Deliverable D2.3

Release of version 2 of the clinical supplementary material and CRF corpora with ongoing automated updates

Project Title	FAIR-ification of Supplementary Data to Support Clinical Research
Project Acronym	FAIRClinical
WP No & Title	WP2: Data gathering and standardisation
Delivery Date	28/02/2025
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 work packages.

In deliverable D2.2 (Release of version 1 of the clinical supplementary material and CRF corpora) we described how a data aggregation, querying and standardisation workflow we presented in deliverable D2.1 (Workflow for downloading and standardising clinical literature and supplementary materials) was run to create two corpora. One corpus consists of full-text case report articles and the other consists of supplementary files associated with the case reports. Both corpora use the BioC JSON format.

Following a review of the version 1 (v1) corpora, we identified two areas for improvement. Firstly, there were limitations with the processing of supplementary text files to BioC format with 75% of PDFs, 42% of Word files, 62% of presentation files and 41% of Excel/table files being automatically processed. This was due to the requirements of Python modules used to process each type of file, which were restrictive with regards to file age, provenance and size. In order to improve the robustness of the workflow and increase the breadth of the files that can be processed, alternative modules could be evaluated and integrated into the workflow. Secondly, the text in BioC format was split at the passage (paragraph) level. The Swiss Institute of Bioinformatics Literature Services (SIBiLS) outputs a BioC format that is split at the sentence level. In order to eliminate any need for post-processing of the corpora to become compatible with the SIBiLS output format, the BioC text could be split by sentence. Here we describe the updates made to the workflow to accommodate these requirements. We also describe the resulting version 2 (v2) corpora generated from running the workflow on the latest literature collection, along with plans for ongoing corpora updates.

2 Description of work accomplished

2.1 Workflow update

Deliverable D2.1 describes our workflow for producing a clinical case report corpus and a supplementary materials corpus. This workflow was adapted to improve supplementary file processing and enable sentence splitting. Figure 1 presents an overview of the updated workflow.

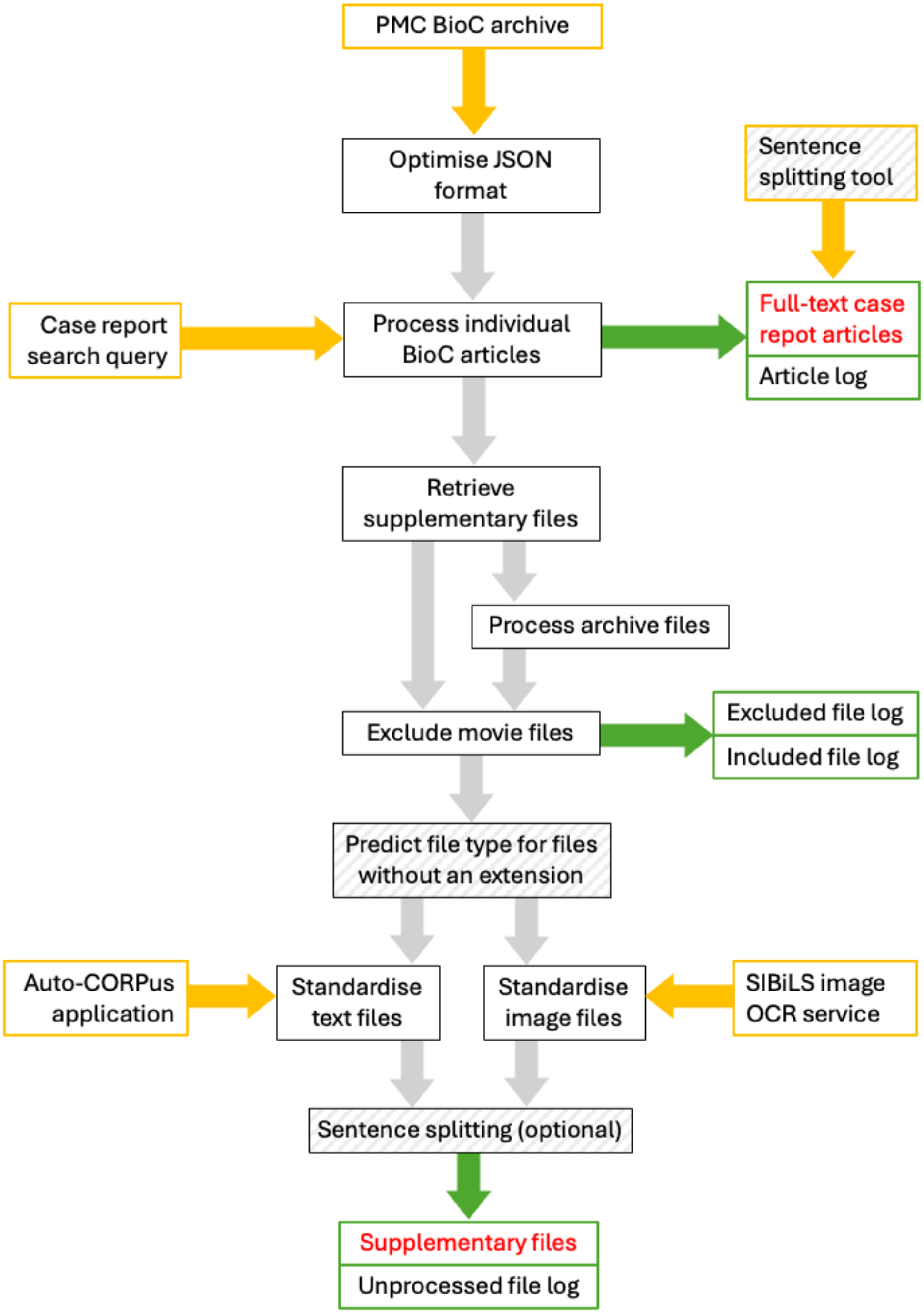


Figure 1: Overview of the updated workflow for downloading and standardising clinical literature and supplementary materials. New processes are displayed in hatched boxes.

2.1.1 Processing additional supplementary files

The algorithms for processing supplementary files were reviewed and optimised to speed up the execution of the workflow and process additional text files that were previously unprocessed. This resulted in the refactoring of the code responsible for the standardisation of text files and the use of updated and new python modules. This included the addition of the "rarfile" python module, and the non-python "unrar" software dependency, for improved rar archive handling. The presentation file handling was improved to process presentation files with *.ppt*, *.pptx*, *.pptm* and *.odp* file extensions. Previously, if a file name did not include a file name extension, the file was ignored. In a new workflow process, the file type is predicted for files without an extension. The "python-magic" python module is used to read the mime type of the file and determine the file type. For example, the "application/pdf" mime type identifies a file as a PDF which can then be processed accordingly. The "marker-pdf" python module is used to process PDFs. At the time of executing the workflow to generate v1 of the corpora, this module was at version 0.2.13. Later versions of this module have improved text detection which can extract PDF text contents that were previously missed. The marker-pdf module was upgraded to version 1.3.4 in the workflow. Finally, the process for extracting text from Word documents was updated by using Office on Windows, or Libre on Linux, to convert *.doc* extension files to *.docx* files. Correctly formatted *.docx* files can then be processed by the "python-docx" python module.

2.1.2 Sentence splitting BioC text

The workflow supports the option to split passages of text by sentences. The text from supplementary files can be split by sentences using a command line argument when executing the workflow. A standalone tool (BioC_utilities.py) can be used to split passages of text in the PMC BioC files from the full-text case report corpus. The SIBiLS sentence splitter algorithm, based on regular expression pattern matching, is implemented in the workflow to split passages of text. The sentence offsets are calculated on the length of characters, including any trailing whitespace characters, within each sentence.

2.1.2 Workflow code availability

The workflow is available on GitHub: https://github.com/FAIRClinical/ClinicalCorporaWorkflow.   In order to be able to reproduce both v1 and v2 of the FAIRClinical corpora, the updated workflow is available from the "Version-2" branch. The updated workflow has been tested on Windows 11 and Linux Mint.

The standalone sentence splitting tool requires an input directory of BioC JSON files and can be run from the command line, thus:

```
python ClinicalCorporaWorkflow/FAIRClinicalWorkflow/BioC_Utilities.py -i
<path_to_input_dir> -o <path_to_output_dir> -s
```

2.2 Execution of the updated workflow

The workflow for downloading and standardising clinical literature and supplementary materials uses input archives containing PubMed Central Open Access full-text articles in BioC JSON ASCII format. These archives are available from an NCBI hosted FTP server (https://ftp.ncbi.nlm.nih.gov/pub/wilbur/BioC-PMC/). The v1 corpora used PMC BioC archives timestamped June 2024 as input. At the time of executing the updated workflow in February 2025 to create the v2 corpora, the PMC BioC archives were timestamped September 2024. This was the second end-to-end execution of the workflow, utilising the improvements described above that were made based on results of the previous execution. In this section we describe the new output corpora and logs files generated by this execution.

As previously described (in Deliverable D2.1), The workflow involves 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

Stage 1: PMC BioC full-text retrieval and processing

The compressed size and number of files in each of the eighteen input archives downloaded were compared with the archives used as input to create the v1 corpora (see Deliverable D2.2). There were no differences between the archives used for v1 (timestamped June 2024) and the archives used in this run of the workflow (timestamped September 2024). In total, 6,674,438 full-text articles were processed.

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 corpus is organised into directories corresponding to each input archive. If sentence splitting of the text passages is required, these directories can be used as input for the standalone sentence splitting tool. A comparison of the v1 and v2 article log files showed that the same number of full-text articles had been identified, a total of 118,653 articles.

Stage 3: Supplementary file retrieval and processing

During stage 3 of the workflow, supplementary files are retrieved and processed. Movie files are removed to reduce the overall storage space required for the corpus. This applies to movie files that are directly downloaded and movie files that are contained in downloaded archives. Table 1 gives the number of directly downloaded movies and movies in archives that were excluded from the supplementary files corpus for each PMC set. A total of 5,890 movies were removed during this run of the workflow. Compared to the v1 corpora, an extra 4 movies were removed during the generation of the v2 corpora. These movies are contained within an archive from the PMC set PMC100XXXXX and were detected due to improved archive file handling.

PMC set	Number of directly downloaded movie files excluded	Number of movie files contained 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	455
PMC105XXXXX	216	315
PMC110XXXXX	0	0
Total	**2,971**	**2,919**

Table 1. The movie files excluded from the supplementary files corpus.

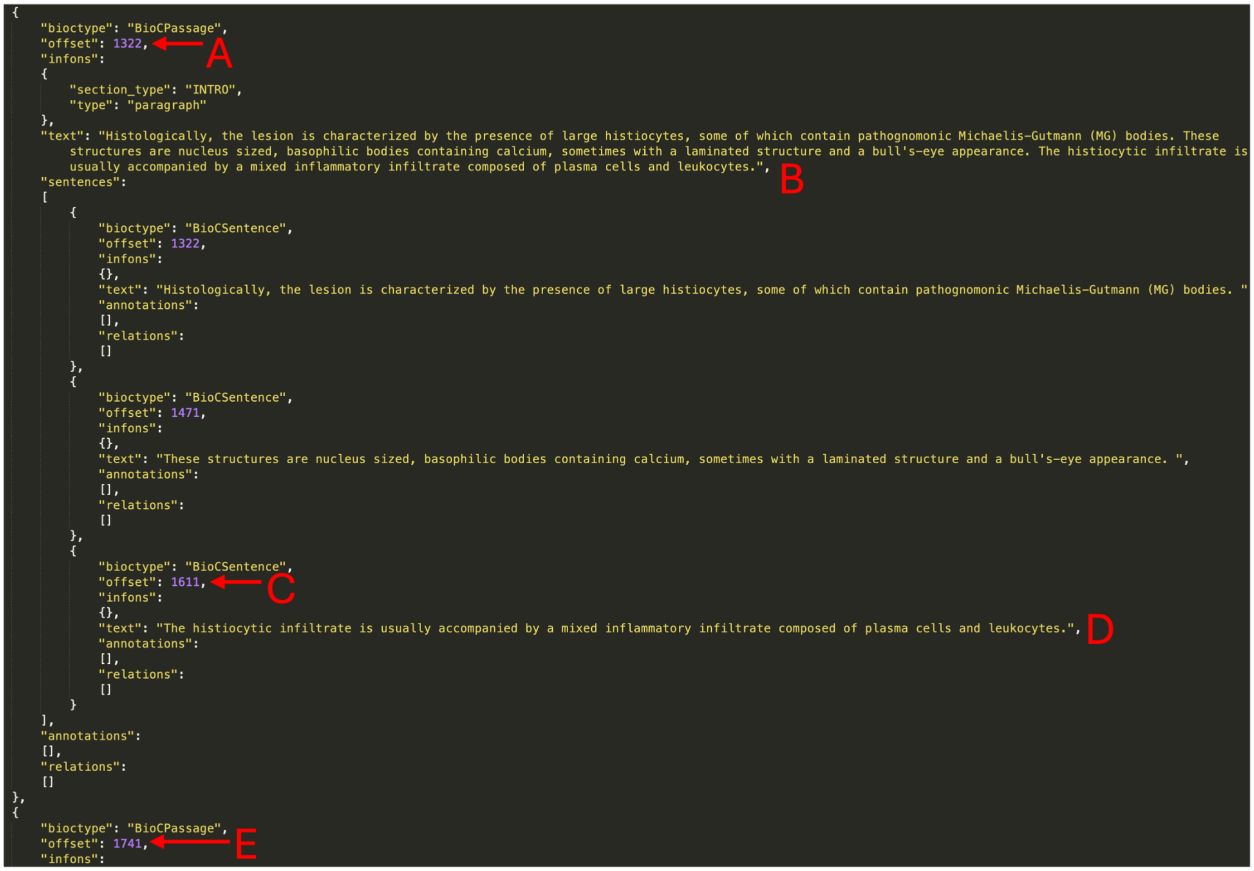
Stage 4: Supplementary file standardisation

During stage 4 of the workflow, a variety of supplementary file types are converted to computer interpretable formats using the Auto-CORPus package for text files and SIBiLS online services for image files. The workflow outputs two JSON formats: the text is output in supplementary-BioC format and tables in a tables-JSON format. During this run of the workflow the sentence splitting option was selected.

2.3 Generated v2 corpora

Since the workflow has been updated (as described in Section 2.1), the v2 corpora was created from scratch instead of extending the v1 corpora. This is because supplementary files that had previously been unprocessed would be reanalysed and potentially now processed due to the workflow changes. The eighteen input PMC BioC archives were 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 was processed, a suite of tests checked 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 log files are also tested.

The outputs from running the standalone sentence splitting tool over full-text articles were assessed for completeness and accuracy. During reviewing the alignment of passage and sentence offsets, it was found that there is a discrepancy between passage-level offsets provided in the original PMC BioC file and the number of characters within each passage. Figure 2 presents an example of this, where the passage offset and passage character length do not match the offset for the next passage. The same is true for the offset for the final sentence in a passage and the sentence character length; these do not match the offset for the first sentence of the next passage. We did not change the original passage offsets in order to ensure compatibility with existing annotations made using the PMC BioC passage offsets.



```
{
    "bioctype": "BioCPassage",
    "offset": 1322,
    "infons":
    {
        "section_type": "INTRO",
        "type": "paragraph"
    },
    "text": "Histologically, the lesion is characterized by the presence of large histiocytes, some of which contain pathognomonic Michaelis-Gutmann (MG) bodies. These
        structures are nucleus sized, basophilic bodies containing calcium, sometimes with a laminated structure and a bull's-eye appearance. The histiocytic infiltrate is
        usually accompanied by a mixed inflammatory infiltrate composed of plasma cells and leukocytes.",
    "sentences":
    [
        {
            "bioctype": "BioCSentence",
            "offset": 1322,
            "infons":
            {},
            "text": "Histologically, the lesion is characterized by the presence of large histiocytes, some of which contain pathognomonic Michaelis-Gutmann (MG) bodies. "
            "annotations":
            [],
            "relations":
            []
        },
        {
            "bioctype": "BioCSentence",
            "offset": 1471,
            "infons":
            {},
            "text": "These structures are nucleus sized, basophilic bodies containing calcium, sometimes with a laminated structure and a bull's-eye appearance. ",
            "annotations":
            [],
            "relations":
            []
        },
        {
            "bioctype": "BioCSentence",
            "offset": 1611,
            "infons":
            {},
            "text": "The histiocytic infiltrate is usually accompanied by a mixed inflammatory infiltrate composed of plasma cells and leukocytes.",
            "annotations":
            [],
            "relations":
            []
        }
    ],
    "annotations":
    [],
    "relations":
    []
},
{
    "bioctype": "BioCPassage",
    "offset": 1741,
    "infons":
```

Figure 2: Discrepancy in offsets between passages using an example PMC BioC JSON file (PMC8000235) which has been split by sentences. A: passage offset, B: passage text of 414 characters, C: final sentence offset, D: final sentence text of 125 characters, E: next passage offset. The next passage offset (E) would be expected to be 1736 (instead of the given offset of 1741) based on passage offset and length (1322 + 414) and final sentence offset and length (1611 + 125).

Table 2 shows the number of full-text articles and the number of downloaded supplementary files per PMC set. Table 3 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 mime type is checked in an attempt to decipher the original file type. There are 37 files downloaded without a file extension in the v2 corpora compared to 52 files without a file extension in v1.

Table 4 shows the total number of supplementary files processed to BioC formats in the v2 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. This functionally was available during the generation of the v1 corpora, however the code refactoring and module improvements described in section 2.1.1 resulted in a better rate of extraction overall during the generation of the v2 corpora. A total of 3,321 files were unprocessed in v1 and 2,141 files were unprocessed in v2.

For each of the processed eighteen PMC sets, two output 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	23	**1,718**
PMC100	10,526	650	359	82	27	662	146	13	**1,939**
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**	**3,931**	**3,408**	**650**	**543**	**2,118**	**1,205**	**519**	**12,374**

Table 2. 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	**348**
PMC080	3	21	21	9	89	45	26	**212**
PMC085	2	42	19	8	56	77	2	**214**
PMC090	3	81	24	169	103	481	198	**1,056**
PMC095	9	54	20	17	85	66	66	**309**
PMC100	1	77	30	26	77	104	146	**434**
PMC105	0	38	21	5	4	44	2	**114**

PMC110	0	0	0	0	0	0	0	**0**
Total	**43**	**388**	**168**	**296**	**776**	**3,637**	**545**	**5,798**

Table 3. The contents of downloaded supplementary data archive files for each PMC set, after movie files are removed.

File type	**Number in corpus**	**Number processed**	**Percentage processed**
PDF	4,319	3,371	78.05%
Word	3,576	1,614	45.13%
Presentation	1,319	1,305	98.93%
Excel/table	946	529	55.91%
Images	5,755	1,348	23.42%

Table 4. File types processed to BioC formats in the supplementary files corpus.

3 Ongoing updates

This second release of the FAIRClinical corpora introduced improvements to the automatic processing workflow, resulting in more files being processed compared to v1 whilst the number of supplementary material items available remained largely the same, with just 2 additional supplementary files downloaded. A total of 7,944 supplementary files were standardised successfully for this output, whilst 7,253 were standardised in the previous v1 corpora release (D2.1).

Improvements made to the workflow codebase are reflected in an increase in the types of files that can be processed, with the increase in the number of presentation file types recognised providing the most significant improvement. The FAIRClinical corpora will continue to be updated with new periodic releases. We will modify the workflow so that it can be run on a single PMC set, to allow only updated or new PMC sets to be processed in the future. We will also ensure the workflow can be run on Mac (in addition to Windows and Linux) to support the reuse of the software by a wider user base. Algorithms for processing text files, in particular PDFs, are regularly improving and will be updated in the Auto-CORPus package which is used for text processing in this workflow. In order to realise the benefits of these types of updates to the workflow, we will reprocess previously seen supplementary files in addition to newly published case reports.