OPEN DATA MICCAI 2024

Unlock Medical Machine Learning with Open Data
7 October 2024
Palmeraie Convention Centre - Marrakesh, Morocco

AFRICAI Imaging Repository White Paper

Martijn P. A. Starmans[1,2] & Apostolia Tsirikoglou[3]

[1]Department of Radiology & Nuclear Medicine, Erasmus MC University Medical Center, Rotterdam, the Netherlands
[2]Department of Pathology, Erasmus MC Cancer Institute, Rotterdam, the Netherlands
[3]Department of Oncology-Pathology, Karolinska Institutet, Stockholm, Sweden

1. Introduction

As the landscape of medical machine learning continues to evolve, access to diverse, representative, and inclusive datasets becomes paramount for addressing global healthcare challenges effectively. The *Open Data* session of the MICCAI 2024 congress[1] aims to empower collaboration and innovation by exploring and utilizing new datasets.

In its dawn, *Open Data* focuses particularly on underrepresented populations and diseases. Despite significant progress in medical imaging research, there remains a critical gap in the availability of diverse datasets, especially from regions such as the African continent. By highlighting datasets from these populations, we strive to address disparities in healthcare and promote inclusivity in machine learning research.

Open Data aims to create a publicly available repository - AFRICAI - of high-quality data, offering hosting services and a session with talks and publications. It is closely related to the principles of *Data-centric AI*, where data and its systematic engineering are at the forefront of model development. After all, there are no *"good"* models without *"good"* data[2].

In this white paper, we briefly explain the AFRICAI repository (2. The AFRICAI repository), provide a tutorial on how to apply for storing your data in the repository and how to upload your data (3. Tutorial: how to upload data), additional information on the costs which are currently free of charge due to sponsoring (4. Costs and sponsoring), and discuss some current limitations and future work how to address these (5. Limitations and future work).

2. The AFRICAI repository

The AFRICAI repository is based on a two-fold solution. Primarily, the Euro-BioImaging XNAT repository will be used, with Zenodo as back-up if the storage could potentially not be maintained anymore (see Limitations and future work).

AFRICAI repository: Euro-BioImaging Medical Imaging Archive XNAT

Background

Euro-BioImaging is a European research infrastructure that offers open access to imaging technologies, training and data services in biological and biomedical imaging. The infrastructure consists of several recognized facilities that offer specialized services, called Nodes. One of these nodes is the Population Imaging Flagship Node Rotterdam[3], which is an initiative of the Dutch

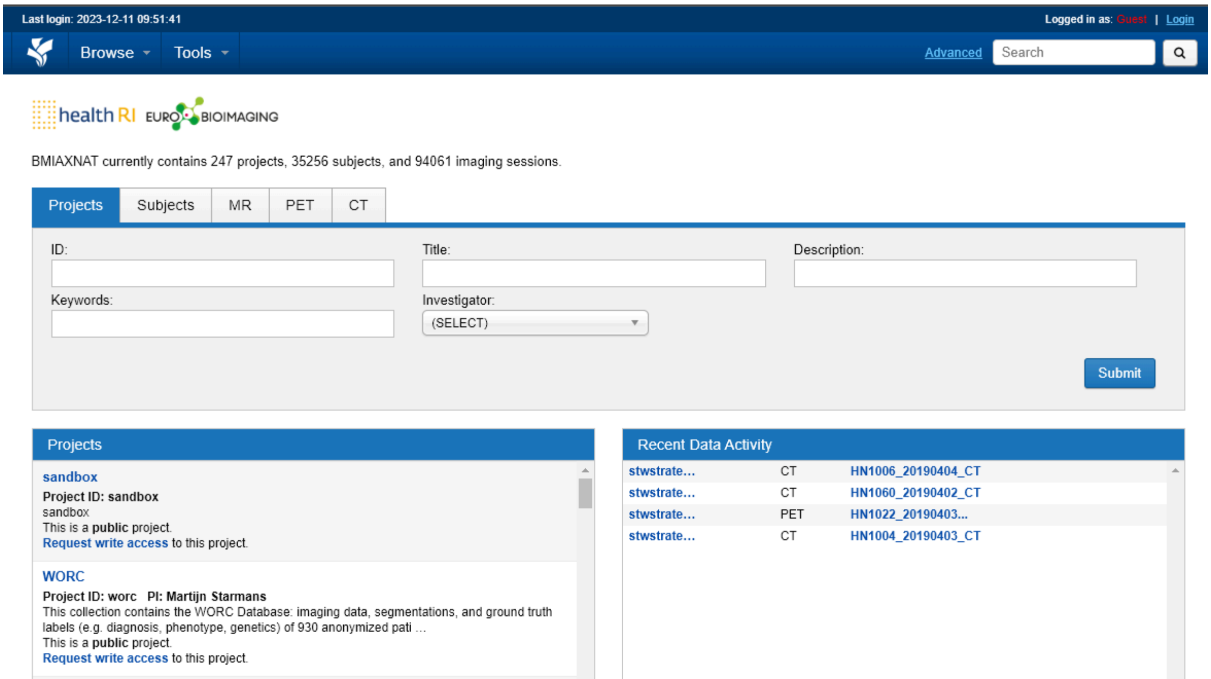
[1] https://conferences.miccai.org/2024/en/
[2] "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI
[3] https://www.eurobioimaging.eu/nodes/population-imaging-flagship-node-rotterdam

Federation of University Medical Centres (NFU) and the Department of Radiology & Nuclear Medicine, Erasmus MC, University Medical Centre Rotterdam.

One of the services of the Node is the Euro-BioImaging Medical Imaging Archive XNAT. The archive is operated by Health-RI[4], the Dutch National Research Infrastructure, with the BIGR[5] group of the Erasmus MC providing second line support and development. The Euro-BioImaging Medical Imaging Archive is among others used in the EU Horizon projects EuCanImage, EOSC4Cancer, RadioVal, and EUCAIM. Due to its proven widespread use, specific design for medical imaging data, and sustainability, the Euro-BioImaging archive is perfectly suitable as AFRICAI storage.

Description and information

The Euro-BioImaging Medical Imaging Archive can be accessed on the web address https://xnat.bmia.nl/. The archive is based on XNAT[6], an open source software solution specifically designed for storing and managing medical imaging data[7,8]. See the figure below for a screenshot of the front page.



Last login: 2023-12-11 09:51:41

Logged in as: Guest

BMIAXNAT currently contains 247 projects, 35256 subjects, and 94061 imaging sessions.

ID:

Title:

Description:

Keywords:

Investigator: (SELECT)

Projects

sandbox
Project ID: sandbox
sandbox
This is a public project.
Request write access to this project.

WORC
Project ID: worc PI: Martijn Starmans
This collection contains the WORC Database: imaging data, segmentations, and ground truth labels (e.g. diagnosis, phenotype, genetics) of 930 anonymized pati ...
This is a public project.
Request write access to this project.

Recent Data Activity

stwstrate...	CT	HN1006_20190404_CT
stwstrate...	CT	HN1060_20190402_CT
stwstrate...	PET	HN1022_20190403...
stwstrate...	CT	HN1004_20190403_CT

Figure 1: Screenshot of the front-page of the Euro-BioImaging Medical Imaging Archive

For more detailed information, we refer to the following publicly available information:

[4] https://www.health-ri.nl/en

[5] https://bigr.nl/

[6] https://xnat.org

[7] Marcus, D.S., Olsen, T.R., Ramaratnam, M. et al. The extensible neuroimaging archive toolkit. Neuroinform 5, 11–33 (2007). https://doi.org/10.1385/NI:5:1:11

[8] Herrick, R., Horton, W., Olsen, T., McKay, M., Archie, K. A., & Marcus, D. S. (2016). XNAT Central: Open sourcing imaging research data. NeuroImage, 124, 1093-1096.

- The Health-RI documentation: https://www.health-ri.nl/en/services/xnat
- Deliverable 3.2 of EuCanImage, which gives more detailed information on access and support: https://eucanimage.eu/wp-content/uploads/2022/07/D3.2.pdf.

Data model

XNAT has a custom data and metadata model. An overview of the XNAT data model is depicted in Figure 2. For brevity, we will only describe the high-level components of the XNAT metadata schema here. For detailed information, we refer the reader to the official XNAT documentation[9].

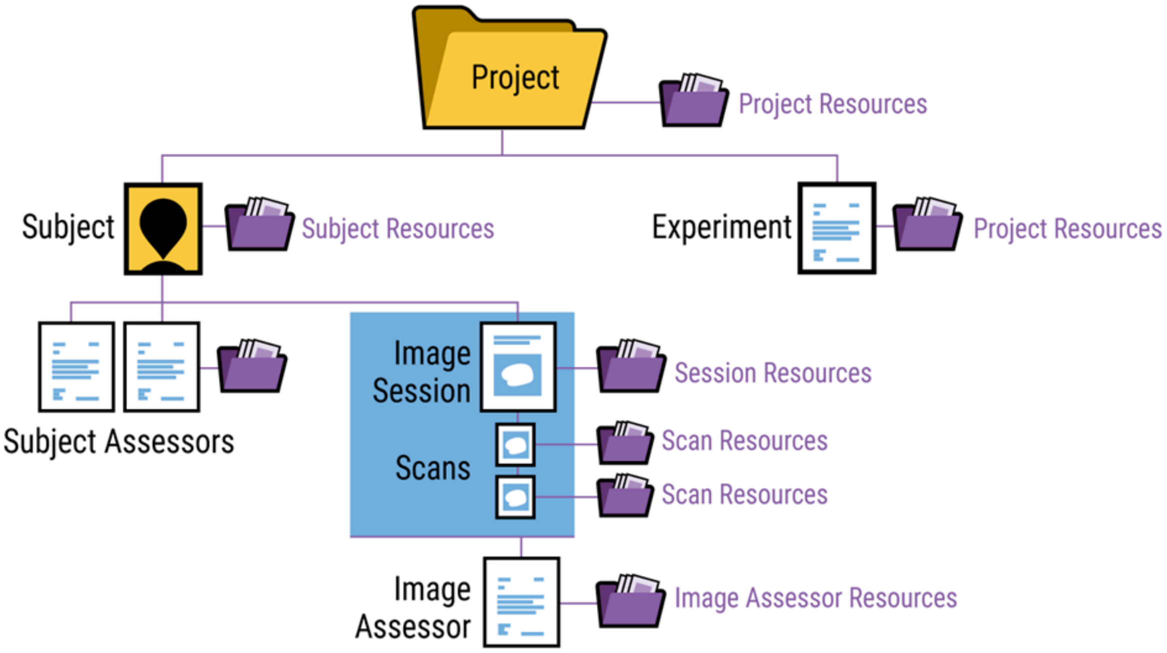


Figure 2: Current XNAT metadata schema. Adopted from https://wiki.xnat.org/documentation/how-to-use-xnat/understanding-the-xnat-data-model.

Project: In XNAT, a project is used to define a collection of data, usually a single (multi-site) study. Permissions for users are managed by the owner of a project, which can give users access to a project with different roles (Owner, Member, & Collaborator) or parts of a project. For the AFRICA repository, each dataset / study will be stored in a separate project. During development of the dataset, we recommend you keep the project private, hence only you as data controller will have access. Once you want to make the data publicly available, you can change the whole project or parts of the project to public.

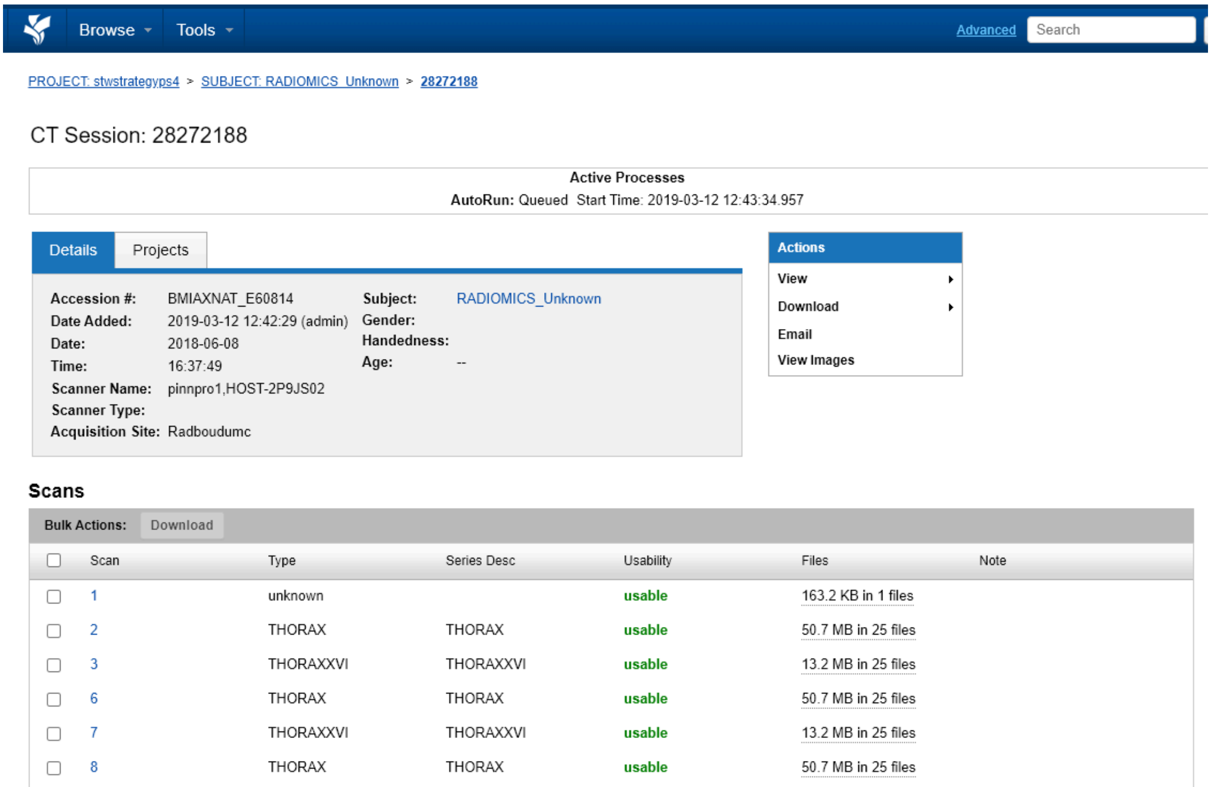
Subject: A subject in XNAT corresponds with an actual subject of a study. Subjects can be registered with other projects and thus be shared between projects. Subjects cannot exist outside of a project.

[9] https://wiki.xnat.org/documentation/how-to-use-xnat/understanding-the-xnat-data-model

Experiment: An experiment corresponds to an event by which data is required. For imaging data, this is usually a scan session, e.g. a CT or MRI session. Non-imaging experiments can also be defined. Experiments can both be defined on a project level, or on a subject level, depending on the type of experiment. Similar to subjects, experiments can be shared between projects and cannot exist outside a project.

Scan: A scan corresponds to a single scanning sequence, e.g. a T1-weighted MRI scan. Typically, an imaging session will consist of multiple scans.

An example of an experiment is shown in Figure 3.



PROJECT: stwstrategyps4 > SUBJECT: RADIOMICS_Unknown > 28272188

CT Session: 28272188

Active Processes
AutoRun: Queued Start Time: 2019-03-12 12:43:34.957

Accession #: BMIAXNAT_E60814
Date Added: 2019-03-12 12:42:29 (admin)
Date: 2018-06-08
Time: 16:37:49
Scanner Name: pinnpro1,HOST-2P9JS02
Scanner Type:
Acquisition Site: Radboudumc

Subject: RADIOMICS_Unknown
Gender:
Handedness:
Age: --

Scans

☐	Scan	Type	Series Desc	Usability	Files	Note
☐	1	unknown		usable	163.2 KB in 1 files	
☐	2	THORAX	THORAX	usable	50.7 MB in 25 files	
☐	3	THORAXXVI	THORAXXVI	usable	13.2 MB in 25 files	
☐	6	THORAX	THORAX	usable	50.7 MB in 25 files	
☐	7	THORAXXVI	THORAXXVI	usable	13.2 MB in 25 files	
☐	8	THORAX	THORAX	usable	50.7 MB in 25 files	

Figure 3: Example of an experiment, a CT session in this case, with multiple scans, found at https://xnat.bmia.nl/data/projects/stwstrategyps4/subjects/RADIOMICS_Unknown/experiments/28272188.

For DICOM data, the XNAT metadata and data models are well-defined and proven. There are specific resources defined for DICOM, API calls for DICOM, metadata is automatically extracted from DICOM upon ingestion and filled in in the XNAT data model. In the resources parts however, users can put any data they want in any format: the data model is flexible in this aspect. For the AFRICAI repository, we recommend the following structure for the resources in a scan:

Scan X

- ○ DICOM (by default in XNAT data model)
- ○ SNAPSHOT (by default in XNAT data model)

- Image (optional)
    - image.nii.gz

- Annotations (optional)
    - segmentation_{contour}_{observer}.nii.gz

For each resource, we created a standardized folder with standardized filenames, including some variables for naming certain contours and observers. Additional resources can be added in a similar way.

Interacting with your data in Python: xnatpy

It is recommended to use the Python client library xnatpy to interact with XNAT through the REST API. The documentation for xnatpy can be found here: https://xnat.readthedocs.io. By using this library the interaction with XNAT can be integrated in a programmatic/automated environment more fluently.

Back-up: Zenodo

Zenodo[10] is one of the most widely used platform for publishing any type of data. It does not offer any of the imaging-specific functionalities of XNAT, e.g., no data model, no viewer, no rest API to access the data. However, it is widely used and freely available.

3. Tutorial: how to upload data

For two examples of how your dataset can be organized in XNAT, we recommend looking at the WORC database[11] or the STW Strategy Head & Neck dataset[12].

This tutorial details the technical parts of how to upload your data to the AFRICA repository. Beforehand, you should make sure you have approval from the relevant ethical (e.g., institutional review board) and legal parties from your institute. For tips on this aspect, we refer you to our AFRICAI webinar which was specifically organized on legal and ethical issues for the AFRICAI repository: https://www.youtube.com/watch?v=FnAiH4AAE0g.

[10] https://zenodo.org/

[11] Starmans, M. P., Timbergen, M. J., Vos, M., Padmos, G. A., Grünhagen, D. J., Verhoef, C., ... & Klein, S. (2021). The WORC database: MRI and CT scans, segmentations, and clinical labels for 930 patients from six radiomics studies. medRxiv, 2021-08. https://doi.org/10.1101/2021.08.19.21262238, https://xnat.bmia.nl/data/projects/worc

[12] Aerts, H. J., Velazquez, E. R., Leijenaar, R. T., Parmar, C., Grossmann, P., Carvalho, S., ... & Lambin, P. (2014). Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature communications, 5(1), 4006., https://doi.org/10.1038/ncomms5006, https://xnat.bmia.nl/data/projects/stwstrategyhn1

1. Requesting a project on the AFRICAI repository

As we currently cannot offer unlimited space, first a request for a proposal needs to be send to both of the MICCAI 2024 Open Data Chairs: Apostolia Tsirikoglou (apostolia.tsirikoglou@ki.se) and Martijn Starmans (m.starmans@erasmusmc.nl). They will judge your proposal, and if accepted, create a project for your study and an account for you on the AFRICAI repository and give you access.

For the proposal, please fill in and send us the following form: https://forms.gle/vLjeChLLRbbDhfPJ8.

2. Uploading imaging data

We highly motivate uploading your imaging data as DICOM, which as standard in the field is easy to work with and provides relevant metadata on the image. Alternative formats such as compressed NifTI are also supported.

Upload DICOM via CTP

The CTP (Clinical Trail Processor) is a DICOM anonymizer, endorsed by the Radiological Society of North America (RSNA), that enables researchers to anonymize imaging data and send it to other imaging archives. CTP has a lot of configurations: we will not focus here on all CTP options, more information can be found on http://mircwiki.rsna.org/index.php?title=CTP-The_RSNA_Clinical_Trial_Processor and http://mircwiki.rsna.org/index.php?title=The_CTP_DICOM_Anonymizer.

We provide a standalone version for CTP, specifically configured to send data to the AFRICAI repository, including an extensive manual at https://gitlab.com/radiology/infrastructure/data-curation-tools/ctp_standalone. Note: important here is that you send the data to your specific project on the Health-RI XNAT, see also the manual on how to do so.

Upload with Python via XNAT REST API

Data, also other than DICOM, can be uploaded using the REST API of XNAT. In order to do this, some expertise in programming and API's is required. The XNAT REST API is documented here: https://wiki.xnat.org/documentation/the-xnat-api.

It is recommended to use the Python client library xnatpy to interact with XNAT through the REST API. The documentation for xnatpy can be found here: https://xnat.readthedocs.io. By using this library the interaction with XNAT can be integrated in a programmatic/automated environment more fluently.

3. Uploading clinical data

Upload as custom variables with Python via XNAT REST API or the XNAT GUI

The neatest way to add your clinical data is through the “custom variables” functionality of XNAT. See for example the WORC database[13], which includes custom variables such as age, sex, dataset, diagnosis. To achieve this, two steps are required:

1. Create a data model for your custom variables. Go to your Project page on XNAT, and click on the “Manage Custom Variables” button in the “Details” tab. Select that you want to add variable(s) to subjects, and click “create a custom variable set”. Give your new set a name and description. The process is quite intuitive: you can give your variables a name, a type, whether they are required or not, and specify possible values.
2. For setting the right values per patient, we recommend two options, although are more possibilities. You can interact with the custom variables in Python through xnatpy (https://xnat.readthedocs.io/en/latest/static/tutorial.html#custom-variables), and thus simply loop through your project to assign the right values. Or, through the XNAT online web interface GUI, you can per patient manually adjust the variables.

Upload as spreadsheet via XNAT GUI

The simplest way to add your clinical data is by uploading a spreadsheet (e.g., Excel, CSV, or any other data format) directly to your project. While this is much easier than adding custom variables, the downside is that users will have to download this spreadsheet and learn to work with it, instead of being able to interact with the clinical data directly through the XNAT REST API. To do so, simply go through your Project page on XNAT, in the “Actions” window select the “Manage Files” option, create a folder, and upload your spreadsheet to it. For standardization purposes, we advise to name the folder “clinical_data”, and your spreadsheet “clinical_data.[extension]”.

4. Make a back-up to Zenodo

Go do your project on the Euro-BioImaging XNAT, make sure you are logged in, and click on the “Download Images” button in the Actions menu on the right side. Under “2: Select Image Data”, make sure you select all the data types you want to include on Zenodo. Under “3: Download Data”, select the option to “Include subject in file paths” and “ Simplify downloaded archive structure”, so the structure of the XNAT data model is maintained in your downloads. If you download as a .zip file, unzip it before proceeding, so you can directly upload the right folder structure to Zenodo. Make sure within your Zenodo submission to refer to your project on the XNAT.

5. Publishing your dataset in a paper

We aim for all papers and datasets accepted for the MICCAI 2024 Open Data session to be published in a special issue, details to be announced later. For more information on this aspect, keep track of the MICCAI 2024 website. To cite your dataset on the AFRICAI repository within your paper, we advise for now to primarily follow the citation from Zenodo, but add the URL to your study on XNAT as well.

[13] https://xnat.bmia.nl/data/projects/worc

4. Costs and sponsoring

Sponsoring for the AFRICAI repository for MICCAI 2024 is provided by the H2020 EuCanImage project (European Union's Horizon 2020 research and innovation programme under grant agreements Nr. 952103). Hence currently, no additional costs for the participants are involved.

5. Limitations and future work

Currently, there are two main limitations of the AFRICA repository, which we are working on improving.

First, while each study will have a unique URL, the repository does not provide persistent identifiers / DOIs. There is currently no standard way of doing this in XNAT in general. Related to this is the lack of proper versioning and associating persistent identifiers with specific versions. The current solution for the AFRICAI repository is; 1) all studies are associated with a published paper that will have a persistent identifier; and 2) publishing the results as back-up on Zenodo will result in a DOI from Zenodo.

Second, there is currently no “cheap” long-term storage solution, the pricing model of the Euro-BioImaging Medical Imaging Archive XNAT is per year. At least until the end of EuCanImage (expected April or Oct 2025) the costs for the repository are covered. This is not a strong limitation, as with our Zenodo back up solution, the data will at least remain after that, and the Open Data chairs will otherwise try to find alternative sponsoring.

Adding persistent identifiers and long-term “cold” storage to the Euro-BioImaging Medical Imaging XNAT archive were already on the roadmap of the Population Imaging Flagship Node Rotterdam and is planned for later this year. We’ve discussed with the team, and they are motivated to have these improvements in place for the AFRICAI repository. Hence we hope to have solved these limitations by MICCAI 2024.