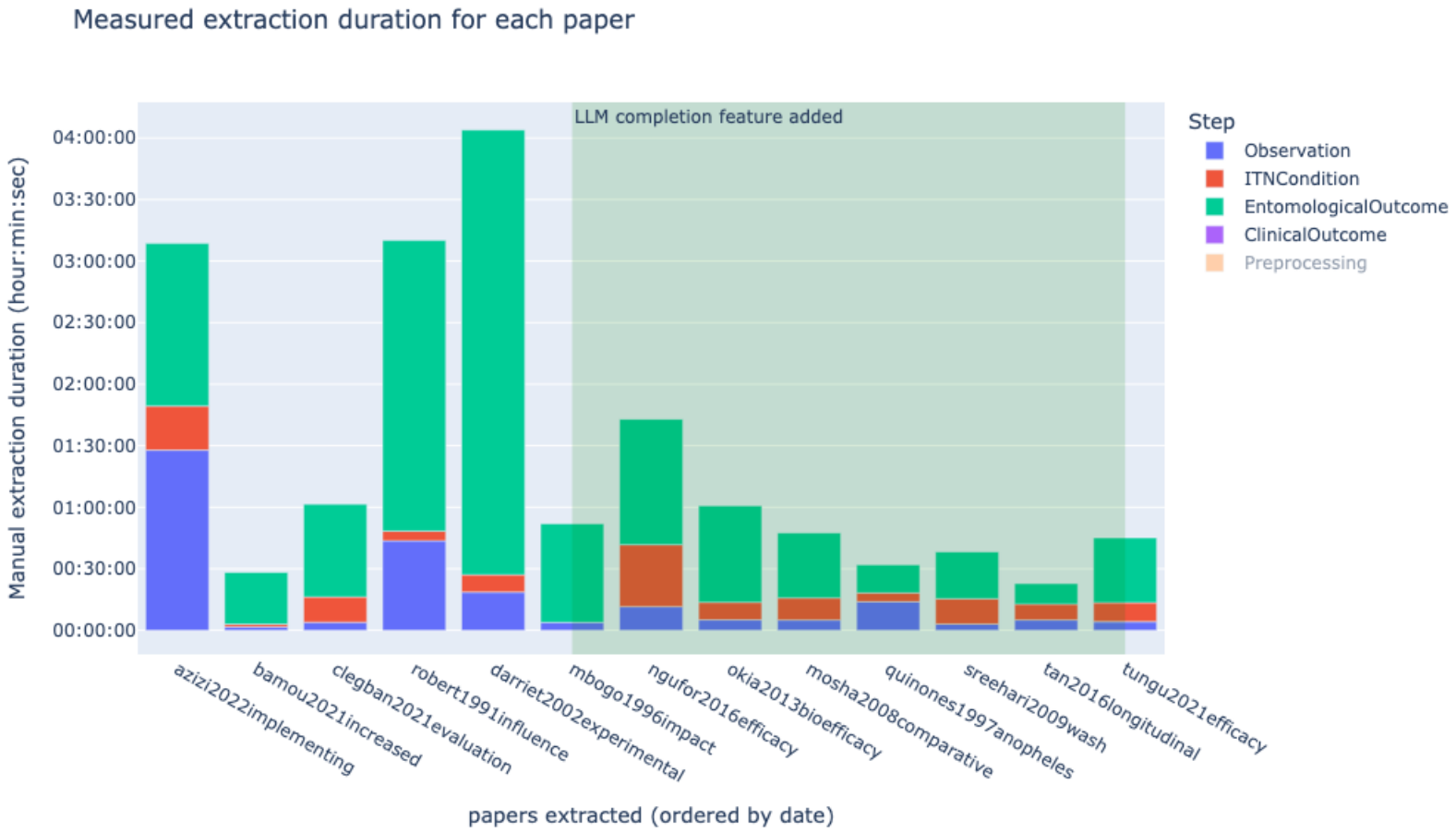
Progress on LLM extraction completion feature

- Consolidated dependencies and services to extralit/extralit.
  - Argilla server
  - Extralit server
  - TBD: PDF preprocessing servers
- Deployed the file storage server Minio
  - Integrated with schemas files and PDF files
  - File versioning capabilities: adding and reading specific versions of a file
- Progress on the "LLM Extraction Completion" feature
  - Refactored the Extraction Table component to be more modular
  - Successfully send an extraction requests from the UI with all of the required metadata
- TODO:
  - A visualized editor of the schema file



Measured time to extract each paper

The LLM completion feature was applied after the `mbogo1996impact` paper.



Comparing Solutions for PDF Preprocessing

Text extraction

Service	Latency	Cost
Nougat OCR	~30 sec /page	$537 /mo*
Mathpix API	*~1 sec /page*	*$0.025 /page*

Using Mathpix OCR API with an LLMWhisperer-style table extraction can preprocess a PDF to extraction-ready in a few seconds after upload, rather than ~13 minutes.

Table extraction

Method	% tables Detected	Table content F1	Manual cor rection
Nougat	45.8%	45.1%	12.2 ± 2.2 min/ paper
Unstructured	100.0%	59.4%	
LLMSherpa	84.7%	58.8%	
Deepdoctection	73.8%	42.7%	
Human-in-the-loop Ensemble	100.0%	73.4%	
*LLMWhisperer / Similar implementation**	*all texts & tables accurately OCR'ed with structure preserved*		*0*

Argilla open-source updates to Extralit

- Argila versions timeline
  - 1.19.0
  - 1.21.0
    - … (Bug fixes & performance)
  - 1.25.0
    - Major refactoring
  - 1.26.0
    - Feature: span labeling
  - 1.29.0

- Extralit versions timeline
  - 1.19.0
  - 1.21.0
    - PDF preprocessing workflows
    - Document extraction workflows
    - …
  - 1.26.0
  - 1.27.0
  - 1.29.0

High Level Project Taskboard (6/13/24)

1. PDF viewer component ✅
2. Table editor & input validation ✅
3. UI improvements for data labeling ✅
4. PDF highlighting annotation for extraction attribution

Argilla

1. Refine extraction schema ▶️
2. Validate gold-standard dataset
3. Extract 87 new papers

Literature extraction

1. Data validation pipeline ✅
2. Data modeling ✅
3. Build RAG + LLM extraction framework ✅
4. ~~Build Workflow Orchestrator~~
5. Added Weaviate and S3 Minio databases ✅
6. Optimize RAG strategies

Data & LLM Orchestration

1. Update evaluation metrics
2. Experiment Tracking for LLM ✅
3. Ablation of RAG parameters
4. LLM cost and prediction metrics ✅

LLMOps

Data Validation with Pandera

DataFrameModel:

Python class with built-in check functions

```python
class ITNCondition(pa.DataFrameModel):
    reference: Index[str] = pa.Field(unique=True)

    Net_type: Series[str] = pa.Field(
        nullable=True,
        description="Name of net - each type of net should have
    )
    Insecticide: Series[str] = pa.Field(
        multiselect_values = {'delimiter': ','},
        nullable=True, ignore_na=True,
        description="Enter the insecticide or insecticide combi
    )
    Net_washed: Series[int] = pa.Field(
        ge=0,
        nullable=True, coerce=True,
        description="Numerical count of number of net washes -
    )
    pHI_category: Series[str] = pa.Field(
        isin=["Good", "Damaged", "Torn", "Serviceable", "All"],
        nullable=True,
        description='One of "Good", "Damaged", "Torn", "Service
    )
```

Single-field checks:

Argilla-compatible

Highlights on single-value inputs based on type, range, format, or set of permissible values

Net_type	Insecticide	Net_washed
LifeNet	deltamethrin	
LifeNet	deltamethrin	
LifeNet	deltamethrin	
LifeNet	deltamethrin	15
LifeNet	deltamethrin	20
LifeNet	deltamethrin	-23

Numerical count of number of net washes - 0 if none and NA

Checks: required, integer: {"nullable": true}, greater_equal: 0

Multi-fields checks within the same table:

Validate the relationship between two or more fields

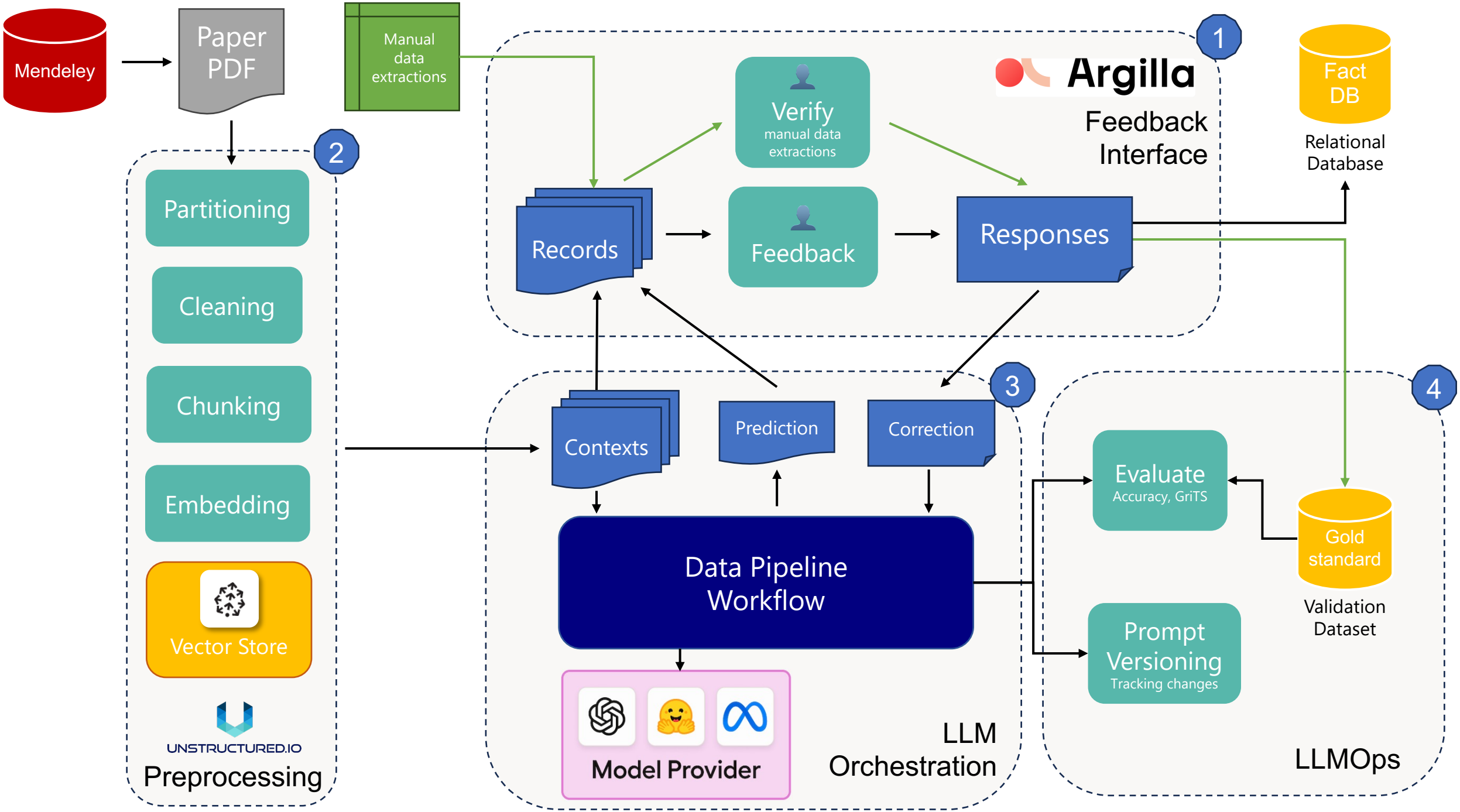
- **Consistency checks:** End_year should be ≥ Start_year
- **Conditional checks**
  - E.g, in ClinicalOutcome, if N_pos is provided, N_people should also be provided and N_pos should not exceed N_people.
- **Composite field validation**
  - KD_rate should equal (KD / Total_mosquitoes) * 100 if both KD and Total_mosquitoes are provided.

Cross-table Checks:

Validate relationships and consistency between fields across different tables

- **Foreign key checks:** Ensure foreign keys correctly reference primary keys in another tables
- **Cross-table consistency checks:** If a Study_type indicate but none ento-outcomes provided.
- **Aggregation checks:**

Python-computed only


Mendeley
Paper PDF
Manual data extractions
1
Argilla
Feedback Interface
Verify
manual data extractions
Records
Feedback
Responses
Fact DB
Relational Database
2
Partitioning
Cleaning
Chunking
Embedding
Vector Store
UNSTRUCTURED.IO
Preprocessing
3
Contexts
Prediction
Correction
Data Pipeline Workflow
Model Provider
LLM Orchestration
4
Evaluate
Accuracy, GriTS
Gold standard
Validation Dataset
Prompt Versioning
Tracking changes
LLMOps