

# Automated Pipeline for XNAT Data Bulk Export

N.G. van der Velden, J.J. Uitterdijk, G.J. Pelgrim, M. Veening,  
R. Vliegenthart, P.M.A. van Ooijen

dash@umcg.nl



Lung cancer, chronic obstructive pulmonary disease, and cardiovascular disease, the so-called Big-3 (B3), are responsible for high rates of morbidity and mortality. B3CARE is a research collaboration project, with a final ambition to establish an integrated B3 computed tomography (CT) screening program and accelerate software development for CT imaging data post-processing. A large-scale, high-quality imaging data biobank is established containing data from different projects all stored in the XNAT opensource platform. The data can be accessed via the Virtual Research Workspace, a two-way verification secured collaboration environment for conducting data analyses jointly and securely.

For the development and implementation of machine learning (ML) solutions with large datasets, the lack of integration causes a slow uptake for the data to actually being used. The full CT datasets are too space-consuming to be stored in the development environment, while manually selecting and downloading specific subject scans from the XNAT server is time consuming.

```
Administrator: Anaconda Prompt
(xnat_bulk_export) H:\My Desktop\Automated ML Pipelines in XNAT\xnat_bulk_data_export>python xnat_bulk_export.py -f H:\My Desktop\Patientnummers_10.xlsx -o H:\My Desktop\CT Scans -p Imalife3 -t Fl_Thorax INSP 3.0 Qr36 -u veldennng
XNAT account password:
Password correct.
Downloading Fl_Thorax INSP 3.0 Qr36 scan files of subject 125586 to H:\My Desktop\CT Scans\125586\F1_Thorax INSP 3.0 Qr36
Downloading Fl_Thorax 3.0 Qr36 scan files of subject 339193 to H:\My Desktop\CT Scans\339193\F1_Thorax 3.0 Qr36
Downloading Fl_Thorax INSP 3.0 Qr36 scan files of subject 370941 to H:\My Desktop\CT Scans\370941\F1_Thorax INSP 3.0 Qr36
Downloading Fl_Thorax INSP 3.0 Qr36 scan files of subject 830219 to H:\My Desktop\CT Scans\830219\F1_Thorax INSP 3.0 Qr36
Downloading Fl_Thorax INSP 3.0 Qr36 scan files of subject 563995 to H:\My Desktop\CT Scans\563995\F1_Thorax INSP 3.0 Qr36
Downloading Fl_Thorax INSP 3.0 Qr36 scan files of subject 136470 to H:\My Desktop\CT Scans\136470\F1_Thorax INSP 3.0 Qr36
Downloading Fl_Thorax 3.0 Qr36 scan files of subject 731797 to H:\My Desktop\CT Scans\731797\F1_Thorax 3.0 Qr36
Downloading Fl_Thorax 3.0 Qr36 scan files of subject 356216 to H:\My Desktop\CT Scans\356216\F1_Thorax 3.0 Qr36
Downloading Fl_Thorax INSP 3.0 Qr36 scan files of subject 190714 to H:\My Desktop\CT Scans\190714\F1_Thorax INSP 3.0 Qr36
Disconnecting from the server...
```

As a solution, we started the development of a Python-based Application Programming Interface (API) that functions as a tool to export a requested DICOM data bulk to a specified network location over a reliable data stream. The API currently works with a command-line interface to construct a pipeline that communicates with XNAT through its Representational State Transfer Application Program Interface (REST API). The user provides a list for the selection of subjects (e.g., CSV, XLSX format), while the other settings can be filled in manually. The performance of the tool was tested with anonymized CT imaging data from the Imalife study (Imaging in Lifelines cohort). A method to establish a desired pipeline for ML purposes will not only grant convenience to researchers, but also helps to ensure compliance with the FAIR principles.