Deliverable D2.1

Workflow for downloading and standardising clinical literature and supplementary materials

Project Title	FAIR-ification of Supplementary Data to Support Clinical Research
Project Acronym	FAIRClinical
WP No & Title	WP2: Data gathering and standardisation
Delivery Date	31/08/2024
Authors	Tim Beck (UK) Thomas Rowlands (UK)
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

While researchers primarily access the biomedical literature through web browsers which display HTML or PDF formats of publications, these formats are not optimised for natural language processing (NLP) and text mining. BioC is a simple data structure, in either XML or JSON format, to store and exchange text documents and annotations between different language processing systems and text mining tools[1]. Digital archives of biomedical literature such as PubMed Central (PMC) and SIB Literature Services (SIBiLS) provide BioC versions of the full-text of publications to support text analytical activities. PMC makes articles available in BioC format from the PMC Open Access Subset and the PMC Author Manuscript Collection[2]. The PMC BioC key file[3] defines the semantics associated with the BioC data. The BioC format will be used in FAIRClinical for the storage and exchange of a full-text corpus of published clinical case reports.

Supplementary data files are provided by journal publishers in a range of formats, including PDFs, word processor documents, spreadsheets, movies, images and plain text files. In order to optimise these formats for NLP and text mining, they will be transformed to computer interpretable formats. The Auto-CORPus tool for converting biomedical literature to the BioC format[4] has been extended to transform PDFs, word processor documents, spreadsheets, presentation and text files to BioC. SIBiLS has extensively processed supplementary image files using optical character recognition (OCR) methods to extract text. These resources will be used to standardise the supplementary files associated with the clinical case report corpus to develop a clinical supplementary materials corpus. Movies will be excluded from the corpus, however, as the technology improves to automatically transcribe audio from movie files, this can be reconsidered in future versions of the corpus.

Here we describe a workflow built to download and standardise clinical literature and supplementary materials. The workflow will be used to create a clinical case report full-text literature corpus and a clinical supplementary materials corpus (deliverables D2.2 and D2.3).

---

[1] Comeau DC, et al. (2013). BioC: a minimalist approach to interoperability for biomedical text processing. Database. 2013:bat064. doi:10.1093/database/bat064

[2] Comeau, DC, et al. (2019). PMC text mining subset in BioC: about three million full-text articles and growing. Bioinformatics. 35(18):3533–3535. doi:10.1093/bioinformatics/btz070

[3] PMC BioC key file: https://ftp.ncbi.nlm.nih.gov/pub/wilbur/BioC-PMC/pmc_ascii.key

[4] Beck T, et al. (2022). Auto-CORPus: A Natural Language Processing Tool for Standardizing and Reusing Biomedical Literature. Front Digit Health. 4:788124. doi:10.3389/fdgth.2022.788124

The workflow is written using the Python programming language and is available from https://github.com/FAIRClinical/ClinicalCorporaWorkflow.

2 Description of work accomplished

2.1 Overview of the workflow

The aim of the workflow is to produce:

1. A clinical case report corpus, consisting of full-text publications in BioC format.
2. A supplementary materials corpus, consisting of supplementary files associated with the clinical case report corpus in formats that are accessible to NLP and text mining algorithms.

The workflow will involve four key stages:

1. PMC BioC full-text retrieval and processing
2. Application of a search query to filter to the publications for clinical case reports
3. Gather and process associated supplementary files
4. Standardise the format of supplementary files to support text analytics

To enable provenance tracking across the key stages of the workflow, logs will be produced to record which files have been included and excluded from the corpora and if any errors were detected during the processing of files. Figure 1 presents an overview of the workflow. Each of the four stages are described below.

2.2 Stage 1: PMC BioC full-text retrieval and processing

PubMed Central Open Access full-text articles are available in BioC JSON or BioC XML, in Unicode or ASCII, from a FTP server hosted by the National Centre for Biotechnology Information (NCBI). The articles are available in compressed archive files, grouped by PMC identifier. Each BioC JSON ASCII archive is downloaded programmatically from the FTP server (https://ftp.ncbi.nlm.nih.gov/pub/wilbur/BioC-PMC/), before being processed. Each archive is processed individually by the workflow.

In an initial housekeeping step, the file extension for each file is changed from .xml to .json to accurately represent the file contents. Once renamed, the files are optimised for use with the BioC Python module 'bioc' (https://github.com/bionlplab/bioc) by expanding their parent JSON array structure so that only BioC JSON content is present.

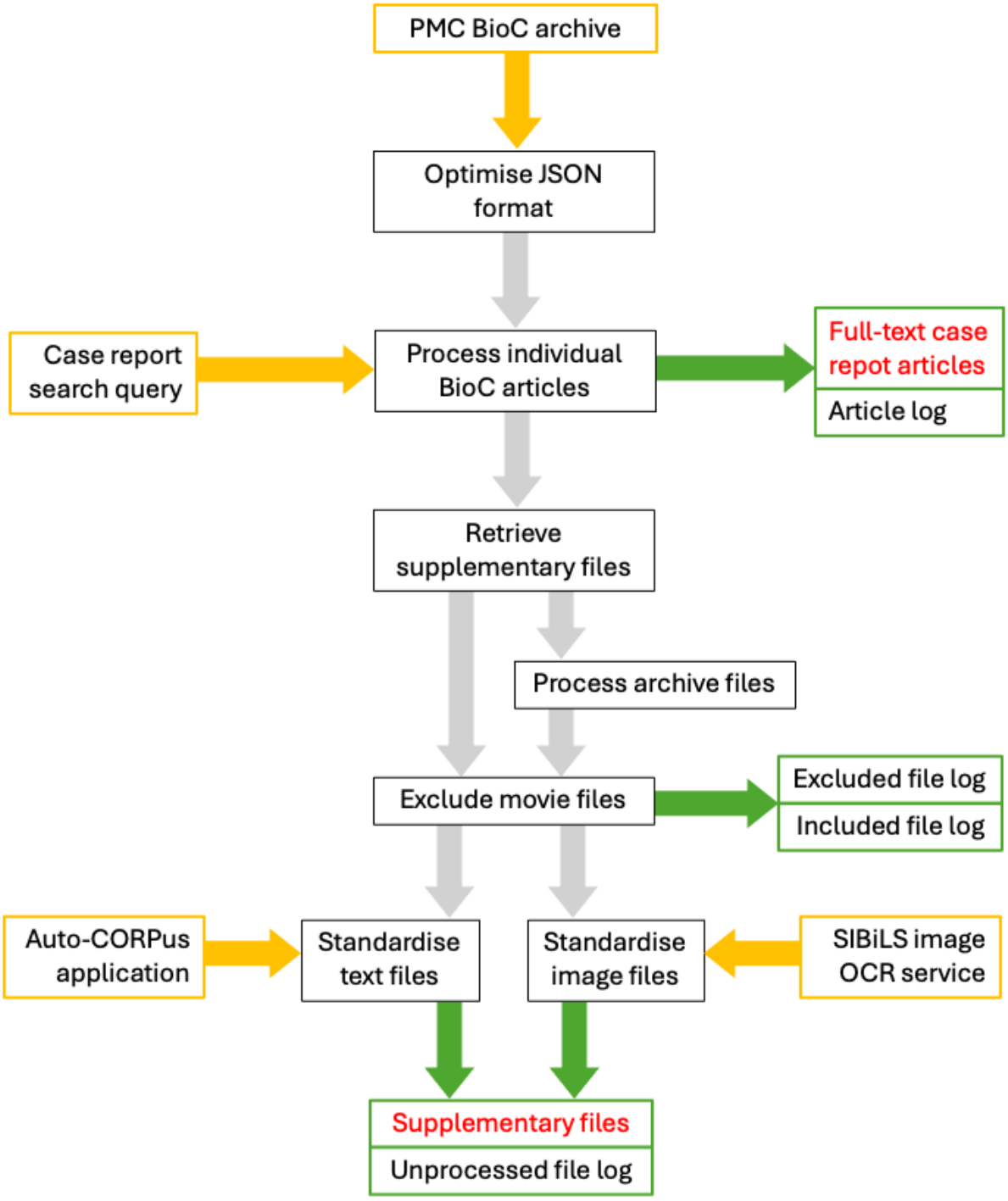

Figure 1: Overview of the workflow for downloading and standardising clinical literature and supplementary materials.

2.3 Stage 2: Application of a search query

The search query filters articles based on the presence of the string "case report" within the article title. However, PMC splits some article titles into a title and a subtitle. Figure 2 shows the different title formats between PubMed and PMC for an example article. In the BioC file, the title section text is queried for the presence of the "case report" string. If the string is not found and the subtitle parameter is present, the subtitle is also searched for the "case report" string. Figure 2 shows an example of how the title and subtitle text is represented in the BioC format. Some of the filtered PMC articles contain just a title or a title and abstract. In order to create a full-text corpus, these articles are excluded by scanning the final section type of the BioC file and removing those files where the final section type is "abstract" or "title".

Articles are output in a compressed archive file that uses the same naming convention as the input archive. Additionally, for each output archive, an article log is output in TSV format. The article log indexes all articles included in the archive and consists of the following fields:

1. PMC identifier
2. Article title
3. Article subtitle (if the reason for inclusion)

2.4 Stage 3: Supplementary file retrieval and processing

Each article is scanned for BioC section types containing a value of "SUPPL", indicating supplementary content is associated with the article. Since the links to external resources are not included in BioC files, they are retrieved from the HTML version of an article using a regular expression search. The supplementary files linked in an article are downloaded programmatically.

For each article containing supplementary files, a new directory is created with the name <*PMCID*>*_supplementary*, where <*PMCID*> is the article's PMC identifier. Within these directories, two subdirectories named *raw* and *processed* are created. The *raw* directory contains each original supplementary file downloaded for the article and the *processed* directory contains the output from the stage 4 file standardisation process.

Due to the increased storage requirements for retaining movie files compared to other file types, movies are removed. The URLs for removed movie files are captured in an *excluded log*

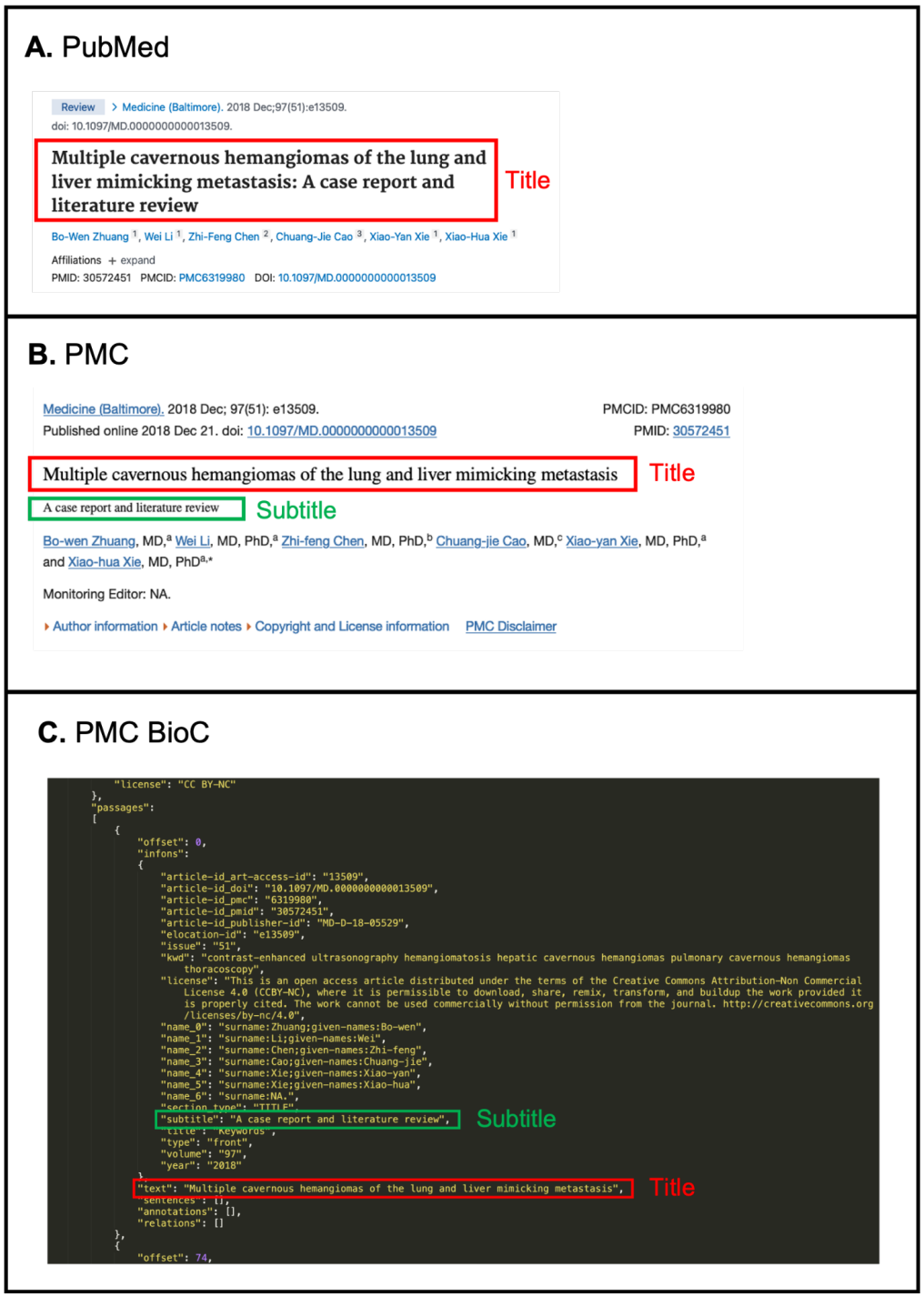

Figure 2. The use of title and subtitle in PubMed and PMC for an example article. A) PubMed displays the title as a single string. B) PMC displays a title and subtitle. C) PMC BioC title text and subtitle parameters searched during stage 2 of the workflow.

file so they can be retrieved if required. Each file extension is analysed and those with movie extensions (mp4, mov, avi, wmv, webm, flv, mpg, movi, m4v, 3gp) are expunged. Since movies can also be contained within archives and compressed archives, these are analysed and movies expunged. If the removal of movie files results in an empty archive (i.e. no other file types are contained in the archive) then the archive itself is expunged. Similarly, if the removal of movie files (and/or archives if necessary) results in the removal of all supplementary files associated with an article (i.e. only movie files are associated with an article), then the *<PMCID>_supplementary* directory is expunged.

Log files are created to list the supplementary files that are included and excluded for each archive being processed. The *included file log* lists the files that remain in the *raw* directories following the movie file removal process. The TSV formatted file contains the following fields:

1. Supplementary data directory
2. PMC identifier
3. URL of the file

The *excluded file log* is also in TSV format and lists the movie and archive files removed, with the following fields:

1. PMC identifier
2. URL of the movie removed (field 3 is null) **or** URL of the archive if movie file removed (field 3 has a value) **or** URL of archive removed if empty after movie removal (field 3 is null)
3. Filename of the movie removed from the archive in field 2

2.5 Stage 4: Supplementary file standardisation

Existing methods for converting a variety of file types into computer interpretable formats are used and extended in the workflow. The Auto-CORPus python package converts HTML text content to a BioC JSON format and HTML tables to a BioC-like JSON format where column and row orders are preserved along with cell contents. The SIBiLS documents and annotations fetch API provides text extracted from images and a separate OCR web service processes unseen images to extract text.

The Auto-CORPus package has been extended to process new file types:

- PDF files are processed using the 'marker-pdf' Python module (https://github.com/VikParuchuri/marker) to extract text which is then converted to BioC JSON.
- Word processor *.doc* and *.docx* files are processed with the 'python-docx' Python module (https://pypi.org/project/python-docx/) and the extracted text converted to BioC JSON.
- Presentation *.ppt*, *.pptx* and *.odp* files are processed using the 'python-pptx' Python module (https://pypi.org/project/python-pptx/) to extract text from slides which is stored in BioC JSON.
- Tables are extracted from spreadsheet *.xls* and *.xlsx* files, as well as from *.csv* and *.tsv* files, using the 'pandas' Python module (https://pandas.pydata.org/) to convert the loaded data frame object into a BioC-like tables JSON format.

Image *.jpg*, *.png* and *.tiff* files are initially queried using the SIBiLS fetch API (https://sibils.text-analytics.ch/api/fetch). If the image has previously been processed then the returned image text is stored in BioC JSON format. If the image has not previously been processed, then the image is sent to the SIBiLS OCR API (https://ocrweb.text-analytics.ch/) to generate new text which is structured and stored as BioC JSON.

The computer interpretable version of supplementary files are stored in the *processed* directory for each article, appending *_bioc.json* or *_table.json* to the filename. If there is an error during file processing, the file is added to the unprocessed file log. The TSV formatted file contains the following fields:

1. Supplementary data directory
2. PMC identifier
3. Filename
4. Error message

3 Next steps

During the development of the workflow each stage was tested separately. A pilot run of the full workflow included an additional article filter of a publication year after 2005 in order to limit the corpora sizes. The next step will be to apply the workflow to the full PMC BioC article collection, without size-limiting filters, to build an initial clinical case report full-text literature corpus and a clinical supplementary materials corpus (deliverable D2.2). Following this, the workflow will be used to update the corpus at regular intervals (deliverable D2.3). Additional

modules will be added to the workflow to enable comparison between the existing corpora and new/updated input PMC BioC files. Applying updates to the existing corpora versus rebuilding the corpora will reduce both the computational resources and time required to run the workflow to completion. The corpora will be processed in the WP3 semantic enrichment and the WP4 semantic structuring tasks.