RT-HaND: Summary of Data Lake Construction & Research Data Pipeline for Head and Neck Cancer

September 2024

# Introduction & Purpose

RT-HaND is a large clinically annotated imaging data lake of Head and Neck Cancer patients.

The complete HNC data lake (RT-HaND) is composed of the XNAT data lake (RT-HaND\_I) (containing imaging and radiotherapy DICOM data) plus a set of (currently) 157 tables in EDW (electronic data warehouse) (RT-HaND\_C) which contain over 7 million clinical data points.

RT-HaND\_C comprises of demographic, disease, treatment and outcome data from 2895 patients seen by the GSTT Head and Neck Oncology team between 1/1/2010 and 4/10/2023. Data is up-to-date (e.g. follow-up dates) to the point of 4/10/2023. Originally this data was planned for storage within forms in XNAT, however, initial work demonstrated these are cumbersome to work with and reduce the interoperability of data. Following review of different options EDW was chosen as the preferred method to store non-imaging clinical data. An advantage is that EDW can ingest data and update itself by “hoovering up” spreadsheets in a set location. This enables a clinical fellow/scientist to keep the database up to date without the involvement of CSC or access to EDW. This section of the project has been led by **Haleema Drake** from CSC and **Tom Young** from Clinical Oncology with the data sourced from existing tables in EDW, some unstructured data from Mosaiq obtained using CogStack and manual curation.

XNAT is an open-source software that enables ingestion, storage, anonymisation and export of DICOM files. Within the Radiotherapy Department, “XNAT” has referred, to date and colloquially, to the project enabling construction of a head and neck cancer (HNC) data repository. XNAT is hosted in the Trust by the Clinical Scientific Computing (CSC-XNAT) (**Dika Vilic & Haleema Drake**) and within this infrastructure, a specific project (RT-HaND\_I) contains the HNC unanonymised data warehouse which enables continuous updating of the data. In addition, the Radiotherapy Physics department hosts a separate instance of XNAT (RT-XNAT) designed to uniquely contain anonymised data and cleaned copies of HNC projects **(T Guerrero Urbano)**. This is linked to CSC-XNAT through opened firewalls and ports enabling data to be sent between the two XNAT instances. The role of RT-XNAT is to enable reusability of data, accessibility of project data (e.g. if asked to reproduce published results) and ensure adherence to data protection principles and information governance. In addition, KCL also has an instance of XNAT (**Marc Modat**). This is linked to CSC-XNAT using “XSync” which enables data to be copied across. XSync is an XNAT plugin which enables automatic synchronization of data from a project in one XNAT system to a project in a second system over the internet. XSync is configurable to ensure that only the desired data is delivered, and if required, data is properly de-identified, and that it is delivered on a pre-set schedule. This is the most efficient method of large imaging data transfer between GSTT and KCL.

# Research Pipeline

## Governance

Researchers can request access to anonymised copies of the data stored within the HNC data repository. Ethics approval via the Guy’s Cancer Cohort or individual project ethics approval is required to access the data as well as appropriate information governance and data sharing agreements if applicable. Guy’s Cancer Cohort is a REC approved research framework (Reference: 18/NW/0297). In addition, applicants are required to submit an XRAP\_RTD\_Combined Request Form both to the RT-HaND admin group and the RT Development group. The form asks researchers to specify the data required ensuring principles of data minimisation are followed. Projects will also require approval from both groups which, amongst other things, will assess the resource implications of the project. These forms can be found as 004a\_RT-HaND: Guy’s Cancer Cohort Application Form and 005\_006\_XRAP\_RTD\_Combined Request.

Figure 1 shows the data flow from patient to research data access with anticipated timelines.

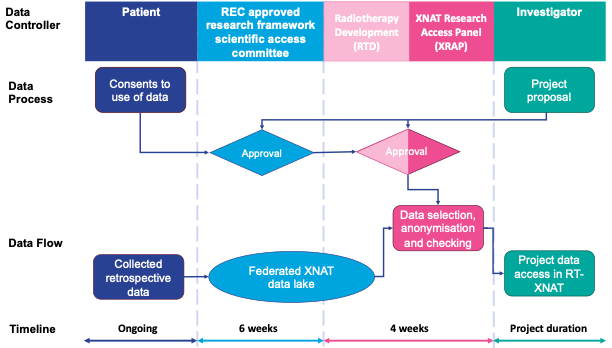


Figure 1: XNAT research access pipeline.

## Practicalities

After approval has been granted, the data will be provided to the researcher. Clinical data will be provided in .csv form, downloaded from EDW (or from a copy of the data in EDW on excel).

The required imaging and radiotherapy data will be downloaded to the XNAT head node by CSC at which point, the RT XNAT administrator would transfer it to be cleaned and anonymised via Python before uploading back to XNAT and applying the XNAT ingestion anonymisation. This process is detailed in 010 RT-HaND\_I SOP for downloading data from the data lake.

This enables the data to then be sent back to XNAT and not be recognised by it as the same data which is key for transferring data onwards (i.e. to KCL via XSync). The RT XNAT administratorwill take responsibility for this. The data transfer pipeline is shown in Figure 2.

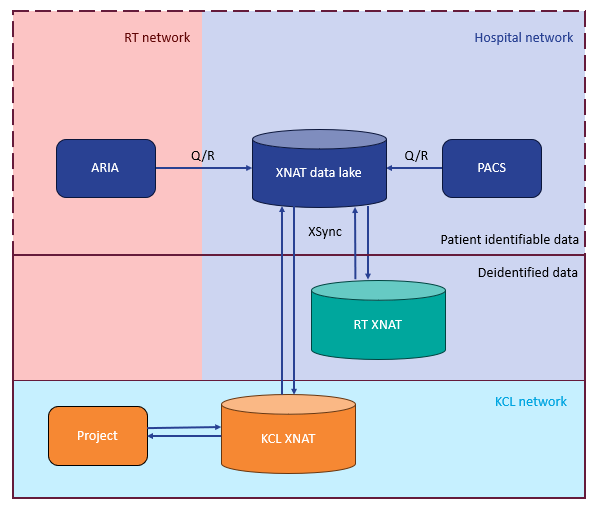


Figure 2: Research data transfer diagram

# Roles and Responsibilities

|  |  |
| --- | --- |
| Role | Responsibility |
| Maintain XNAT server including upgrades and link with KCL | CSC |
| Maintain RTXNAT server including upgrades and link with XNAT | CSC |
| Update RTHNDL data lake (PACS & ARIA) | Refer to PACS & RT documentation for specific responsibilities. |
| Download data for research purposes | CSC |
| Anonymise data for research purposes | XNAT admin |
| Refine data for research purposes | Individual researcher |
| XSync to KCL | CSC |
| Contribute to repository of code transforming and cleaning data | Head and Neck and RT teams; researchers. |
| Maintain back-ups of research data | Individual researcher |

# Glossary

|  |  |
| --- | --- |
| Cogstack | Natural Language Processing AI tool which “reads” electronic health records to structure unstructured data. |
| CSC | Clinical Scientific Computing, under Medical Physics. |
| EDW = Electronic Data Warehouse | A query-able SQL database owned by informatics containing copies of data from historic clinical systems. The Head and Neck clinical data repository has been downloaded from EDW, cleaned and updated in Excel using CogStack and manual curation. The opdated copies are then reuploaded to EDW as a centralised query-able, automatically updatable store. |
| XRAP = XNAT Research Access Panel | A committee composed of members of the Head and Neck, Radiotherapy, Radiology and CSC teams which permits access to anonymised data from the Head and Neck Data Lake repository following appropriate regulatory and governance approvals. As of August 2024, the committee consists of:  Dika Vilic ([Dijana.Vilic@gstt.nhs.uk](mailto:Dijana.Vilic@gstt.nhs.uk)) (CSC)  Haleema Al Jazzaf ([Haleema.AlJazzaf@gstt.nhs.uk](mailto:Haleema.AlJazzaf@gstt.nhs.uk)) (CSC)  Victoria Butterworth ([Victoria.Butterworth@gstt.nhs.uk](mailto:Victoria.Butterworth@gstt.nhs.uk)) (RTPHYS)  Tom Young ([Thomas.Young@gstt.nhs.uk](mailto:Thomas.Young@gstt.nhs.uk)) (RAD ONC)  Teresa Guerrero Urbano ([Teresa.GuerreroUrbamo@gstt.nhs.uk](mailto:Teresa.GuerreroUrbamo@gstt.nhs.uk)) (RAD ONC)  Phil Touska ([Philip.Touska@gstt.nhs.uk](mailto:Philip.Touska@gstt.nhs.uk)) (RADIOLOGY) |
| XSync | An XNAT function which allows 2 XNATs to connect via the internet or network and “mirror” the contents of a project from one XNAT to another. CSC XNAT and KCL XNAT are connected via XSync and the intention is for RTXNAT to also be connected via XSync. |