Deliverable D2.2

Release of version 1 of the clinical supplementary material and CRF corpora

Project Title	FAIR-ification of Supplementary Data to Support Clinical Research
Project Acronym	FAIRClinical
WP No & Title	WP2: Data gathering and standardisation
Delivery Date	31/10/2024
Authors	Tim Beck (UK) Thomas Rowlands (UK)
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 FAIRClinical project aims to represent unstructured clinical case reports in a more structured format and improve the FAIR-ness of supplementary data files. Collections of case reports and example supplementary data files are therefore required for the development and testing of algorithms. Work package 2 (WP2) is tasked with building extensive full-text case report and supplementary file corpora which will be processed by other project work packages.

In deliverable D2.1 we described a workflow we have developed for downloading and standardising clinical case report literature and supplementary materials. An overview of the workflow, showing the inputs and outputs, is presented in Figure 1. The workflow involves four 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. Gathering and processing associated supplementary files
4. Standardising the format of supplementary files to support text analytics

Here we describe the execution of the workflow to create two output corpora, one corpus of full-text case report files and one corpus of supplementary files, along with the accompanying logs to track the provenance of the data.

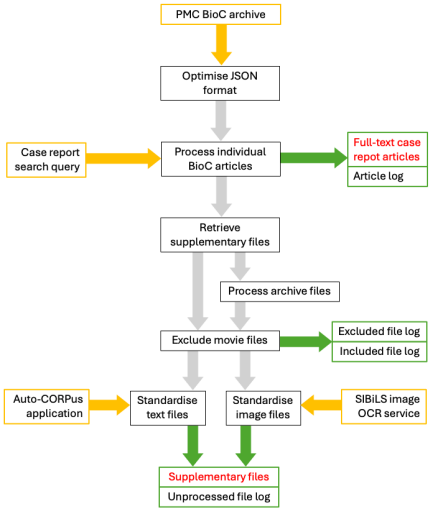

Figure 1: Overview of the FAIRClinical WP2 workflow.

2 Description of work accomplished

2.1 Overview

The workflow for downloading and standardising clinical literature and supplementary materials was run during October and November 2024, using the input PMC BioC archives available at that time from the NCBI website. The two output corpora were generated, along with the accompanying logs. In this report we describe the generated corpora and logs. This was the first end-to-end execution of the workflow and we identified areas for improvements. In this section we describe a change that was made to the format of the unprocessed file log and applied to this deliverable. In section *'3 Next steps'* we describe future changes that will be applied to the next version of the corpora (deliverable D2.3).

2.2 Execution of the workflow

Stage 1: PMC BioC full-text retrieval and processing

During stage 1 of the workflow, PubMed Central Open Access full-text articles in BioC JSON ASCII format are retrieved and used as the input for the workflow. These files are provided in eighteen compressed archives from an NCBI hosted FTP server (https://ftp.ncbi.nlm.nih.gov/pub/wilbur/BioC-PMC/). The archives are updated periodically and the archives with a timestamp of *2024-09-22* were used in this execution of the workflow. Table 1 gives the compressed size of each archive and the number of articles it contains. In total, 6,674,438 full-text articles were processed.

Archive	Archive file size	Number of articles
PMC000XXXXX_json_ascii.tar.gz	4.41GB	390,302
PMC030XXXXX_json_ascii.tar.gz	4.72GB	333,438
PMC035XXXXX_json_ascii.tar.gz	5.33GB	362,456
PMC040XXXXX_json_ascii.tar.gz	5.53GB	376,207
PMC045XXXXX_json_ascii.tar.gz	5.55GB	371,532
PMC050XXXXX_json_ascii.tar.gz	4.62GB	402,141
PMC055XXXXX_json_ascii.tar.gz	5.85GB	408,379
PMC060XXXXX_json_ascii.tar.gz	6.24GB	394,384
PMC065XXXXX_json_ascii.tar.gz	5.78GB	363,626
PMC070XXXXX_json_ascii.tar.gz	6.66GB	408,788
PMC075XXXXX_json_ascii.tar.gz	6.64GB	395,927
PMC080XXXXX_json_ascii.tar.gz	7.25GB	414,789
PMC085XXXXX_json_ascii.tar.gz	7.25GB	434,278
PMC090XXXXX_json_ascii.tar.gz	7.19GB	443,715
PMC095XXXXX_json_ascii.tar.gz	7.50GB	443,114
PMC100XXXXX_json_ascii.tar.gz	7.39GB	434,912
PMC105XXXXX_json_ascii.tar.gz	5.11 GB	295,519
PMC110XXXXX_json_ascii.tar.gz	16.6MB	931
Total number of articles		**6,674,438**

Table 1. The PMC BioC archives processed by the FAIRClinical workflow for standardising clinical literature and supplementary materials

Stage 2: Application of a search query

During stage 2 of the workflow, the PMC BioC files are queried to identify case report articles. The final full-text case report corpus is composed of files in the PMC BioC format. The use of a BioC 'key' file allows the creator of BioC files to specify details of how the data in the JSON file should be interpreted. PMC have defined a BioC key file and Figure 2 presents a concise version of this, which describes the fields present in the PMC BioC format used in the full-text case report corpus.

```
encoding:                ASCII
collection:              The Open Access subset of PMC articles.
source:                  PMC
date:                    yyyymmdd. Date articles downloaded from PMC.
document:                PMC article
id:                      PMC id
passage:                 One portion of a PMC article. Often a paragraph or subsection
title.
infon front:             Text is the title of the document.
infon abstract:          Abstract for the article.
infon abstract_title_1:  Title for a section of the abstract in a structured abstract.
infon title:             Section title for a section not part of the main text.
infon title_1:           Section title
infon title_2:           Subsection title, and so on. For example, title_3 would be a
subsubsection title.
infon fig_caption:       Caption for a figure.
infon fig_title_caption: The title of a figure caption, if it is explicitly indicated
in the XML. If the figure caption title is merely indicated by bold text, then it
appears as part of the caption.
infon footnote:          Text in a footnote in the article.
infon footnote_title:    Title of a footnote.
infon paragraph:         Ordinary paragraph of text in the article.  To understand
where in the article it belongs, the previous title passages would need to be tracked.
infon table_caption:     Caption for a table.
infon table:             Contents of a table. The elements of a table concatenated
together with space between them.
infon table_footnote:    Footnote to a table
infon table_title_caption: Title for the table caption.
infon ref:               A source referenced by the article. The text will be the title
of the source, if indicated in the XML.
offset:                  The first passage (title) has an offset of zero. Each
following passage has an offset increased by the number of bytes in the text of a
passage plus one space.
text:                    The ASCII text of the passage.
```

Figure 2. Excerpt from the PMC ASCII key file. Some field descriptions have been shortened. The full key file can be accessed at https://ftp.ncbi.nlm.nih.gov/pub/wilbur/BioC-PMC/pmc_ascii.key.

Stage 3: Supplementary file retrieval and processing

During stage 3 of the workflow, supplementary files are retrieved and processed. Due to the increased storage requirements for retaining movie files compared to other file types, movies are removed. Movies are defined as files with the following extensions: mp4, mov, avi, wmv, webm, flv, mpg, movi, m4v and 3gp. This applies to files that are directly downloaded and files that are contained in downloaded archives. All removed movie files are added to the excluded file log. Table 2 gives the number of directly downloaded movies and movies in archives that were excluded from the supplementary files corpus for each PMC set. In total, 2,971 directly downloaded movies and 2,915 movies in archives were excluded.

Stage 4: Supplementary file standardisation

During stage 4 of the workflow, a variety of supplementary file types are converted to computer interpretable formats. The Auto-CORPus package processes PDFs, word processor files, presentation files, spreadsheet files, and text-based formats for storing tabular data (e.g., CSV and TSV files). The SIBiLS documents and annotations fetch API provides text extracted from images and the SIBiLS optical character recognition (OCR) web service processes unseen images to extract text. The workflow outputs two JSON formats: the text is output in supplementary-BioC format and tables in a tables-JSON format. Figure 3 shows the key file for the supplementary-BioC format and Figure 4 shows the key file for the tables-JSON format.

PMC set	Number of directly downloaded	Number of movie files contained

	movie files excluded	in archives excluded
PMC000XXXXX	93	0
PMC030XXXXX	34	4
PMC035XXXXX	32	0
PMC040XXXXX	24	7
PMC045XXXXX	28	17
PMC050XXXXX	42	1
PMC055XXXXX	87	1
PMC060XXXXX	156	162
PMC065XXXXX	124	151
PMC070XXXXX	213	232
PMC075XXXXX	210	542
PMC080XXXXX	309	244
PMC085XXXXX	314	164
PMC090XXXXX	325	336
PMC095XXXXX	377	288
PMC100XXXXX	387	451
PMC105XXXXX	216	315
PMC110XXXXX	0	0
Total	**2,971**	**2,915**

Table 2. The movie files excluded from the supplementary files corpus.

source: Auto-CORPus (supplementary)
date: yyyymmdd. Date document processed by Auto-CORPus.
document: Supplementary text file
inputfile: Local path to the text file processed by Auto-CORPus
textsource: The name of the service used to provide the text.
id: Unique document identifier.
passage: One portion of a supplementary text file.
infon iao_name_1: Information Artifact Ontology label "supplementary material section"
infon iao_id_1: Information Artifact Ontology identifier "IAO:0000326"
offset: The first passage has an offset of zero. Each following passage has an offset increased by the number of characters in the text of a passage.
text: The text of the passage in UTF-8 character encoding

Figure 3. The Auto-CORPus supplementary-BioC key file

source: Auto-CORPus (supplementary)
date: yyyymmdd. Date document processed by Auto-CORPus.
document: Biomedical literature table
inputfile: Local path to the HTML file processed by Auto-CORPus.
textsource: The name of the service used to provide the text.
id: Unique table identifier.
passage: One portion of a biomedical literature table, such as a caption, table content (in columns and rows), or a footer. Details provided by section_type.
offset: The first passage has an offset of zero. Each following passage has an offset increased by the number of characters in the text of a passage.
infon section_title_1: A local label applied to table sections: table_title, table_caption, table_content, table_footer.
infon iao_name_1: An Information Artifact Ontology term name to describe the section type and the name for iao_id_1. An additional IAO name is iao_name_2, and so on.
infon iao_id_1: An Information Artifact Ontology term identifier to describe the section type and the identifier for iao_name_1. An additional IAO identifier is iao_id_2, and so on.
text: The text of the passage
column_headings: A column heading cell text and identifier pair. Column headings are in the order they occur in the table reading from left to right. Included when section_label = "table_content".
cell_text: Table column heading or data cell text. Included when section_label = "table_content".
cell_id: Unique cell identifier. Can be used to identify cells in annotations. In the format of table id, row id integer ("1" is the column header row) and column id integer ("1" is the left-most column) separated by periods e.g. "1_1.1.1". Included when section_label = "table_content".
data_section: One section of a table that shares column headings. Groups section titles and results rows. Included when section_label = "table_content".
table_section_title_1: Section title. Included when section_title_1 = "table_content".
data_rows: A data row cell text and identifier pair. Data cells are in the order they occur in the table reading from left to right. Included when section_label = "table_content".

Figure 4. The Auto-CORPus tables-JSON key file

If there is an error with processing a supplementary data file, the filename is added to the unprocessed file log. We have extended the format of the unprocessed file log from that described in D2.1. We required a more detailed log that identified the individual files in a downloaded archive that could not be processed, rather than simply identifying an error with the archive itself. We therefore added an optional field to the log which records the filename of the archive content. The TSV formatted unprocessed file log contains the following fields:

1. Supplementary data directory
2. PMC identifier
3. Filename of downloaded file
4. Filename of archive content (optional)
5. Error message

2.3 Generated corpora

The eighteen PMC sets are processed separately on a standard Windows desktop (*AMD Ryzen 7 7800X3D CPU, 32GB RAM)* with a minimum of 50G storage, taking approximately seven days to complete and generate the full corpora. After each set is processed, a suite of tests check the integrity and quality of the extracted case reports in BioC format, the downloaded supplementary files and the processed supplementary files. The accuracy of the logs are also tested. Since these v1 corpora are generated during the first end-to-end execution of the workflow, a failed test could mean a technical change to the workflow is required. In these cases, the workflow is rerun for a single set and the outputs compared against a benchmarked output for that set. The workflow is then resumed.

Table 3 shows the number of full-text articles and the number of downloaded supplementary files per PMC set. Table 4 shows the number of additional files that are extracted from downloaded archives per PMC set. The files classified as "Other" did not have a file extension that identified them as PDF, Word, Excel/table, PowerPoint, image or archive file. These included file types that could be identified by their file extension, such as HTML, XML, JSON and MD files, and files that did not include a file extension. If a file extension is not provided, the file is skipped during the processing step.

For each of the processed eighteen PMC sets, two out archives are created - one of full-text case report articles, and one of related supplementary files. One set (PMC110XXXXX_json_ascii.tar.gz) did not contain any case report articles, so output archives were not produced for this set. The final 34 output archives are made available to FAIRClinical researchers from the project's data storage space.

PMC set	Full-text articles	Downloaded supplementary files							
		PDF	Word	Excel/ table	Power point	Image	Archive	Other	Total
PMC000	4,945	5	32	2	1	21	4	1	**66**
PMC030	3,700	7	8	0	7	4	1	8	**35**
PMC035	3,932	5	27	2	2	20	1	15	**72**
PMC040	4,636	9	16	11	2	15	9	52	**114**
PMC045	5,261	22	20	2	1	16	2	10	**73**
PMC050	4,691	18	36	8	4	23	13	4	**106**
PMC055	6,088	23	68	13	2	12	8	59	**185**
PMC060	7,097	38	95	5	42	10	61	89	**340**
PMC065	6,562	75	137	29	74	40	53	95	**503**
PMC070	6,919	209	269	52	55	77	68	54	**784**
PMC075	8,452	317	265	61	67	83	233	23	**1,049**
PMC080	9,221	446	441	50	76	191	123	24	**1,351**
PMC085	9,998	404	442	102	103	220	104	20	**1,395**
PMC090	9,523	461	514	104	35	204	159	22	**1,499**
PMC095	9,106	666	393	93	38	353	152	22	**1,717**
PMC100	10,526	649	359	82	28	662	145	13	**1,938**
PMC105	7,996	576	286	34	7	167	68	7	**1,145**
PMC110	0	0	0	0	0	0	0	0	**0**
Total	**118,653**	**3,930**	**3,408**	**650**	**544**	**2,118**	**1,204**	**518**	**12,372**

Table 3. Numbers of full-text case reports and supplementary data files downloaded for each PMC set, after movie files are removed.

PMC set	Unprocessed archives	Archive contents						
		PDF	Word	Excel/ table	Power point	Image	Other	Total
PMC000	4	0	0	0	0	0	0	**0**
PMC030	0	0	1	0	0	0	0	**1**
PMC035	0	0	0	1	0	0	0	**1**
PMC040	2	0	0	48	0	0	1	**49**
PMC045	0	0	0	0	0	2	0	**2**
PMC050	0	0	0	0	0	2,749	17	**2,766**
PMC055	0	4	1	2	0	1	0	**8**
PMC060	9	0	14	10	28	5	38	**95**
PMC065	3	2	1	0	53	2	29	**87**
PMC070	7	17	3	1	54	15	12	**102**
PMC075	0	52	13	0	227	46	8	**346**
PMC080	3	21	21	9	89	45	26	**211**
PMC085	2	42	19	8	56	77	2	**204**
PMC090	3	81	24	169	103	481	198	**1,056**

Formatted Table

PMC095	9	54	20	17	85	66	66	**308**
PMC100	1	50	30	26	76	104	146	**432**
PMC105	0	38	21	5	4	44	2	**114**
PMC110	0	0	0	0	0	0	0	**0**
Total	**43**	**361**	**168**	**296**	**775**	**3,637**	**545**	**5,782**

Table 4. The contents of downloaded supplementary data archive files for each PMC set, after movie files are removed.

Table 5 shows the total number of supplementary files processed to BioC formats in version 1 of the supplementary files corpus. Where extractable text can not be found in a file, this is recorded in the unprocessed file log. Analysis of the unprocessed files found that reasons for failed processing involved file corruption (the file will not open), or the file not being compatible with the Python module used to process them. For example, a Word file produced using pre Microsoft Word 2007 can not be processed using the 'python-docx' module.

File type	Number in corpus	Number processed	Percentage processed
PDF	4,291	3,215	74.92%
Word	3,576	1,493	41.75%
Presentation	1,319	813	61.63%
Excel/table	946	389	41.12%
Images	5,755	1,343	23.33%

Table 5. File types processed to BioC formats in the supplementary files corpus

3 Next steps

3.1 Processing the unprocessed files

Some supplementary files are not processed because of limitations of the Python modules selected to process certain file types (modules used: marker-pdf, python-docx, python-pptx, pandas). There are limits on the age, provenance and size of some files. Other files are not processed because they do not have a file extension from which to identify the type of file it is. We will investigate using additional Python modules to improve the robustness of the workflow and increase the breath of the files that can be processed. This will involve implementing a method to attempt to predict file types when a file extension is absent.

3.2 Sentence splitting

The BioC standard spits text into passages (paragraphs). Further splitting at the sentence level is supported but optional. The full-text text report corpus uses the PMC BioC format which uses passage-level splitting. SIBiLS outputs a BioC format that is split at the sentence level. In order to eliminate the need for post-processing to become compatible with the SIBiLS output format, we will introduce sentence splitting in future versions of the corpora.

3.3 Version 2 of the clinical supplementary material and CRF corpora

The number of open source PMC articles is steadily increasing, and so too is the number of PMC BioC articles that are the input for the workflow. In order to avoid repeating the processing of articles and supplementary files that are in version 1 of the corpora, we will engineer new capability to only process unseen articles and files. Along with further optimisation of the workflow, this will reduce processing time and accelerate the release of version 2, and subsequent versions, of the corpora.