Deliverable D5.1

UML specification describing the development of the new index and its integration within SIBiLS

Project Title	FAIR-ification of Supplementary Data to Support Clinical Research
Project Acronym	FAIRClinical
WP No & Title	WP5: Supplementary data search engine
Delivery Date	30/11/2024
Authors	Anaïs Mottaz (CH) Patrick Ruch (CH) Julien Gobeill (CH)
Reviewers	
Funding Acknowledgement	This work is supported by the CHIST-ERA grant CHIST-ERA-22-ORD-02, by the Luxembourg National Research Fund (FNR, INTER/CHIST23/17882238/FAIRClinical), by Swiss National Science Foundation (SNSF, 20CH21_217525), by the Agence Nationale de la Recherche (ANR-23-CHRO-0008-01), and by Engineering and Physical Sciences Research Council (EP/Y036395/1).

1 Introduction

The SIB Literature Services (SIBiLS) is a custom search engine API [1] that provides search services for exploring biomedical literature. In addition to mirroring MEDLINE and PMC contents with daily updates from the National Library of Medicine, SIBiLS annotates these contents with a wide range of terminological resources. The resulting collection contains more than 2 billion RDF annotations, making SIBiLS the world's largest semantically enriched library for life sciences.

The goal of WP5 is to implement the onboarding of external annotations to improve the searchability and semantic interoperability of SIBiLS and in particular the case reports and case report forms documents. The information extraction components developed in WP3 and 4 will be applied to the existing content of SIBiLS as well as to newly acquired supplementary data files from WP2. SIBiLS will be updated to incorporate the new annotations from WP3 and 4, and specific indexes will be created in the MongoDB and Elasticsearch/Lucene backend systems that serve SIBiLS.

This document contains a general schema and more detailed UML specifications describing the workflow and interactions for integrating and updating external annotations into the SIBiLS system.

2 System Description

This section provides an overview of the SIBiLS workflows and how these workflows will be impacted to harvest annotations provided by third party groups.

2.1 Overview of the workflow

The schema in Figure 1 outlines the general SIBiLS workflow and the integration of external annotations from WP3 and WP4 into this workflow. First, it collects and parses documents from heterogeneous sources (e.g. Medline, PMC+) into a uniform content, followed by internal annotations using over 20 controlled vocabularies (e.g. MeSH, GO). External annotations from WP3 and WP4 are uploaded in BioC [2] format to the SIBiLS public FTP server. These annotations are included into SIBiLS via the daily updates. Finally, SIBiLS makes the enriched annotations available through its API and frontend search engine to provide end users with seamless access to searchable, up-to-date annotated documents.

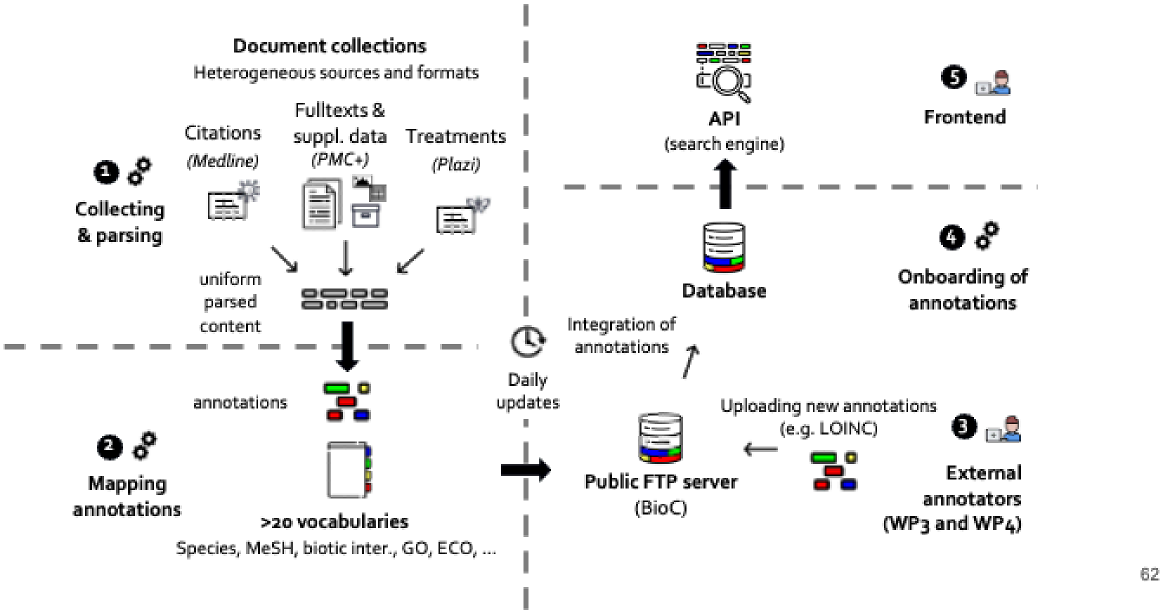

Figure 1: Overview of the workflow for onboarding external annotations into SIBiLS

2.2 High-Level Architecture

Figure 2 shows the diagram for creating annotations and updating the SIBiLS system. The documents are retrieved in BioC format and the sentences are annotated using sentence splitting, Named Entity Recognition (NER) and terminology mapping. The annotations are then converted back into BioC format and uploaded to the SIBiLS FTP server. In addition, relation extraction (RE) is applied to the annotations to detect relationships. These are also uploaded to the SIBiLS FTP in BioC format. The resulting annotations and relationships are integrated into the source documents and updates are sent to MongoDB and ElasticSearch so that the enriched data is stored and searchable in SIBiLS. The main elements are the SIBiLS FTP server, the annotation workflow, the annotations integration workflow and the SIBiLS backend components, which are described in the following subsections.

2.2.1 SIBiLS FTP server

The *SIBiLS FTP* server acts as the central repository for the management of external and internal annotation data. It contains the external BioC inputs from WP2 and the pre-existing BioC annotation data (*BioC Ext/SIBiLS*) as well as the new annotations (*BioC new annotations*).

The FTP also contains the internal SIBiLS source documents, which are divided into: *bib*, the bibliographic data for each publication, *sen*: the splitted sentences, and *ana*, the annotation data.

2.2.2 External annotations workflow

The *External annotations* are divided into two workflows: the *Entities* workflow, which corresponds to WP3, where the BioC files from WP2 are downloaded in a compressed .json.gz format. Processing involves splitting the document content into individual sentences (*Sentence splitting*) and preparing the text for entity recognition. *NER* (Named Entity Recognition) identifies entities within sentences and assigns IOB (Inside-Outside-Beginning) tags, followed by terminology mapping to normalize the recognized entity with an existing terminology (*Normalization*). The result is then converted into an annotation (*Sentence annotation*) that is added to the BioC file before it is uploaded back to the *SIBiLS FTP*.
In the *Relations* workflow, which corresponds to WP4, BioC files with annotated entities are downloaded to extract relationships between these entities (*RE*), which are then integrated into BioC and uploaded back to the *SIBiLS FTP* server.

2.2.3 External annotation integration

The integration process ensures that new annotations are fetched, processed and integrated into the *SIBiLS* system. The first steps include fetching the updated BioC annotation files and retrieving the corresponding source documents (i.e. *sen* and *ana*). The following steps consist of matching the annotations with the corresponding sentences in the source documents and updating the SIBiLS components, in particular *ana*, *MongoDB* and *ElasticSearch* for storing and making annotations available to the SIBiLS API and frontend.

2.2.4 SIBiLS backend components

The *SIBiLS backend* includes *MongoDB*: the structured database for storing the processed data and *ElasticSearch*: the search engine for querying and retrieving annotation data, which is used by the SIBiLS API and search engines.

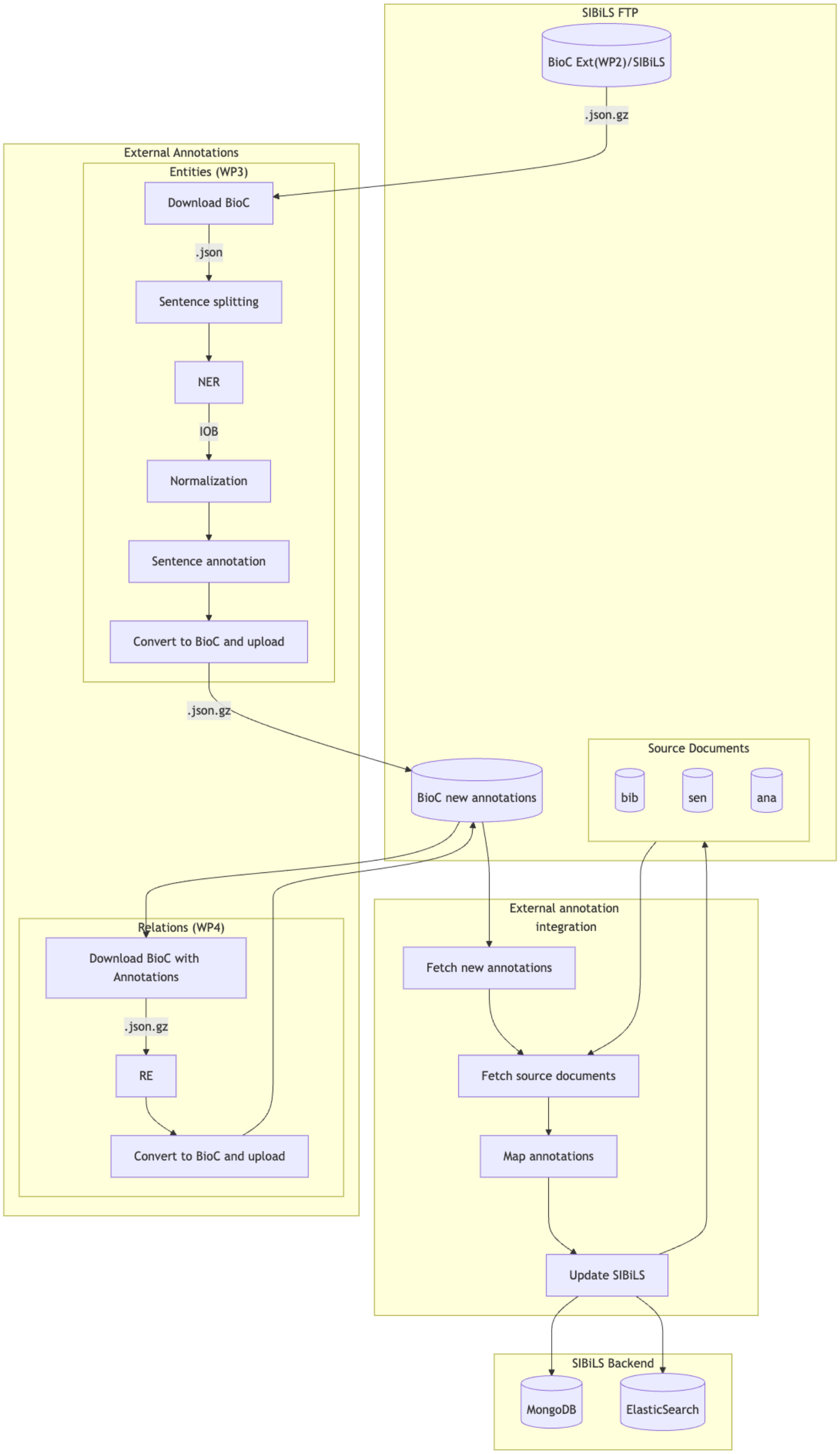

Figure 2: Activity diagram describing the workflow for creating and integrating the new annotations generated by WP3 and WP4 in SIBiLS

2.3 Detailed processing of new annotations

The UML activity diagram on Figure 3 represents a more detailed description of external annotation onboarding and integrating into SIBiLS. This process ensures that annotations are properly aligned to SIBiLS storage and search documents, making new annotations generated by WP3 and WP4 accessible through SIBiLS.

The workflow begins with the detection of new or modified annotations through daily logs and progresses to validate the existence of required documents, map annotations accurately, and integrate the updates into the system. Along the way, any missing documents are processed to be integrated as a new document into SIBiLS. Missing terminologies from external annotations are automatically detected to be manually checked and integrated into SIBiLS.

2.3.1 Fetch New Annotations

The *fetch new annotations* component identifies new or updated BioC files by analyzing *daily logs* to *check new/modified BioC*. Once changes are detected, the system creates a list of documents to update.

2.3.2 Fetch Source Documents

This component checks the availability of required *source documents* using the generated list of updates. It searches for the corresponding sentence-level documents (*sen*). If a document is missing, the system triggers the *create new source documents* process to generate the required documents, which are then added to the *source documents* repository. If the document exists, the workflow proceeds directly to *the map annotations* phase.

2.3.3 Map Annotations

The *map annotations* phase aligns new annotations with the corresponding text in source documents. It begins by mapping paragraphs or sentences to the *sen* file. Next, annotations are aligned to the text during the *align annotations on sen* step. Once aligned, the system verifies the presence of the terminology in the system. If the terminology exists, *create annotations* generates the enriched annotations. If terms are missing, the system logs the issue for further resolution.

2.3.4 Update SIBiLS

The final phase integrates the newly created annotations into the SIBiLS system. The *ana* files are updated and the annotations are indexed into *ElasticSearch* and *MongoDB*. This ensures that the new annotations generated by WP3 and WP4 are findable and accessible through SIBiLS.

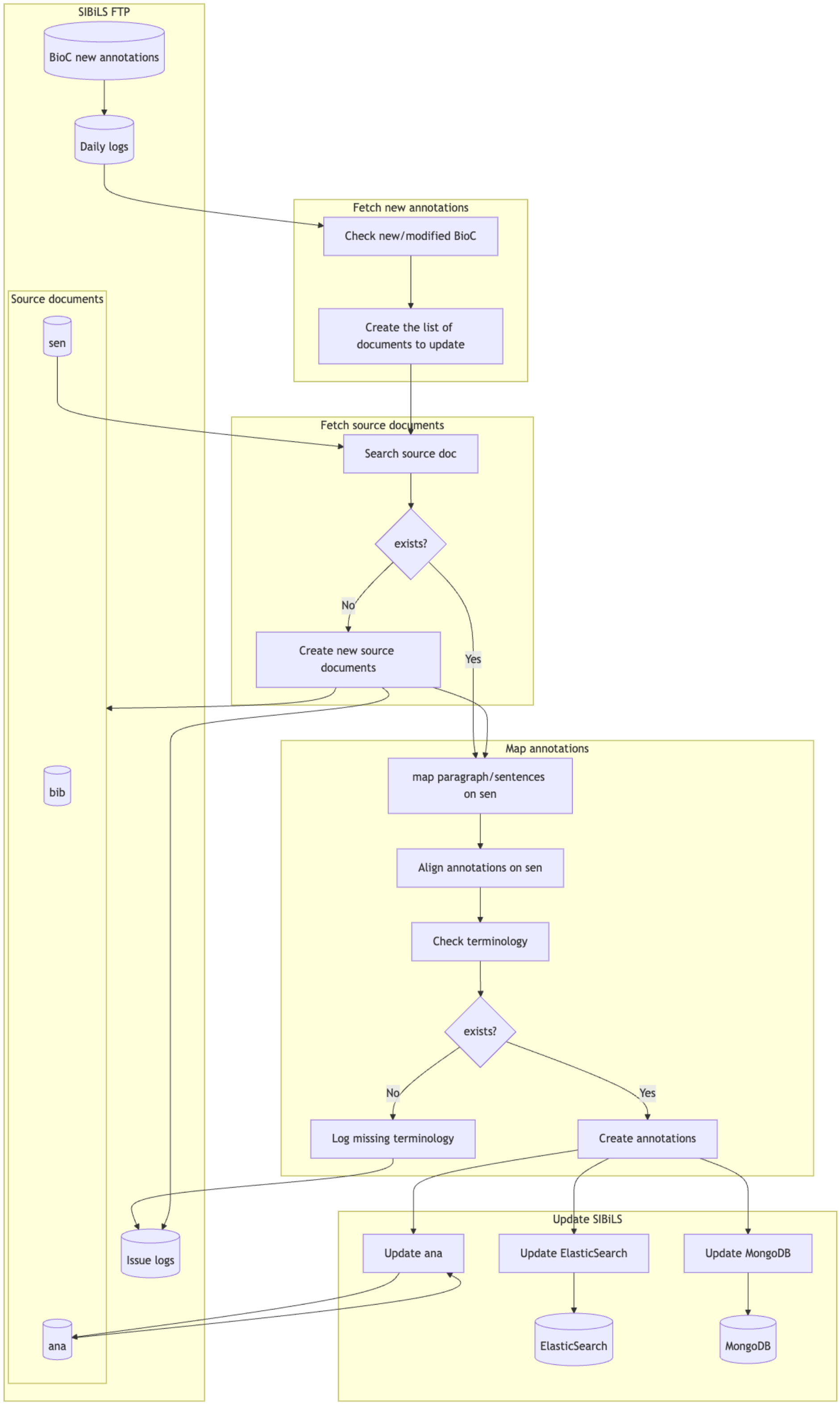

Figure 3: Activity diagram describing the steps for integrating the new annotations into SIBiLS

3 Next steps

The next steps involve the implementation and testing of the described schema to ensure its functionality and reliability. During the implementation phase, each component of the workflow, from fetching new annotations to updating the SIBiLS system, will be developed and integrated. This will be followed by testing individual components as well as validating the flow between modules. Special focus will be placed on handling edge cases, such as missing documents or terminology, to ensure the system is able to correctly log any issues.

4 References

[1] Gobeill J, Caucheteur D, Michel PA, Mottin L, Pasche E, Ruch P. SIB Literature Services: RESTful customizable search engines in biomedical literature, enriched with automatically mapped biomedical concepts. Nucleic Acids Res. 2020 Jul 2;48(W1):W12-W16. doi: 10.1093/nar/gkaa328. PMID: 32379317; PMCID: PMC7319474.

[2] Comeau DC, Wei CH, Islamaj Doğan R, Lu Z. PMC text mining subset in BioC: about three million full-text articles and growing. Bioinformatics. 2019 Sep 15;35(18):3533-3535. doi: 10.1093/bioinformatics/btz070. PMID: 30715220; PMCID: PMC6748740.