Uploading AI-recognized lab results from scanned documents to the system

---

Glossary

1. **Lab result** - information on the results of laboratory tests performed on a specific patient on a specific date, including:
    1. Test code according to LOINC (mandatory parameter)
    2. Test result as a number or text (mandatory parameter)
    3. Unit of measurement (optional parameter)
    4. Test name (optional parameter)
    5. Reference values (optional parameter)
2. **Structured Lab Data - lab results** received from Payers in the form of structured files
3. **Unstructured (Scanned) Lab Data - lab results** received in the form of scanned patient medical records, which also include information on laboratory test results. The data is contained as images and cannot be directly used in the [redacted].
4. **Lab data available for use in the system** - all lab results which are regularly (daily) accumulated and uploaded to the [redacted] by the Ocean module. Currently, only **Structured Lab Data** is used to obtain this data. It is this data that is used in all processes, modules and interfaces of the system.
5. **Collection** - logically merged specific pages from the scanned documents received. Merging is done by patient, date of service, physician/provider, and document type. In the context of this document, collections will refer to collections with the document type "Lab:General Panel"
6. **Recognized Lab Data - lab results** in which using AI tools and post-processing it was possible to obtain **lab result** properties (LOINC code, value etc) in a structured/machine-readable form

Purposes and Rationale

Background

According to the conducted {researcj link}, the Sxope system uses only Structured Lab Data in the key processes of creating Care Gaps, Evidences, and preparing data for ML models.

At the same time, Unstructured Lab Data contains up to 10% of new and unique Lab results. (37% of lab results from Unstructured Lab Data are missing in Structured Lab Data). Unstructured Lab Data also arrives 3-6 months faster.

Purposes

Implement a new regular (minimum daily) process to recognize lab results from Unstructured Lab Data using AI and load them into the [redacted] for further use on an equal basis with Structured Lab Data

Requirements

User stories

1. As a user of the system, I want to use Recognized Lab Data as well as Structured Lab Data in all existing processes and modules of the [redacted]
2. As a user of the system I should be able to view the scanned document (ideally page and polygon) on the basis of which Recognized Lab Data was created.

Non functional requirement

The System should be able to process the current minimum amount of data per day

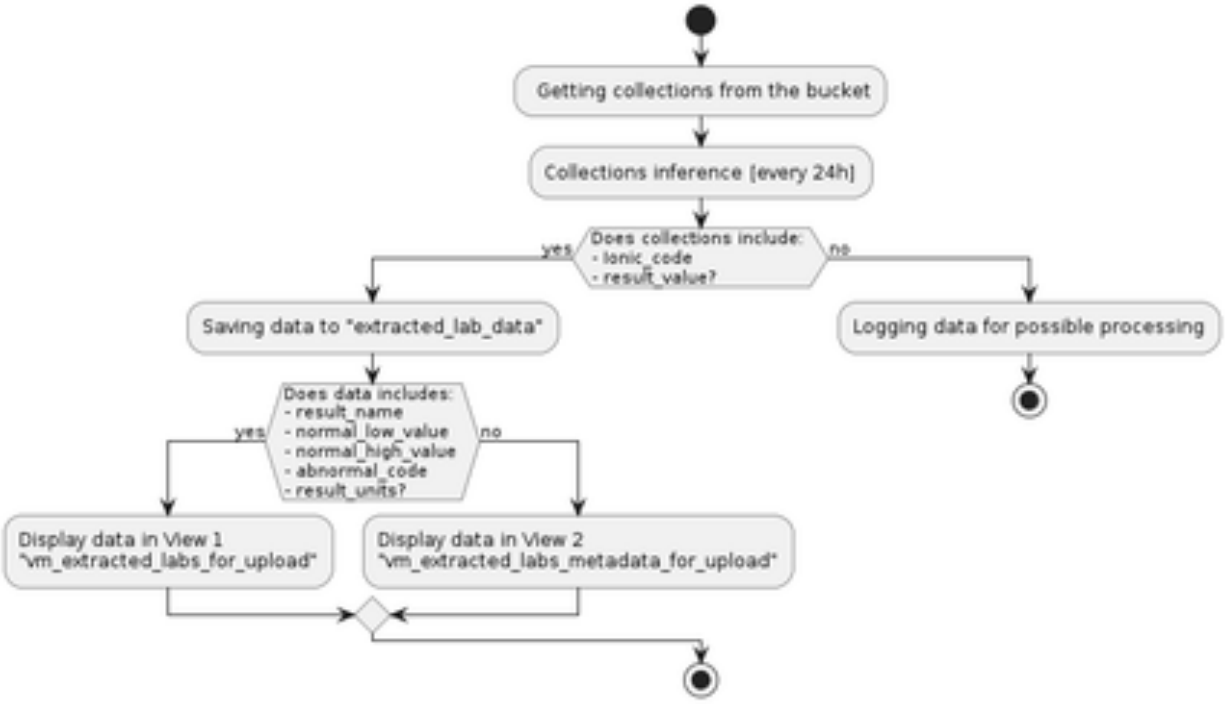
- On average 350 collections per day (this is an average of 1,500 pages, approximately 6,000 tests);
- The maximum for the last year was 1712 collections (7786 pages, approximately 30,000 tests).

Decomposition of tasks

Decomposition overview

#	Descritption	Jira

1	Providing inference for each document online, and PostProcessing once a day	
2	Collecting new collections with lab results(pdf), processing them with AI models, and storing model outputs in BQ	
3	Creating a BQ summary table for uploading **Recognized Lab Data** to the [redacted] (Validation, enrichment, ML data cleansing	
4	Loading **Recognized Lab Data** into the Spanner database with final lab data by Ocean module.	
5	Creating an API interface for viewing detailed information (collection id, page, polygons) of **Recognized Lab Data**	
6	Creating an interface for viewing **Recognized Lab Data** details in the [redacted] UI	



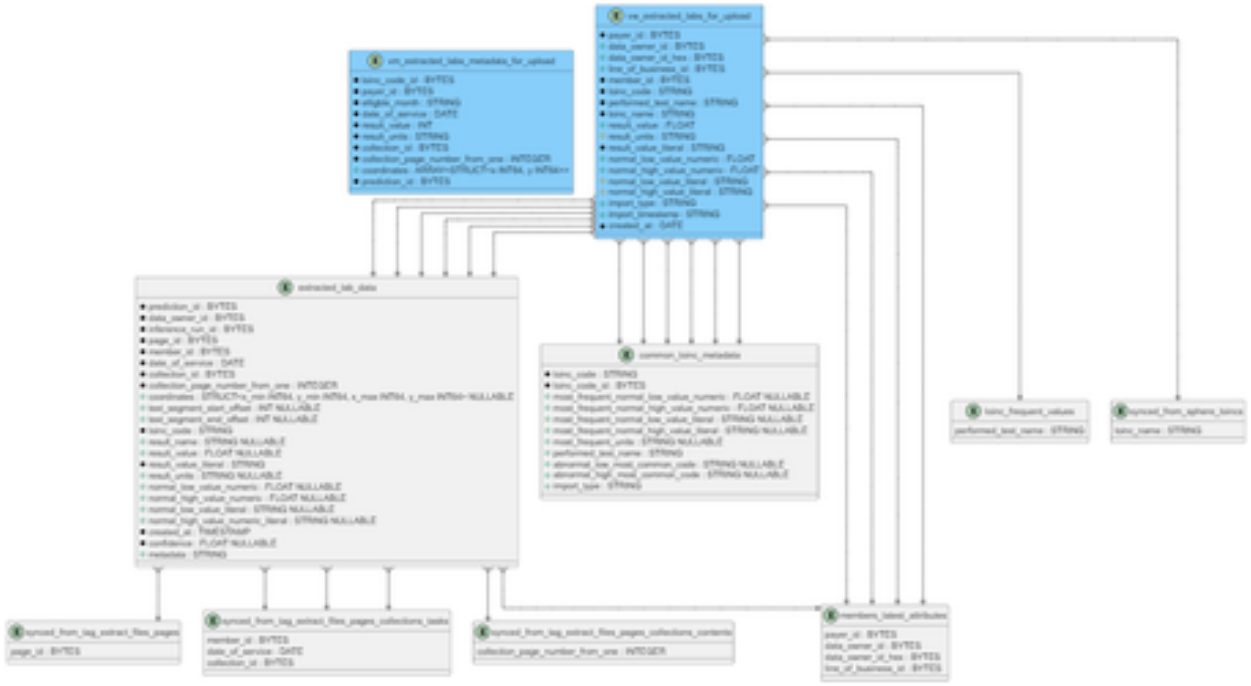
Entity Relationship

View 1. `vm_extracted_labs_for_upload` - Required to upload lab data to Ocean;

View 2. `vm_extracted_labs_metadata_for_upload` - Used to store metadata about where a value is found in documents;

Table 1. `extracted_lab_data` - Data obtained as a result of labs inference;

Table 2. `common_loinc_metadata` - Contains missing data that ML was unable to collect as a result of inference. Obtained by collecting data in BQ based on statistics.



*click to zoom**

Scenarios to test

№	Scenario Name	Scenario's description	Test schema	Comments
1	Not exist - Added by NER	• Data doesn't exist on the snapshot • NER team added some data (import_type = "Document Recognizer)	1. Verify that all the field names match the vw_extracted_labs_for_upload 2. If don't match, compare to extracted_lab_data	
2	Was Added by NER - Updated by NER (better quality)	• Data was added by NER team • The same data was recognized one more time and the updated version was sent to view	1. Check if on the snapshot there is still one row for this member_id + dos + loinc_code 2. Check that the last version of the data is presented	

3	Was Added by NER - Updated by NER (worse quality)	• Data was added by NER team • The same data has been recognized one more time and the updated version was sent to view	1. Check if on the snapshot there is still one row for this member_id + dos + loinc_code 2. Check that the **old** version of the data is presented	
4	Was added by 'labs' - Updated by NER	• The data has been uploaded with import_type "labs" • NER team recognized data for the same member_id + dos + loinc_code	1. Check that only 'labs' data is kept and it wasn't disturbed	