Provisioning anonymised routinely collected radiology data from the Scottish Population: an extensible big data software architecture

# Abstract

<write this at the end?>

iRDMP

# Background

## Advantages and challenges of using routinely collected clinical images for research

Clinical images, especially when linked to other routinely collected health data, are extremely useful for many types of research: examining early/preclinical diagnosis [1], disease progression [2, 3], genotype-phenotype associations [4], development of risk profiles [5, 6], computer vision methods for biomarker extraction [7, 8], machine learning approaches [9], and discovery and classification of disease types [10]. The emerging field of Radiomics has the potential to bridge the gap between medical imaging and personalised medicine [11]. However, collecting images for specific research projects is expensive and constrains the scale of many studies. Research cohorts are usually comprised of a narrow subset of people with a particular condition, which can make both generalising findings and repurposing of images for other research problematic. Use of routinely collected images, in contrast, opens up the potential for very large-scale studies, which not only efficiently and effectively complement smaller disease-based cohorts of patients but are also extremely flexible when linked to extensive electronic medical records allowing for a wide range of disease areas to be examined. However, whereas research images are typically collected using specific image acquisition protocols under ideal conditions, routinely collected clinical images are much more heterogeneous.

Using clinical images for research and linking them to other routinely collected clinical data is challenging because:

1. Existing software used to query/search for images from PACS (Picture Archive Communication System) are designed for clinical care rather than research. They make it easy to find all images for a particular patient, but they are not designed to facilitate searching for all images with particular characteristics such as slice thickness/scanning protocol/contrast agent/patient medication or linking to other datasets.
2. Reuse of clinical images for research requires de-identification, yet identifiable data can be present in many areas of the associated DICOM (Digital Imaging and Communications in Medicine) file metadata and/or may be present within the pixel data itself and therefore ‘burned on’ to the actual image.
3. Anyonymisation of images can reduce the ability to perform linkage to other datasets e.g. demography,prescribing, hospital admissions etc.
4. Reuse often requires approval from multiple Data Controllers, and the complexity of de-identification increases the risk of rejection of applications for research given the amount of work the Data Controller may have to do to ensure that no identifiable data is released.
5. For machine learning projects, where large numbers of images are required, the image extraction costs for research can be prohibitive.

## Scottish Clinical and Research Data

**Scottish Clinical PACS System:** Scotland has a single National PACS Clinical System which contains all of the radiological images collected from 14 different health boards. This includes 23 million different radiologicalexaminations from a population of 5.4 million and over 1.7 petabytes of data collected since 2010. It includes a range of modalities (including computed tomography (CT), magnetic resonance imaging (MRI), ultrasound, nuclear medicine imaging and plain film radiography). This system is a live environment which is used directly for clinical care.

**Provision of routinely collected text-based clinical data for research:** Scotland has a relatively stable population with long-established use of a unique healthcare identifier (the Community Health Index [CHI] number) that is also increasingly seeded in data in other sectors such as social care. An NHS Scotland service, called the electronic Data Research and Innovation Service (eDRIS) [12], provides a National Safe Haven environment (hosted by Edinburgh University) to support research access to anonymised extracts of linked data from different sources to answer specific research questions. The linkable phenotypic data includes a range of national datasets including, for example, prescribing, death data and hospital admissions.

**Incorporating clinical imaging data into the wealth of available datasets for research:** A project has been underway to obtain a research copy of the data held within the Scottish National Clinical PACS system to enable the clinical imaging data to be linked with the other routinely collected datasets and be made accessible for research (given appropriate data governance approvals). The research copy of the Clinical PACS system is called the Scottish Medical Imaging (**SMI**) Database and the data is held in the non-proprietary DICOM format.

The management of imaging data for research presents a substantial set of challenges beyond those encountered in the management of purely text-based records. Some of these are variations on familiar challenges, such as de-identification, whilst others are novel and intrinsic to this type of dataset, such as size and compute requirements for big data processing.

This paper describes the architectural solution and software platform developed to support the management of hosting, extracting and linking the SMI data which addresses the challenges identified of using routinely collected imaging data for research listed above.

## Project Approach

There were 4 phases to the project to date:

**Requirements gathering:** An initial requirements gathering exercise was undertaken at the project inception elucidating requirements from the research community who will use the data extracts provided by the platform, the National Health Service Data Governance representatives as the Data Controllers of the data and the National Safe Haven staff who will use the platform to build cohorts and provision relevant data extracts to researchers for analysis. We also looked into other open source and freely available platforms for hosting and/or anonymising imaging data to see if any of these could be used entirely or in part within our solution. We investigated both functional and non-functional requirements of the solution.

**Development of the Architecture:** We developed a range of option appraises and designed an architectural solution to meet the requirements.

**Development of Prototype and MPV Software:** We developed prototype software to run on a Regional Safe Haven environment managed by the University of Dundee, whilst the SMI data transfer project was taking place in parallel. This prototype supported 2 consented research projects which were predicting dementia from CT and MRI images. We then expanded the prototype and developed a Minimum Viable Product (MVP) to run on the National Safe Haven. In general terms the MVP is “a product with just enough features to satisfy early customers, and to provide feedback for future product development”[13].

**Testing and Exemplar Project on Sample Data:** The MVP software was then tested on 10 test cases and 1 full end to end exemplar project using a subset off all of the data: ~3 million cases with an estimated total size of 180TB. This is all of the images generated across Scotland in February for a 7 year period. The full set of historical data is still in the process of being transferred and could therefore not be used for complete testing at this stage.

# Architectural Solution

We first describe at a very high level a summary of the requirements, then our architectural solution and explanation of why this solution met the requirements and why our solution is different to that of other open-source solutions for the large scale hosting of imaging data.

## Summary of requirements

**Main Requirement:**

1. To provide a secure method for hosting >2 Petabytes of identifiable imaging data and provision de-identified subsets of this data, linked to other datasets, for specific cohorts within a virtual Safe Haven Environment for researchers to analyse but not remove the data.

**Data Governance Requirements:**

1. To adhere to the Scottish Safe Haven Charter [14] for the use of unconsented data for research and work within the existing National Safe Haven architecture.
2. To satisfy data governance requirements so that there is clear separation of roles for the users of the platform and only the minimum amount of data can be seen to fulfil each role. Therefore, researchers can only see a de-identified subset of data which is required to answer their specific research question. Researchers cannot build cohorts directly themselves from the raw underpinning data. The eDRIS team of Research Co-ordinators and Data Analysts (termed Research Co-coordinators throughout this document) who provision data extracts for the researchers can only see a de-identified version of the metadata about the images in order to fulfil the role of cohort building. Only individuals who are maintaining the infrastructure or fulfilling the role of de-identification analyst can view identifiable data and only when carrying out specific tasks that require them to view identifiable data.
3. To protect identifiable data from unauthorised access.

**Cohort Building Requirements:**

1. To support the National Safe Haven Research Co-ordinators to build cohorts based on data from different sources such as:

* Image metadata (DICOM tag data), e.g. select MRI images of the head
* Image pixel data, e.g. select lung scans images where the airways are less than 3mm using an algorithm which extracts features from pixel data.
* Other health data sets not held in the image store (such as prescribing data), e.g. select all image where a patient has been given a particular drug within 3 years prior to the scan date.
* Other non-health related data, e.g. select images where the patient lived in a care home at the time of the scan
* Structured reports, e.g. where ….
* Metadata about an image captured as part of the research output of a project which used the environment. There are several instances where a research project which uses the environment may curate or add value to images through their expertise. An example might be where a project requires chest CT scans and funds a radiologist to view the images and generate a gold standard curated set of 5000 images by recording whether or not the image shows evidence of coronary artery calcium. If the research group who creates the gold standard data wishes, the system should be able to record such information and use it to build further cohorts for other research projects e.g. select images where a radiologist has recorded the image as showing signs of coronary artery calcium and controls.

1. To provide technological solutions which the Research Co-ordinators are familiar with so that the same skill set can be employed e.g. use of structured SQL databases rather than un-structured databases and no programming expertise
2. To provision summary, curated data for cohort building rather than requiring the Research Co-ordinators to require domain knowledge of the DICOM standard and the intricacies of the alternative use of the standard by different vendors, health boards and users.
3. To provide the data to the Research Co-ordinators in a form which is easily linkable to other datasets for efficient cohort building.

**Data Requirements:**

1. To preserve original data in the state it is provided. This requirement must be balanced against the data security and information governance requirements which arise from holding unconsented patient data. The goal of only discarding operationally non-critical data helps ensure that nothing is thrown away which later turns out to be important, as it will be very difficult (if not impossible) to re-download missing data from the source system.
2. To minimise the number of copies of the data where possible. This is both for data governance reasons and because of the large volumes of data i.e. cost of storage and challenge of maintenance.

**System consistency:**

1. To ensure all procedures are traceable and reproducible through auditing and atomicity of operations. For example, information about the what, how or when a system user interacted with the system is stored and the interaction can be replicated.
2. To maximise data integrity by ensuring no operation whatsoever can damage production data or leave a production dataset in an indeterminate state.
3. To modularise the software components to mitigate point of failure risks, maximize reusability and dissemination of implementations.

**Efficiency and Maintainability:**

1. To reuse as many applicable, open source or freely available tools as possible i.e. do not try to re-invent the wheel.
2. To be cost effective to run and for data to be securely available for research projects in a timely fashion. A pragmatic consequence of this high level requirement is that the system should not try to de-identify all 1.5 Petabytes of data prior to the platform being utilised for research, rather it should support reactive de-identification based upon the image types required for research projects which us the system. The DIOCM standard has 5000 different metadata tags and unspecified numbers of additional private tags. The format supports hierarchical tags (sequences) which can include a mixture of sensitive and insenstive data. The DICOM standard is used differently by different vendors, for different modalities and by different health boards resulting in highly heterogeneous data. Although there are many software programs which claim to de-identify imaging data there is a risk that such programs do not do this for all variations of the DICOM data used from such a diverse set of real world, routinely collected imaging data. Any programme of work to develop de-identification protocols for individual machines covering the entirety of the dataset would take an infeasible amount of time. It may also be the case that a significant proportion of the imaging might never be requested for extraction, rendering unnecessary any work done to create related de-identification protocols.
3. To be vendor agnostic and open source .
4. To be able to support modular enhancements over time from other sources. There is vast expertise and other software tools which could improve the platform for the whole community. This needs to be balanced with the IG considerations and the stability of a production system.

**Research Use:**

1. To provide tools within the Safe Haven Analytical Platform which can view and manipulate images.
2. To be able to develop software within the Safe Haven Analytical Platform and a secure method for extracting the code from the environment (without being able to extract the image or image metadata).
3. For the access to the data within the Safe Haven Analytical Platform to be high performance to support deep learning and machine learning.
4. The de-identification process should not remove metadata which is required by software to view and manipulate images.
5. Different research projects may require different de-identification protocols depending on the research question they are asking. E.g. ???

## Architecture – Logical Overview

The high-level platform architecture is shown in Figure 1. Software processes are shown as blue boxes, data stores as orange cylinders, the metadata catalogue as a purple cylinder and access for different roles are shown as people. The SMI Data Repository includes the identifiable zone and a de-identified zone. The SMI Analytic Platform is the Safe Haven environment where researchers can access their relevant data extracts.

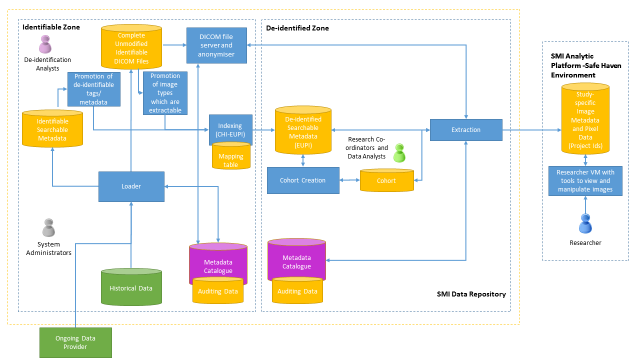
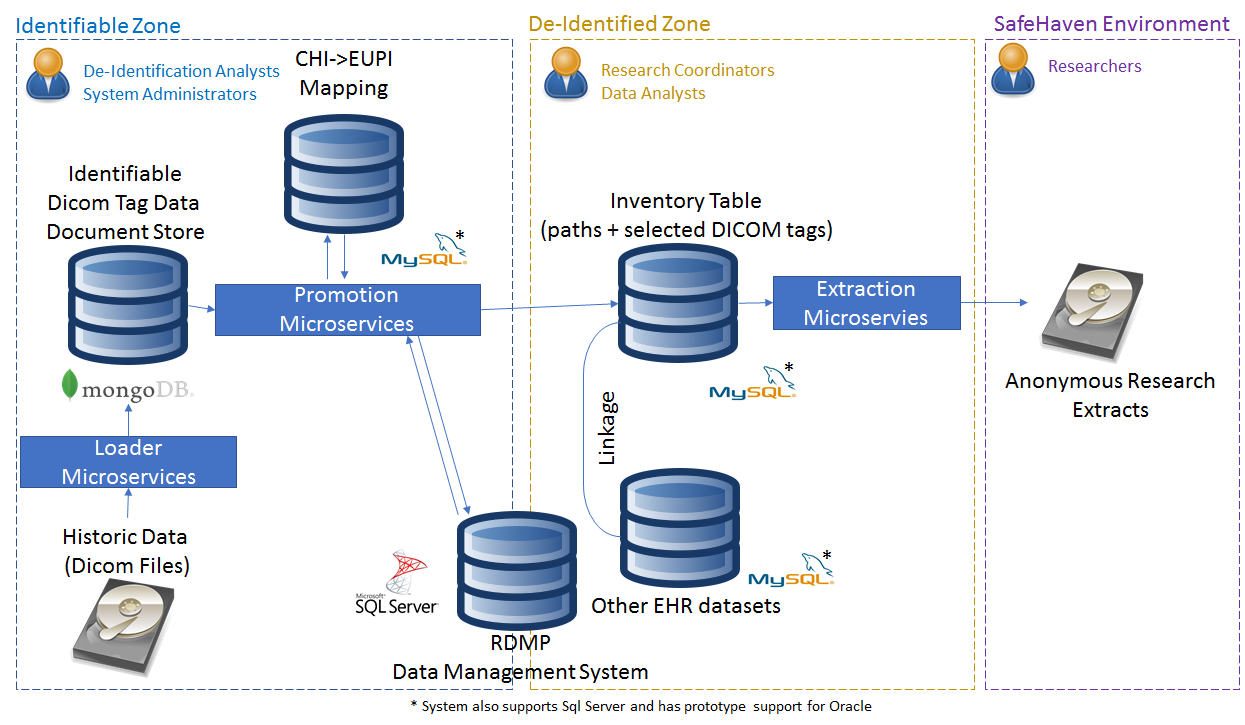


Figure 1 - Overview of Architecture



Each different zone has audited, controlled access with a clear separation of roles and functions. Only System Administrators are able to access the identifiable zone to carry out maintenance and security functions without being able to view the raw data. Only De-identification Analysts are able to view the potentially identifiable data in their duties. Research Coordinators are able to query de-identified text based data for cohort building, linkage and extraction (using EUPI – Encrypted Universal Patient Identifier) in the de-identifiable zone. Researchers are able to access de-identified (using Project-IDs) metadata and de-identified pixel data for their particular cohort within the Safe Haven VM environment.

## High level Summary

Within the identifiable zone, imaging data is pulled from the data stores by the **Loader** process and stored within the **Complete Unmodified Identifiable DICOM Files** data store and the metadata (DICOM tags) copied from these files and stored in the **Identifiable** **Searchable Metadata** data store (a Mongo database).

The **Identifiable** **Searchable Metadata** is analysed for potentially identifiable data. Metadata which can be de-identified and is useful for cohort building is sent to the **De-identified Searchable Metadata** (a SQL database) and stored using EUPIs within the De-identifiable Zone.

The De-identified Searchable Metadata database is queryable like any other database by Research Co-coordinators to construct cohorts using their existing working procedures. The extraction process calls the DICOM file server and de-identifier which will de-identify all of the pixel and non-pixel data in the DICOM files used for the cohort and copy these files into the Researcher Safe Haven environment.

A summary of each of the data stores and process is provided in Tables 1 and 2.

|  |  |
| --- | --- |
| Data Store | Description |
| Complete Unmodified Identifiable DICOM Files | Non-proprietary standard DICOM files are stored unaltered in a file archive. There are many reasons why we wish to keep the identifiable data and store the original DICOM images: Some DICOM anonymisation tools have been known to output DICOM files that other tools are unable to process. As different tools often have slightly different interpretations of the DICOM standard we were reluctant to alter the original files and risk introducing such corruptions over the whole data collection.It is conceivable that future data de-identification strategies will wish to make use of some identifiable data and removing that data would therefore limit future options.If a program is developed to strip all of the identifiable data from the DICOM files and tags there is the risk of rendering the whole dataset unusable if this is done incorrectly, linkage to other datasets therefore will be either incorrect or impossible.The NHS may wish to use the data as a secondary offline DR system or use the data to populate a clinical system from an alternative provider. In which case it needs to be technically feasible to generate the data in identifiable form in a format that is non-proprietary and as close to the DICOM files as when they were originally captured. |
| Identifiable Searchable Metadata | All tag metadata from the DICOM files is extracted to a Mongo database in a searchable format. Not all the file metadata from this store will be copied into ‘De-identified Searchable Metadata’ because its quality and identifiability risks are unknown.  This data is stored in an identifiable format because De-identification Analysts need to know what the identifiable data is so that they can remove it *e.g.*   * If the patient name is Mrs Jones then if searching for identifiable data in the clinical report the De-identification Analyst will need to know to look for the text “Jones” in order to remove this data. * Or if checking if an image is identifiable they might need to know the CHI number in order to check this is not burnt into the pixel data. |
| De-identified Searchable Metadata | This is a SQL database which contains image metadata which is suitably cleansed and de-identified, i.e. has been confirmed to be well-populated, of high quality and does not contain identifiable data. This is used by the Research Co-ordinators for cohort creation and extraction to the Safe Haven. The data is indexed using EUPIs.  A metadata field may not be simply a copy of data from a single DICOM tag. It may be transformed and curated data. For example, some DICOM vendors store “h” whereas others stored “head” to mean a head scan. The cleaned and homogenised metadata may contain only “head”. Other metadata fields may be a single summary value which summarises data stored in multiple different DICOM tags. For example, ??? |
| Cohort | This is the list of image IDs and metadata columns required for the Research Project. It may also contain data linked from other datasets (via the EUPI). |
| Audit | This database contains all of the audit information from the different processes. |
| Study-specific Image Metadata and Pixel Data | This is the data (pixel and non-pixel) required for a specific research project indexed by project identifiers. |
| CHI to EUPI Mapping Table | Scotland uses the CHI unique identifier for health data. Adhering to the Guiding Principles of data linkage for research [15] the National Safe Haven separates out the roles of indexer and linker. Research Co-ordinators link data from a range of sources provided to them with the CHI replaced by the EUPI identifier. The imaging data also follows methodology. The mapping table is securely held and the data only accessible via the automated conversion process. An individual can be given multiple CHIs if they access healthcare in different regions and it takes time to resolve therefore the mapping table is updated monthly. |
| Metadata Catalogue | This purple block is both a process and a data store. It is the co-ordinating process of the components of the architecture, keeping data and processes in sync and storing audits and configurations. |

Table 1 - Overview of Data stores

|  |  |
| --- | --- |
| Processes | Description |
| Loader Process | A set of configurable components which manage the load of a set of DICOM files from retrieval to storage. |
| Promotion of de-identifiable tags/metadata | This process pushes metadata from the Identifiable Searchable Metadata (once it has been fully checked and marked for promotion i.e. white listed) to De-identified Searchable Metadata.  Only a small subset of metadata has been initially included in the de-identified dataset. It is not feasible or desirable to proactively analyse the complete Identifiable Searchable Metadata in order to promote all tags. This is in part due to the difficulty in wholly determining that a tag of a certain type does contain identifiable information for a) the whole of the current archive and b) future PACS images that will be taken.  A tag can be promoted on two conditions:  1) it is determined to not contain identifiable information,  2) or the identifiable information it does contain can be de-identified.  Dicom tags were promoted based on anaylsis by a domain expert and the initial schema was created. [Ref ] . Initial effort focused on CT as it represts 30% of the historical data available.  Sophisticated techniques such as Natural Language Processing methodologies are used to determine condition 1 or find a solution for de-identification for condition 2. The solution for condition 2 is known as the anonymisation profile and is saved in the Metadata Catalogue. Once a tag can be holistically flagged as safe for promotion it is moved to the De-identified Searchable Metadata data store.  Once the highly likely used metadata has been promoted to the De-identified Searchable Metadata database (such as modality, scan location, contrast agent etc.), the Identifiable Searchable Metadata database is more likely to be queried reactively as a result of specific researcher requirements encountered when attempting to identify a cohort for a particular study rather than trying to de-identify it wholesale on a first pass. |
| Promotion of image types which are extractable | This process white-lists images which are extractable in the sense that pixel data can completely be de-identified. Some images, particularly ultrasounds, may have identifiable information such as patient name or CHI watermarked on the image. A solution to redact the identifiable information from the image is developed here and the resulting anonymisation profile (the technique to anonymise the image) is subsequently saved to the Metadata Catalogue. The anonymisation profile will be used later when extracting images. |
| Indexing (CHI-EUPI) | This process is called when metadata is promoted to the de-identifiable zone to replace identifiable CHIs with EUPI. It is an automated process so that no individuals can see this mapping. |
| Cohort Creation Process | A set of software tools (or manual SQL queries if the user prefers) which query the DICOM metadata within De-identified Searchable Metadata in order to select images relevant for a particular cohort (by applying filters which describe researcher requirements). The resulting cohort forms the basis for both the initial and subsequent releases of data to the Safe Haven for the relevant study, and as such it is critically important that the cohort is identified and managed correctly. |
| Extraction Process | This process uses the data within the Metadata Catalogue, the Cohort database and De-identified Searchable Metadata to determine which files to extract for a particular research project. It calls the DICOM file and server anonymiser to de-identify the relevant files used to build the cohort for release to the researcher. After the cohort output and the de-identified DICOM files are curated, the process triggers a release into the Researcher Safe Haven environment. |
| DICOM file server and anonymiser | The DICOM file server and anonymiser:   * Obtains the file(s) from the file archives * Anonymises the pixel data of the file if need be * Anonymises the metadata in the file (leaving only the whitelisted tags) * Converts the file to an alternative format if required * Returns the final file(s) to the user * Checks that the researcher has permission to access the requested file |
| Researcher VM with tools to view and manipulate images | There are 2 main use cases: small scale studies where a researcher team may wish to open and mark up each image by eye and large scale studies where software and algorithms will be developed by the users of the system to analyse the images for their specific project. The different tools available within the Safe Haven meet both sets of requirements. The researcher VM image includes a standard set of tools (such as MicroDICOM (simple DICOM viewer), ClearCanvas (open-source PACS client, cf. Carestream), XNAT), and the capability for users to add their own. The VM provides access to the associated data from study-specific image metadata and pixel data but does not allow row level or pixel data to be extracted. Access to the internet is restricted when analysing the data. |
| Metadata Catalogue | This purple block is both a process and a data store. It is the co-ordinating process of the components of the architecture, keeping data and processes in sync and storing audits and configurations. |

Table 2 - Overview of Processes

|  |  |
| --- | --- |
| Role | Description |
| Researchers | Carry out the research on a dataset extracted from the SMI DB and other linked data. Any project may have a variety of researchers including clinicians, statisticians, radiographers, image analysis and machine learning experts etc. They view and work on the PACS images within a Safe Haven environment. |
| Research coordinators/cohort builders | Work with the Researchers to produce the data extract that allows the research study to be carried out. Research coordinators understand where the data is stored, how to link across datasets and will run software, write scripts, query databases etc. in order to produce the final cohort datasets. |
| Data analysts | Work with de-identified PACS data to produce more usable versions of the data for research coordinators to work with. Over time data analysts (working with domain experts) may produce additional mapping tables and categorisation systems that make it easier for Researchers and Research Co-ordinators to work with the data. |
| De-identification analysts | Are responsible for ensuring as much data as possible is made available to research coordinators for the creation of cohorts but that no identifiable data reaches the coordinators. Much of the de-identification task is automated but the system needs to be continually monitored and new DICOM tags added to the whitelist (or blacklist) as required. |
| System Administrators | Are part of the infrastructure team and are responsible for building and maintaining the underpinning infrastructure, security, network separation, monitoring and supporting automated processes. Supported automated processes would involve checking for example, whether there were errors in the data load process or data extraction process. They have privileges and expertise to debug and/or restart these processes. |
| Software developers | Produce any new software required within the Safe Haven. The software is developed and tested outside of the production environment. Deployment of software updates will be carried out by System Administrators. |

*Table 3: Roles*

## Software Architecture

The platform has been built upon the open source Research Data Management Platform (**RDMP**)[16]. The RDMP stores, manages, cleans, de-identifies and processes data to create reproducible, auditable data extracts for research and in the last 4 years has been used to support over 500 projects, generating over 1300 data extracts of mainly phenotypic text-based data for epidemiological research projects and clinical trials. The RDMP already provides many of the core components such as auditing, logging, and anonymisation required for populating the relational database in a platform agnostic way; and linkage and extraction; ~~workflow ??? required for cohort building using text data~~ therefore it was efficient to build upon this platform to also handle imaging data.

The system has been designed with a plugin architecture. All the operation and tasks in the RDMP platform are mainly based on pipelines of components. These components are freely pluggable within a pipeline and reusable across pipelines. External developers can extend the functionality of RDMP by writing components which can be loaded by the RDMP platform at start-up. Description

# System Analysis

## Current status

To date ~3 million cases with an estimated total size of 180TB have been transferred, and the rest of the historical images will be transferred into the EPCC environment by the end of 2018. Going forward, a technical solution and agreement is in place for future images to be transferred as they are collected from 2019.

## Comparisons to other existing systems

Given the different imaging platforms in active development to support research projects, we investigated alternative platforms so that we did not re-invent the wheel. In general terms, other solutions have concentrated on consented cohorts from researcher collected research images rather than much larger unconsented data from routinely collected “real world” images. The architectural solution developed by others is generally a large anonymised database (sometimes distributed) containing all of the images with permissions to see, extract and run pipelines on the imaging data configured for each research group. The metadata provided is limited and relatively clean in comparison to routinely collected data. The architectural challenges and solutions are therefore very different. For example, a key functionality of the platform is the efficient and effective selection of anonymised cohorts from noisy, heterogeneous, petabytes of identifiable data.

If we were to have created an architecture where we stored a de-identifiable, clean, homogenised copy of all of the pixel and metadata within the de-identifiable zone we could have employed one of the many excellent open source platforms for managing large volumes of imaging data such as OMERO[19, 23] or XNAT[22, 24]. There are several reasons why we did not choose to create a completely de-identifiable copy of all of the data (pixel and metadata) within the de-identifiable zone and therefore did not use such platforms to manage the core data repository:

* We envisage that the methods to de-identify data will change over time as our understanding increases and technological solutions improve. It is impractical to re-create a >1 petabyte of de-identified images each time our methodology improves.
* It is unnecessary to undertake the effort to validate any de-identification method on all 5000 standard DICOM tags and a large number of vendor specific private tags when only a small fraction of these will be required by research teams. It is unknown which ones will be required upfront.
* A proportion of all of the images will never be extracted/released for research projects as they will not meet the cohort requirements. Only de-identifying imaging data re-actively when it is required for a specific project removes the requirement to carry out a needless timely and computationally expensive de-identification process on images which are never required. (The flip side of this is that a particular image may be de-identified several times, once for each project. This is an issue which we plan to resolve in the next stage of development - as discussed further in the discussion section).
* It is risky to test a specific de-identification tool on sample data and trust that it will therefore also be successful for variations of routinely collected data from multiple sources and vendors. The architecture was designed to reduce this risk: by default blacklisting all data until proven otherwise, in which case the metadata and/or image time is then “promoted” to a white list.
* The data is currently 1.6 Petabtyes and expected to grow at ~400TB per year. There is significant cost of maintaining 2 copies of the data both in terms of hardware but also the maintenance required to update a duplicate as new data arrives.
* Hosting duplicate versions of the data introduces additional data security and governance risks.
* Different research projects require different de-identification – so one size fits all does not fit all.
* Following the data protection principle that individuals should see the minimum data to fulfil their job role, there is no need for Research Co-ordinators to see the pixel level data to build cohorts therefore, only text based metadata is provided for cohort building.

Although the solutions developed by other groups will not fully meet the requirements of this programme, one of our core principles is to reuse as many applicable, open source or freely available tools as possible i.e. do not try to re-invent the wheel. Therefore, where relevant we have included other software within our architecture (see section below).

## Justification of different tools within the plugin architecture

The modular services/messaging/plugin architecture allows independence over the choice of underlying tools. This means that anyone of the components or tools employed within the architecture can be reasonably easily swapped out for another solution.

### Choice of Micro Services

A Microservice architecture can be defined as sub-dividing a single application into a set of loosely coupled elements with limited responsibility and communicating with each other. Advantages of a Microservice architecture as opposed to a monolithic approach have been known in the IT industry in recent years (see examples at Amazon, Netflix, LinkedIn, etc.).

### For the communication mechanism between the microservices Rabbit MQ was selected as a robust open source queue messaging system. Non structured database solution for identifiable metadata

The data in the DICOM tags is largely unstructured and deeply hierarchical. Moreover, its structure may change overtime (e.g. new tags). As a consequence the use of a document oriented, flexible and dynamic data storage was deemed necessary and mongoDB, being open source and widely used.

Load time?

Scalability

### Structured database solution for de-identifiable metadata

Why SQL Server? The main interaction with the de-identifiable metadata catalogue comes from the RDMP tool, both when loading the DICOM whitelisted tags into the de-identifiable store and when providing the filtering query for the extraction process. The RDMP database platform of choice is MS SQL Server.

There are many different software programs which de-identify imaging data. We tested the feasibility of 3 different widely used programs (DICOM Confidential [26], XNAT [22, 24], CTP [21]) in deciding which to adopt as part of the pipeline. A summary of each of the programs is provided in Appendix B.

For a meaningful comparison of the tools, a set of criteria were devised and each de-identification program was examined in turn against these criteria using a rating of 1-5 (where 5 is the best). We grouped the results into 3 different categories: core functionality, user friendliness and support. Table 4 shows a summary of the scores for each category with the detailed analysis provided in supplementary material A.

In summary, DICOM Confidential was ruled out due the quality of the documentation and the lack of a large community supporting it. We found that some of the images produced by DICOM confidential were corrupted and chose not to pursue the matter to too much detail as the functionality of the other 2 tools also appeared superior.

There was little difference in the functionality of CTP and XNAT. They are both well-supported tools which were found to be able to perform the tasks required. The overall score of CTP was higher than XNAT. We thought that the XNAT image “bundling” for applying rules to subsets of images would be a useful which CTP does not provide. The pixel level anonymisation capability appeared to be much better supported and straightforward in CTP, and this is very important to the solution. As such, we chose CTP.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Core Functionality  (max possible = 45) | User friendliness  (max possible = 30) | Support  (max possible = 25) | Total |
| XNAT | 37 | 21 | 22 | 80 |
| CTP | 41 | 24 | 25 | 90 |
| DICOM Confidential | 35 | 24 | 14 | 73 |

Table 4 – Score of different de-identification tools

### NIFTI storage as a method of de-identification

NIFTI (Neuroimaging Informatics Technology Initiative) is an alternative to DICOM as a file format to store medical images. Originally created for neuroimaging, NIFTI stores image data as a single 3D image (.nii file), whereas DICOM stores a separate image file for each slice of the scan. In addition, the NIFTI format only stores pixel data and metadata related to the image itself, not any patient or study information as you would find in a DICOM image. This makes NIFTI an attractive method to “anonymise” DICOM images. Not all image modalities and compressions are supported however, and conversion tools require extensions to interpret the private tags that some image scanners write into the DICOM files to describe the pixel data.

NIFTI has become popular in some machine learning applications and is preferred over DICOM due to the ease of dealing with only 1 file representing the whole 3D scan.

Emily to fill in more

### De-identification of metadata within De-identifiable SQL DB

What did we do here?

### Software deployed in the Safe Haven Analytical Environment

Deployed software into the Safe Haven for viewing and manual annotation of these images. This software includes: MicroDICOM (simple DICOM viewer), ClearCanvas (open-source PACS client, cf. Carestream), ITK-SNAP (annotation/visualisation) and the VAMPIRE team's retinal image annotation tools.

OMERO[19, 23], XNAT[22, 24], VAMPIRE[25].

## Functionality testing

1 – test the data ingest, largely work with the EPCC team.

2 – test the progress we have made with the algorithm plugin (EPCC or HIC support required)

3 – test against the requirements from last November – everyone

4 – test against the “test case projects”

5 – finally test with Lungsolve exemplar

## Test Cases

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Example | Challenges | Current Capability | Future Work |
| Derive cohort from complex DICOM metadata | Extract all the T2 (the timing of radiofrequency pulse sequences which highlight fat and water) MRI brain scans which are in transverse orientation, where they are susceptibility weighted imaging (SWI) sequences and are not a calibration sequence. | Metadata about each image is stored in DICOM tags. There are 5000 standard DICOM tags and there can be any number of private DICOM tags, used by machine manufacturers to store additional data above that of the standard tags. Each different health board, manufacturer and machine uses the DICOM tag standard in slightly different ways e.g. the text in the body type DICOM tag could be “H” or “Head” for a head image, depending on the way the data was captured. The same information can also be found within different DICOM tags depending on the source e.g. identifying the image as a SWI sequence requires checking three different fields for the occurrence of one of four possible strings and then filtering out some specific mismatches. This leads to problems of standardisation, metadata and definition of data dictionaries. In Scotland there are 4 different Radiology Information System which hold data in different ways. There are additional complexities due to conflicting requirements for standardisation for the purposes of cohort building and research use. |  |  |
| Derive cohort from structured reports | Structured reports are summary information mainly stored in free text format which have been populated by a clinician about the study. The structured report is stored in a DICOM file related to a series of images. They include patient information such as why the scan was requested in the first place. A cohort derived from structured reports might seek to extract all the images where a CT scan was performed because a lung tumour was expected. | There are 2 main challenges with the free text format of structured reports: ensuring anonymisation and inferring the correct meaning. |  |  |
| Derive cohort from image pixel data | Extract all of the x-ray images of the knee where the depth of cartilage is less than 2mm. This information is not captured in the DICOM metadata and instead would be obtained using an image processing algorithm. This requires an automation process where potentially relevant images are opened in the identifiable zone and the algorithm applied to the pixel data returning the cartilage depth. | Applying automatic medical image analysis to large volumes of data, especially 3D like MRI, can give issues of scale and system complexity especially when reproducibility for research is required (i.e. without re-running computationally expensive processes). |  |  |
| Derive cohort mainly from other data sources and simple DICOM metadata | Extract all of the CT and MRI brain images (basic DICOM tag metadata) where the patient is between 50 and 75 years old, has been prescribed a Dementia drug in the 5 years prior to the image capture date (information obtained from National Prescribing Records) and where the patient is not deceased (information obtained from the Death Registry) | The data required to link to imaging data can be from a range of different sources. There are many manual processes for extracting and linking the data which are time consuming, inconsistent and error prone, and not scalable. |  |  |
| Share image files across research projects |  | Simply copying pixel data for each research project may not scale for imaging data, where storage could quickly become infeasible as the SH hosts ever greater numbers of studies each requiring large imaging datasets. An efficient method of sharing the pixel data between multiple studies may be required. However, each study will have different metadata, e.g. study-specific patient identifiers in the image header, so a solution which combines shared pixel data with study-specific non-pixel metadata is needed. |  | We will investigate different solutions such as a Virtual File Server (already developed in prototype), requiring each research group to purchase more disk space should their project require it, pulling images in batches/caching or another technical solution entirely. Different strategies for serving images may be required, such as a file share for machine learning consumption but a DICOM server when using a DICOM image viewer. The output will be a classification of processing patterns and optimal strategies to address these requirements. |
| **Anonymise images and metadata for viewing by researchers and cohort builders** | DICOM files can potentially contain identifiable data in any tag or burnt into the pixel data itself e.g. the CHI number or patient’s name and postcode. | For large numbers of images it isn’t feasible to manually check every single image and metadata field for identifiable data to ensure that an automated anonymisation program has successfully removed all potentially identifiable data. This can be problematic when metadata is stored in different ways by different manufacturers, machines and health boards as the anonymisation needs to be validated for each storage method. | Following analysis of different tools during our MVP development we will utilise an existing tool (CTP) for anonymisation of the data for researchers. We will develop methods for recording which fields and imaging types are “extractable” and a pipeline to automate the execution of CTP. We will develop efficient algorithms to anonymise subsets of metadata from the MongoDB into the anonymised metadata DB. We will optimise and parallelise the anonymisation process where possible. |  |
| **DICOM viewers within the SH to enable manual feature extraction** | Imaging data requires a wider range of applications in the SH than the smaller set of analysis programmes provided for text-based records, with specific applications varying based on researcher requirements e.g. to measure the width of a stenotic artery. There are many available software tools such as such as XNAT[22], ClearCavas[27], dcm4che[28], MicroDicom[29], OMERO[19] | We will install and configure tools for image viewing, manipulation and feature extraction for researcher use in the SH. |  |  |
| **SH Environments allowing programming and HPC** | Researchers doing machine learning, deep learning, computer vision and algorithm development the SH need suitable hardware, such as GPUs, the ability to develop software programs within the SH and to install their own software. |  |  |  |

# Discussion

Although the platform has been designed to work within Scottish data governance constraints and used by the eDRIS team, it could be used in other environments.

This is not a tool for managing and viewing images for research like systems such as XNAT, OMERO, MicroDICOM (simple DICOM viewer), ClearCanvas (open-source PACS client), ITK-SNAP (annotation/visualisation). This is a platform and pipeline for extracting images from a directory of images based upon cohort selection criteria, anonymising them and copying the images into a secure location for analysis.

Theoretically a tool/system for managing and viewing images from a single data store could have been configured/enhanced to include a permissions layer on the top of the data source so that they only had access to the images they were allowed to view and these permissions could be set up by the Safe Haven team. This model was discounted for several reasons:

* Risk of hacking
* Risk of de-identification going wrong
* Speed of access
* Researchers wishing to use their favourite tools to manage and manipulate imaging data.
* Cohort building functionality not available within such tools.
* Speed of de-identification on the fly – or needing 2 copies of all of the data.

The introduction section of this paper identified 4 challenges to using routinely collected clinical images for research. To address challenge 1, the platform uses a Data Controller approved, standardised de-identification workflow. To mitigate the risks to the Data Controllers of providing unconsented routinely collected images for research (thus addressing challenge 2) the platform works in accordance with the Scottish Government guiding principles for secure linking, anonymising and analysing data sets for research, where a subset of data for a specific cohort are linked for an approved research project and access provided via a Safe Haven (**SH**) environment (the NHS Scotland term for a Trusted Research Environment)[15]. Access to the data can be revoked by the Data Controller at any time and researchers cannot extract/output any information other than aggregate level results from the environment. There is a separation of the roles of indexer, linker (carried out by a trusted third party or Safe Haven staff) and researcher. The platform stores the SMI data in the standard DICOM format (rather than a proprietary format) and provides the capability to search for and build research cohorts reducing the cost of providing images for research projects and also providing a mechanism to return relevant images from different data sources (addressing challenges 3 and 4).

Applicability/potential of the architecture and platform to be utilised in other environments/use cases: There are many different platforms in active development to support multiple research projects using clinical imaging data. The architecture have not just been designed to fulfil Scottish data governance principles and data structures, there is a much wider applicability. There are many other Safe Havens nationally and internationally [17, 18] where such a solution might be applicable and a there is a trend towards the creation of new Safe Havens. Data extracts produced by such an automated platform do not necessarily have to be accessed within a researcher analytic platform.

## Still to do

### Image provisioning with limited storage

The size of the dataset is within one order of magnitude of the size of the total available storage. Storage capacity will become problematic when 10s or 100s of research projects are running concurrently, with each requiring a significant volume of imaging data. Simply copying the imaging data to a research project’s file area could result in total system storage requirements in the petabyte range.

### Virtual File Server

The Virtual File Server component is a solution for provisioning project-specific versions of images without making project-specific copies of the pixel data for each image, substantially reducing storage requirements. Project-specific DICOM metadata is stored for each project, but the pixel data is retrieved from the ‘Partially de-identified DICOM files’ store; therefore, there is only ever one copy of the pixel data within the data management system (the researchers could create copies in their own file area, but would be subject to disk quotas in their area).

## Feedback and enhancement improvement of system from other sources



# Conclusions

# Appendix A – Summary of the hardware infrastructure

The environment has 2PB of useable high performance, replicated storage platform that is closely linked to a computation cloud/High Performance Computing (HPC) environment to enable high throughput analysis of clinical imaging. The infrastructure uses building blocks of ultra-dense commodity disk storage systems connected to storage servers and linked to interface nodes to provide access to the storage blocks by 10 GB Ethernet. File system metadata and image metadata will be stored on ultra-fast SSDs with an open source data management tool providing the necessary access controls and audit logs. The environment currently supports 135 data linkage research projects.

# Appendix B – Analysis of Anonymisation Tools (Or supplementary material)

## Summary of the tools

### CTP:

CTP is the short name for the Medical Imaging Resource Center (MIRC) Clinical Trials Processor[21]. It used by radiology sites participating in multisite clinical trials to manage, process and transmit the medical images and their associated information. It features processing pipelines that the imaging datasets can be passed through as preparation for their use in research, including stages for anonymisation of the DICOM data. It is written in Java.

### XNAT

XNAT is an imaging informatics platform developed by the Neuroinformatics Research Group at Washington University, USA. It is designed with extensibility and customisability in mind (usually via third party or user-defined plugins), but has a core set of functional tasks common to most uses: data upload; data organisation and sharing; data viewing and downloading; secure and managed access; searching large data sets; and running complex processing on the data. It is written in Java.

### DICOM Confidential

DICOM Confidential is a DICOM anonymisation tool first developed at the University of Edinburgh in 2010 used by some imaging centres in Edinburgh. It comes with a graphical user interface and supports a common set of anonymisation tasks and can also be extended. It is written in JAVA.

## Core functionality

The core functionality was assessed. Here we provide examples of how the operations are undertaken via the respective tools, to give a flavour of the syntax necessary. The XML syntax for DICOM Confidential is long and not easily human readable so here we simply list the name of the transformation that would be used.

### Map a tag value to a new value using an external lookup

CTP can do this straightforwardly but only for a lookup file – this may raise some issues for our database solution. XNAT can do this but requires an external app to inject the new details into the script. DICOM confidential can do this via an external file.

|  |  |
| --- | --- |
| Task | Map (1010,0020) Patient ID “Case Report 1” to “1234” via an external lookup |
| CTP | @lookup(ElementName, KeyType) |
| XNAT | // Example for patientID  patientID := (1010,0020) // Sets the variable patientID to the initial value  describe patientID "ChangeME" // Sets an external label for the patient ID for use by the app  (1010,0020) := patientID // DICOM attribute gets set with the new variable |
| DICOM Conf | uk.ac.ed.dcmconf.transformer.idmapper.StudyIDMapper |

### Replace a value

|  |  |
| --- | --- |
| Task | Replace value for (0010,0010) PatientName to “XXX” |
| CTP | PatientName = “XXX” |
| XNAT | (0010,0010) := “XXX” |
| DICOM Conf | uk.ac.ed.dcmconf.transformer.field.StringOverwriter |

### Blank a value

|  |  |
| --- | --- |
| Task | Blank out the value for (0008,1050) PerformingPhysicianName |
| CTP | @empty(PerformingPhysicianName) |
| XNAT | (0080,1050) := “” |
| DICOM Conf | uk.ac.ed.dcmconf.transformer.field.StringOverwriter |

### Reduce granularity of a date

No direct command for this in XNAT but easy to write one.

|  |  |
| --- | --- |
| Task | Change (0008,0020) StudyDate to the first of the month |
| CTP | @modifydate(this,\*,\*,1) // assuming **this** is set to StudyDate |
| XNAT | myStudyDate := (0008,0020)  myMonthYear := substring(myStudyDate, 4, 10)  (0008,0020) := concatenate("01/",myMonthYear) |
| DICOM Conf | uk.ac.ed.dcmconf.transformer.field.DateTransformer |

### Remove all private data

|  |  |
| --- | --- |
| Task | Remove all private data i.e. (7FE1,x) for all x |
| CTP | Checkbox on UI “Remove Private Groups” |
| XNAT | **-** (7FE1,XXXX) |
| DICOM Conf | There is a transformer to do this. |

### Remove a tag

|  |  |  |
| --- | --- | --- |
| Task | Remove tag (0008,0081) InstitutionalAddress | |
| CTP | @remove() | |
| XNAT | **-** (0008,0081) | |
| DICOM Conf | uk.ac.ed.dcmconf.transformer.object.AttributeRemover |

### Additional core functionality

Other factors considered were:

* Whitelisting tags
* Adding bespoke anonymisations
* Defining subsets of images to which the rules will apply
* The ability to anonymise pixel data

## User friendliness

The evaluation of the user friendliness of the tools was considered from two points of view:

1. The person writing and maintaining the rules
2. The system administrator installing and maintaining the software

By their nature these criteria will be more subjective.

## Support

The criteria considered included existence of an active group of developers regularly working on the software, is there good documentation etc.

## Evaluation analysis

To give a more quantitative analysis following the evaluation, each tool was given a score between 1 (poor) and 5 (great) as to how straightforward the task was to achieve, or if there is support for the required feature and so on. Table 1,Table 2 and Table 3 show the scoring for the categories of Core Functionality, User Friendliness and Support respectively.

Table 1: Core functionality

|  |  |  |  |
| --- | --- | --- | --- |
| **Category** | **Tool** | **Comment** | **Score** |
| Remove a tag | CTP |  | 5 |
| XNAT |  | 5 |
| DC |  | 5 |
| Replace tag with hardwired value | CTP |  | 5 |
| XNAT |  | 5 |
| DC |  | 5 |
| Reduce granularity of a date | CTP |  | 5 |
| XNAT | No direct date command but do-able | 4 |
| DC |  | 5 |
| Map a value using database lookup | CTP | Mapping by file is default, database lookup needs bespoke extension | 4 |
| XNAT | Needs a bespoke plugin to do it, but it is supported | 4 |
|  | Mapping by file is default, database lookup need bespoke extension. | 4 |
| Remove all private data | CTP |  | 5 |
| XNAT |  | 5 |
| DC |  | 5 |
| Whitelist tags | CTP | Use the remove unchecked elements option. | 5 |
| XNAT | Can be done on initial import[[1]](#footnote-1) but is was not clear how to do it via scripts | 3 |
| DC | Looks like there is a transformer to do this (untested by us) | 5 |
| Add bespoke anonymisations | CTP | Supported and well documented. In Java. | 5 |
| XNAT | Supported, java | 5 |
| DC | Supported, Java. | 5 |
| Define subsets of images to which rules apply | CTP | Supports within tag if statements that allow some degree of conditionals | 2 |
| XNAT | Can create "bundles"[[2]](#footnote-2) which makes subsets available to others - should be possible to use this or similar to define subsets for our use | 3 |
| DC | Does not look possible | 2 |
| Pixel data anonymisation | CTP | Claims to do it well and have rules of all known cases | 5 |
| XNAT | Images can be updated[[3]](#footnote-3) | 3 |
| DC | Claims to support this. Not been able to get working. | 3 |

Table 2: User friendliness

|  |  |  |  |
| --- | --- | --- | --- |
| **Category** | **Tool** | **Comment** | **Score** |
| User friendliness of rules text | CTP | XML but quite flat and readable XML | 3 |
| XNAT | Straightforward syntax | 4 |
| DC | XML too ugly for users. Refers to full classpath etc. | 2 |
| User friendliness of rule GUI | CTP | Fairly clean but doesn’t parse the operation text | 4 |
| XNAT | Fine | 4 |
| DC | Poor, clunky and does not specify what parameters are needed. | 3 |
| Ability to write new GUI as part of other tools | CTP | Would be easy to write new GUI to spit out the same XML format | 5 |
| XNAT | Is supported via plugin development framework | 3 |
| DC | Would be easy to write new GUI to spit out the same XML format. | 5 |
| Ease of use of imagined best GUI for eDRIS staff | CTP | This is one of by possible concerns. The rules are written as text, e.g. @modifydate(this,\*,\*,1) and @lookup(this,pid). These are fine for developers, but it would be hard to use a GUI to hide this and expose a simpler parameters and values viewpoint. | 3 |
| XNAT | The rules use tags to reference what to change e.g. (0008,0080) := “Hospital A” so shares issue with CTP | 3 |
| DC | Rules have clear parameters and default values so a well written GUI could really help the user create rules | 5 |
| Command line invocation | CTP | Points to a directory and processes all files in it | 5 |
| XNAT | Linux only but expected it can be adapted for other platforms given the underlying software runs anywhere | 3 |
| DC | Points to a directory and processes all files in it. | 5 |
| Programmatical invocation | CTP | Not formally documented but it would be fairly easy to take the command line code and repackage that | 4 |
| XNAT | Yes, via plugin development[[4]](#footnote-4) and API [[5]](#footnote-5) | 4 |
| DC | Not formally documentation but it would be fairly easy to take the command line code and repackage it. | 4 |

Table 3: Support

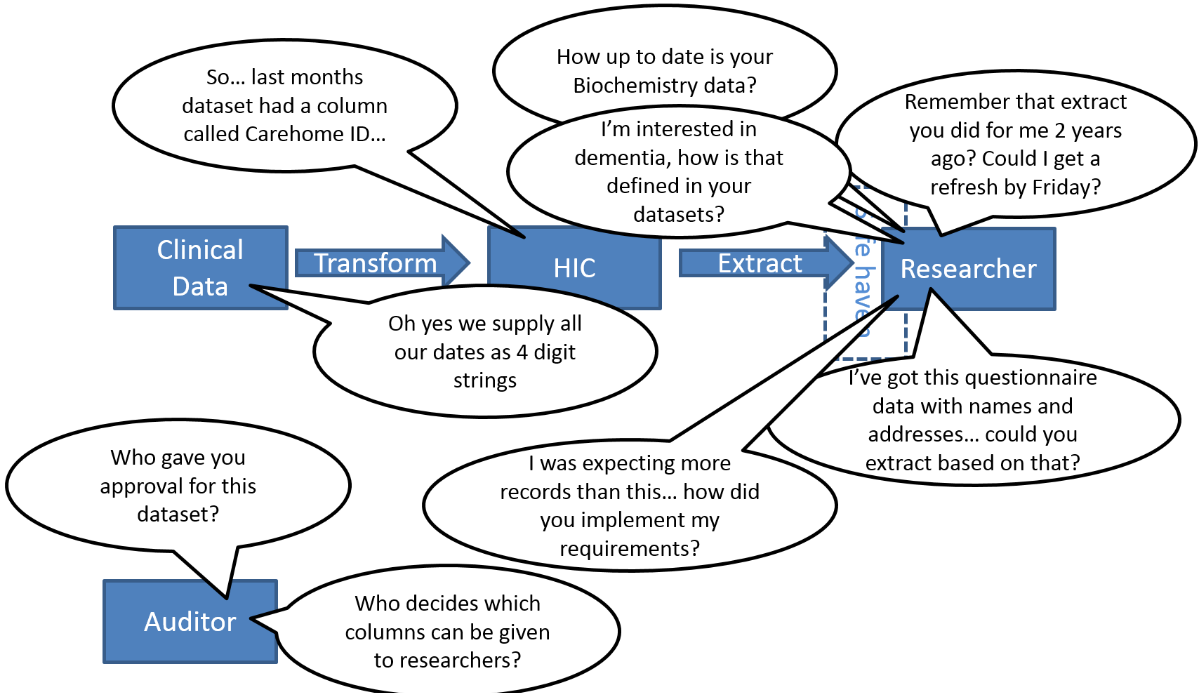
|  |  |  |  |
| --- | --- | --- | --- |
| **Category** | **Tool** | **Comment** | **Score** |
| Active development | CTP |  | 5 |
| XNAT | Active group of developers | 5 |
| DC | Original developer still support it. | 3 |
| Responsive to queries | CTP | Developer made a change overnight to the command line version following a request | 5 |
| XNAT | No personal experience of this but the community appears to be very active | 4 |
| DC | Original developer responded to personal emails | 4 |
| User documentation | CTP | Good webpage explaining the anonymisation operations | 5 |
| XNAT | Very good, comprehensive | 5 |
| DC | Very poor and we had to ask the developer for it. | 2 |
| Open source | CTP | RSNA MIRC public license [[6]](#footnote-6). | 5 |
| XNAT | Yes – just requires the inclusion of copyright notice in redistributions | 5 |
| DC | Available on-line for free but just JAR files and not the latest code. Developer gave me JAR files with the code. | 3 |
| Runs without bugs or obscure limitations | CTP | No bugs or limitations discovered in initial experiments | 5 |
| XNAT | Currently only linux command-line. No bugs spotted. | 3 |
| DC | Does not run on 64 bit windows. DICOM viewer unable to view the output images. | 2 |

# Draft pad – bits of info and text that require merging into paper

* It may not scale to support Petabytes of Images
* It would require a duplication of all of the data – an identifiable dataset and an anonymised dataset. For a range of reasons discussed in the anonymisation section we can’t discard the raw identifiable data in case it is needed later, and we can’t allow researcher access to a database which contains identifiable data underneath for data governance and security reasons.
* It does not facilitate complex cohort building and automated linking to other routinely collected EHRs without creating a huge data warehouse of permanently linked data from multiple sources. This is infeasible for data governance reasons especially when linking is required with sensitive data such as Police Data
* It does not align with the guiding principles of data linkage for research [30]
* Different groups require the images to be available in different structures for use by their own machine learning software or to be opened by their favourite image viewer and so the files need to be extracted out of the database
* The metadata is dirty – 5000 different tags plus private tags – don’t know how to standardise – how you do the standardisation differs for each research project Issues with research data images in contrast to real world clinical images

## Why is research data curation important?

Data management and data curation of long-term study and research databases are time-consuming and complex activities that demand the attention of experts with very specific skills. Some of the most costly and complex data management activities emerge from consideration of two common scenarios. The first considers a single cohort used in a longitudinal study accruing data in distinct phases where the new data must be reconciled and merged with the existing data sets. The second scenario occurs when distinct cohorts from different studies of the same disease are merged to create greater scale in the research data. Again the data must be merged and reconciled in order to create an aggregate data set that is valid in its totality.



*Figure 1 - Common questions asked of a research data host*

Existing data management approaches are focused on the initial generation and preparation of project research data and on preservation techniques that promote reuse of the data at the end of individual research projects. These approaches do not consider longer term studies and research programmes and fail to account for the key data merge, transformation and enrichment processes that are applied over life-time study lengths and that shape the data to support analysis and results. Failure to capture the project level transformation processes represents a major loss for long-lived research data sets, as data improvements identified by individual studies and cohorts are not fed back into larger aggregated data sets to extend the data and improve the data quality. For detailed information, see [1].

## What is the Research Data Management Platform?

The Research Data Management Platform (hereafter referred to as the RDMP) is a framework and suite of tools for the management and curation of longitudinal research datasets. It not only performs many typical ETL (Extract, Transform and Load) tasks, but also includes tools for management of the research lifecycle (see [1] for more details). The framework software focuses on ensuring thorough documentation of datasets, reliable loading of often poorly structured/volatile data, cohort linkage and reproducibility of project extracts.

### Roles of Users

There are a range of different roles of those interacting with the SMI platform:

To address challenge 1, these resources will use a Data Controller approved, standardised de-identification workflow. To mitigate the risks to the Data Controllers of providing unconsented routinely collected images for research (thus addressing challenge 2) these resources will work in accordance with the Scottish Government guiding principles for secure linking, anonymising and analysing data sets for research, where a subset of data for a specific cohort are linked for an approved research project and access provided via a Safe Haven (**SH**) environment (the NHS Scotland term for a Trusted Research Environment)[15]. Access to the data can be revoked by the Data Controller at any time and researchers cannot extract/output any information other than aggregate level results from the environment. There is a separation of the roles of indexer, linker (carried out by a trusted third party or Safe Haven staff) and researcher. Creating the SMI with DICOM format images and software tools to search for and build research cohorts will reduce the cost of providing images for research projects and will also provide a mechanism to return relevant images from different data sources (addressing challenges 3 and 4).

For those Research Coordinators that wish to use it, the RDMP suite provides a more integrated tool for cohort generation that will simplify the workflow and increase the reusability of cohort generation queries and sub-queries.

1. Snyder, P.J., et al., *Nonvascular retinal imaging markers of preclinical Alzheimer's disease.* Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring. **4**: p. 169-178.

2. Kickingereder, P., et al., *Radiomic subtyping improves disease stratification beyond key molecular, clinical, and standard imaging characteristics in patients with glioblastoma.* Neuro-Oncology, 2017: p. nox188-nox188.

3. Chaddad, A., et al., *Predicting survival time of lung cancer patients using radiomic analysis.* Oncotarget, 2017. **8**(61).

4. Rios Velazquez, E., et al., *Somatic Mutations Drive Distinct Imaging Phenotypes in Lung Cancer.* Cancer Research, 2017. **77**(14): p. 3922.

5. McGarry, S.D., et al., *Magnetic Resonance Imaging-Based Radiomic Profiles Predict Patient Prognosis in Newly Diagnosed Glioblastoma Before Therapy.* Tomography : a journal for imaging research, 2016. **2**(3): p. 223-228.

6. Yu, W., et al., *Development and Validation of a Predictive Radiomics Model for Clinical Outcomes in Stage I Non-small Cell Lung Cancer.* International Journal of Radiation Oncology • Biology • Physics.

7. Shen, C., et al., *2D and 3D CT Radiomics Features Prognostic Performance Comparison in Non-Small Cell Lung Cancer.* Translational Oncology. **10**(6): p. 886-894.

8. Gillies, R.J., P.E. Kinahan, and H. Hricak, *Radiomics: Images Are More than Pictures, They Are Data.* Radiology, 2015. **278**(2): p. 563-577.

9. Parmar, C., et al., *Machine Learning methods for Quantitative Radiomic Biomarkers.* Scientific Reports, 2015. **5**: p. 13087.

10. Hsu, C.-Y., et al., *Radiomics Features Differentiate Between Normal and Tumoral High-Fdg Uptake.* Scientific Reports, 2018. **8**(1): p. 3913.

11. Lambin, P., et al., *Radiomics: the bridge between medical imaging and personalized medicine.* Nature Reviews Clinical Oncology, 2017. **14**: p. 749.

12. *electronic Data Research and Innovation Service (eDRIS)*. Available from: <http://www.isdscotland.org/Products-and-Services/eDRIS/>.

13. *Minimum Viable Product (MVP)*. Available from: <https://en.wikipedia.org/wiki/Minimum_viable_product>.

14. Government', S. *Charter for Safe Havens in Scotland: Handling Unconsented Data from National Health Service Patient Records to Support Research and Statistics.* 2015.

15. *SHIP Guiding Principles and Best Practices*. 2010; Available from: <http://www.scot-ship.ac.uk/sites/default/files/Reports/Guiding_Principles_and_Best_Practices_221010.pdf>.

16. Nind, T., et al., *The Research Data Management Platform (RDMP): A novel, process driven, open-source tool for the management of longitudinal cohorts of clinical data.* GigaScience, 2018: p. giy060-giy060.

17. Lea, N.C., et al., *Data Safe Havens and Trust: Toward a Common Understanding of Trusted Research Platforms for Governing Secure and Ethical Health Research.* JMIR Medical Informatics, 2016. **4**(2): p. e22.

18. Burton, P.R., et al., *Data Safe Havens in health research and healthcare.* Bioinformatics, 2015. **31**(20): p. 3241-3248.

19. Williams, E., et al., *The Image Data Resource: A Scalable Platform for Biological Image Data Access, Integration, and Dissemination*. 2016, Cold Spring Harbor Laboratory.

20. L. Engstrom, C.M., *Dementias Platform UK Imaging Informatics* 2017.

21. Aryanto, K.Y.E., et al., *Implementation of an anonymisation tool for clinical trials using a clinical trial processor integrated with an existing trial patient data information system.* European Radiology, 2012. **22**(1): p. 144-151.

22. Marcus, D.S., et al., *Open Access Series of Imaging Studies: Longitudinal MRI Data in Nondemented and Demented Older Adults.* Journal of Cognitive Neuroscience, 2010. **22**(12): p. 2677-2684.

23. Swedlow, J.R., *Informatics and Quantitative Analysis in Biological Imaging.* Science, 2003. **300**(5616): p. 100-102.

24. Abbott, D. *What is Digital Curation?* DCC Briefing Papers: Introduction to Curation 2014; Available from: <http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation/what-digital-curation>.

25. MacGillivray, T.J., et al., *Retinal imaging as a source of biomarkers for diagnosis, characterization and prognosis of chronic illness or long-term conditions.* The British Journal of Radiology, 2014. **87**(1040): p. 20130832.

26. Rodríguez, D., et al., *An open source toolkit for medical imaging de-identification.* European Radiology, 2010. **20**(8).

27. *ClearCanvas*. Available from: <https://www.clearcanvas.ca/>.

28. *dcm4che*. Available from: <https://www.dcm4che.org/>.

29. *MicroDicom*. Available from: <http://www.microdicom.com/>.

30. Government, S., *Joined-up data for better decisions - Guiding Principles for Data Linkage*. 2012.

1. <https://wiki.xnat.org/docs16/xnat-configuration-framework/series-import-filter-configuration?src=sidebar> [↑](#footnote-ref-1)
2. <https://wiki.xnat.org/docs16/3-administrator-documentation/managing-sites-projects/sharing-stored-searches> [↑](#footnote-ref-2)
3. <https://wiki.xnat.org/documentation/how-to-use-xnat/image-session-upload-methods-in-xnat/where-anonymization-happens-in-xnat> [↑](#footnote-ref-3)
4. <https://wiki.xnat.org/display/XNAT17/Creating+an+XNAT+Plugin+Project> [↑](#footnote-ref-4)
5. <https://wiki.xnat.org/documentation/the-xnat-api> [↑](#footnote-ref-5)
6. <http://mirc.rsna.org/rsnapubliclicense.pdf> [↑](#footnote-ref-6)