Part of

FAIR-ification of Supplementary Data to Support Clinical Research

Description

CHIST-ERA Data Management (2025)

Description

This is the updated version of Data Management Plan (DMP) for the FAIRClinical Project.

Researchers

Anaïs Mottaz (0000-0003-0080-9451), Tim Beck (0000-0002-0292-7972), Venkata Satagopam (0000-0002-6532-5880), Patrick Ruch (0000-0002-3374-2962), Nona Naderi (0000-0002-1272-7640), Thomas Rowlands (0000-0002-7912-4203), Basel Alshaikhdeeb (0000-0002-7518-2676)

AFL-3.0 DOI: - 28/02/2025

Funder	Grant
CHIST-ERA||CHIST-ERA	FAIRClinical (CHIST-ERA-22-ORD-02)

Description

1 Data Description And Collection Or Reuse of Existing Data

1.1 What data (for example the kind, formats, and volumes), will be collected or produced?

1.1.1 Give details on the kind of data

Secondary data

Derived or compiled (e.g., text mining, 3D models)

1.1.2 Give details on the data format

FairClinical is targeting secondary data of publicly open-accessed scientific publications where the aim specifically is to process supplementary data attached to such publications. Supplementary data files are provided by journal publishers in a range of formats, including PDFs, word processor documents (i.e., doc or docx), spreadsheets (i.e., xls or xlsx), images (i.e., jpg, gif, png, TIF, tiff), plain text files (i.e., txt, csv, xml, rtf, html), and movies. The BioC format will be used in FAIRClinical for the storage and exchange of a full-text corpus of published clinical case reports. Note that, movies will be excluded, however, as the technology to automatically transcribe audio from movie files improves, such files can be reconsidered in future versions of the corpus.

1.1.3 Justify the use of certain formats

widely supported format across disciplines

While researchers primarily access biomedical literature through web browsers that display HTML or PDF formats of publications, these formats are not optimized 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. 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. The PMC BioC key file defines the semantics associated with the BioC data. In order to optimize the formats of supplementary files 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 has been extended to transform PDFs, word processor documents, spreadsheets, presentations, 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.

1.1.4 Give details on the volumes

GB (gigabyte)

The collected data is going to be in Gigabytes due to the volume of publications and the variety of data formats of supplementary data that will be tackled. During the first stage 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. In total, 6,674,438 full-text articles were processed which have a size of 119.62 Gigabytes.

1.2 How will new data be collected or produced?

1.2.2 Explain how data provenance will be documented

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)

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

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.

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. 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. 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".

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 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.

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

2.1.1.4 Consider what other documentation is needed to enable re-use

FairClinical repositories (i.e., GitHub, GitLab) would contain the source code along with references of any other APIs or existing models that might be involved.

3 Reused Data

3.1 Reused Data

3.1.1 How will existing data be re-used?

• To reproduce and validate findings

• To compare and combine with other data

• To follow-up research on a specific area

• To develop new products/ services

• To contribute to the wider community (researchers, public authorities, citizen scientists)

Section 1.2.2

3.1.2 Where can re-used data be found?

APIS

SIB Literature Services (SIBiLS)

3.1.3 Which data will be re-used?

Clinical Case Report Forms (CRFs)

4 Storage And Backup During The Research Process

4.1 Storage And Backup During The Research Process

4.1.1 How will data security and protection of sensitive data be taken care of during the research?

4.1.1.1 Explain how the data will be recovered in the event of an incident and describe the main risks and how these will be managed

Multiple cloud systems are going to be used along with virtual machines

where the deposition of produced data will be distributed.

5 Legal And Ethical Requirements, Codes Of Conduct

5.1 Legal And Ethical Requirements, Codes Of Conduct

5.1.2 How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?

5.1.2.1 Data ownership and accessibility

5.1.2.1.1 Who will be the owner(s) of the data?

Haute Ecole Spécialisée de la Suisse Occidentale

Patrick Ruch

5.1.2.1.2 Explain what access or restrictions will apply to the data?

Open

Since the data is public and open-accessed hence, there will be no

restricted access.

5.1.2.2 Intellectual property rights

5.1.2.2.1 Explain which intellectual property and how will they be dealt with

Copyright

The software code and written documentation will be protected under copyright law, ensuring that ownership and distribution rights are maintained

5.1.2.3 Third-party data restrictions

5.1.2.3.1 Are there any restrictions on the re-use of third-party data?

No

5.1.3 Ethical issues

5.1.3.1 What ethical issues and codes of conduct are there, and how will they be taken into account?

Other

No ethical issues

6 Data Sharing And Long-term Preservation

6.1 Data Sharing And Long-term Preservation

6.1.1 How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?

6.1.1.1 Explain how the data will be discoverable and shared

• Deposit in a FAIR-enabling data repository

• Indexed in a catalogue

6.1.1.2 Outline the plan for data preservation and give information on how long the data will be retained

SIBiLS is hosting the publications, whereas, university of Luxembourg will

host the OMOP output and the source code will be stored in github/gitlab.

6.1.1.3 Explain when the data will be made available

By the end of the FairClinical project, the OMOP will be available.

6.1.1.4 Indicate the expected timely release

2026-12-01

6.1.1.5 Will exclusive use of the data be claimed?

No

6.1.1.7 Indicate how the data will be used

The enriched metadata of the clinical case report form will be deposited into the SIBiLS which will enable future researchers to search for such CRFs.

6.1.1.8 Is it necessary to restrict access to certain communities or to apply a data sharing agreement?

No

6.1.2 How will data for preservation be selected, and where data will be preserved long-term?

6.1.2.1 Indicate what data must be retained or destroyed for contractual, legal, or regulatory purposes

Retained

All data wil be retained and no need for destroying

6.1.2.2 Indicate how it will be decided what data to keep

None

6.1.2.3 Describe the data to be preserved long-term

Derived or compiled (e.g., text mining, 3D models)

6.1.2.4 Indicate where the data will be deposited

SIBiLS

6.1.2.5 Indicate how the data will be shared

Repository

6.1.2.6 Indicate whether potential users need specific tools to access and (re-)use the data.

Search API

6.1.3 How will the application of a unique and persistent identifier to each data set be ensured?

6.1.3.1 What type of persistent identifier (PID) will be used?

• DOI

• URI

7 Data Management Responsibilities And Resources

7.1 Data Management Responsibilities And Resources

7.1.1 Who will be responsible for data management?

7.1.1.1 Outline the roles and responsibilities for data management/stewardship activities

Patrick Ruch

Giving access to virtual machines, APIs and enabling deposition

7.1.2 What resources will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

7.1.2.1 Explain how the necessary resources to prepare the data for sharing/preservation have been costed in

Use of institution infrastructure