**Data Deposit Report**

**Phase1: Extraction of funding information**

**Part A : Extracting the funding information from PubMed**

Phase 1 codes can be run in two modes –

**I . Pilot**

1. The full text of research papers has to be manually downloaded in xml format from PubMed (<https://www.ncbi.nlm.nih.gov/pmc>)
2. Extraction query (e.g., for the date 2021/01/01)

**Graphical user interface, text, application, email

Description automatically generated**



1. XML file download

Added the manually downloaded xml files in the data folder

Graphical user interface, text, application

Description automatically generated**Table

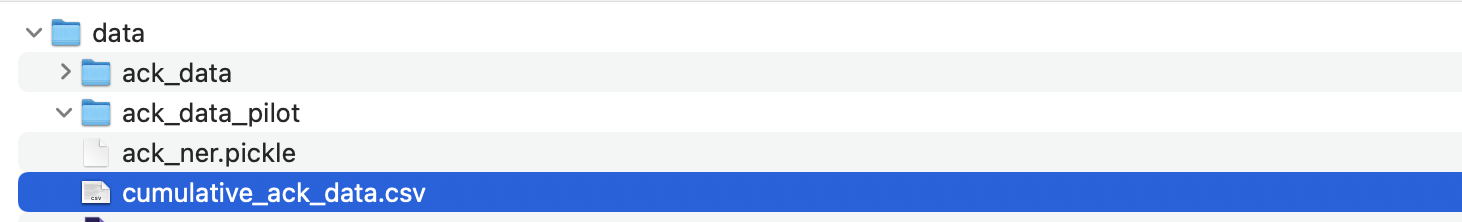
Description automatically generated**

**II . main**

The code will automatically retrieve xml files containing the full text of research papers based on the selected time period

**Part B: Combing the funding information for all the research into a single “.csv” file**

A “.csv” file with funding data for all the research papers will be stored together.

****

A screenshot of a computer

Description automatically generated with medium confidence

**Phase2: Application in the selected NER tools on the funding information (Approximate Execution Time – 1.5 hours)**

Selected NER tools will be applied on the funding data stored in “cumulative\_ack\_data.csv”

This will generate two files highlighted below-

Background pattern

Description automatically generated



The ack\_ner.pickle will be used for result matching in phase 3

**Phase3**

**Part A: Result Matching (Approximate Execution Time – 30 minutes)**

This phase will compare the organization names identified by the NER tools with a reference list of organizations extracted from the Crossref Funder Registry

There are two sources from which we will use organizations to perform result matching

1. Organization names from Crossref

Table

Description automatically generated



1. NER tool output

Background pattern

Description automatically generated



The output will be a “.csv” containing the result matching counts. The observations will be reported from this file



Graphical user interface, application, table, Excel

Description automatically generated

**Part B: Applying NER on Crossref (Approximate Execution Time – 30 minutes)**

The data pertaining to the list of organizations will be extracted from the file “registry.rdf”. The code will transfer all the names of organizations into a flat list and apply the 4 selected NER tools on this list.

The result will be a CSV file named “crossref\_ner.csv” displaying the list of organizations which match between Crossref and NER tool output.

Table

Description automatically generated with medium confidence