Shanoir NG – Import

Software Design Description

Table des matières

[1 Microservice context 2](#_Toc498069305)

[1.1 Microservice presentation 2](#_Toc498069306)

[1.2 Main functionalities 2](#_Toc498069307)

[1.3 Application users 2](#_Toc498069308)

[2 Functional architecture 2](#_Toc498069309)

[2.1 Imports 2](#_Toc498069310)

[2.1.1 Import from DICOM CD/DVD 2](#_Toc498069311)

[2.1.2 Import from PACS 2](#_Toc498069312)

[2.1.3 Import from ShanoirUploader 2](#_Toc498069313)

[2.1.4 Import processed dataset (NifTi) 2](#_Toc498069314)

[2.1.5 Import processed dataset (Special case for Nifti files) 3](#_Toc498069315)

[2.1.6 Import from files 3](#_Toc498069316)

[2.2 Steps 3](#_Toc498069317)

[2.2.1 Choose modality 3](#_Toc498069318)

[2.2.2 Upload files 3](#_Toc498069319)

[2.2.3 Select series 4](#_Toc498069320)

[2.2.4 Select clinical context 4](#_Toc498069321)

[2.2.5 Describe dataset 5](#_Toc498069322)

[2.2.6 Select/create data processing 6](#_Toc498069323)

[2.2.7 Anonymization 6](#_Toc498069324)

[2.2.8 Nifti conversion 6](#_Toc498069325)

[2.2.9 Select settings for research study 6](#_Toc498069326)

[3 Software/technical Architecture 7](#_Toc498069327)

[3.1 Anonymization 7](#_Toc498069328)

[3.1.1 Anonymized fields : 7](#_Toc498069329)

[3.1.2 Prerequisite 7](#_Toc498069330)

[3.1.3 Steps 7](#_Toc498069331)

[3.1.4 Technical Specification of Shanoir NG Anonymization 7](#_Toc498069332)

[3.1.5 Email 12](#_Toc498069333)

[3.1.6 New folders 12](#_Toc498069334)

# Microservice context

## Microservice presentation

This microservice is a part of the Shanoir-NG application and is used to import data.

## Main functionalities

Shanoir NG main functionalities are:

* Import from DICOM CD/DVD
* Import from PACS
* Import from ShanoirUploader
* Import processed dataset (NifTi)
* Import from files

## Application users

Target population is:

* Administrators
* Experts
* Users

In other words, if the user’s role is guest, he could not see the import menu.

# Functional architecture

## Imports

### Import from DICOM CD/DVD

Steps of this import are:

1. Upload DICOM zip
2. Select series
3. Select clinical context
4. Finish import and go back to the first step

### Import from PACS

Steps of this import are:

1. Query PACS
2. Select series
3. Select clinical context
4. Finish import and go back to the first step

### Import from ShanoirUploader (TODO)

Steps of this import are:

* Choose modality
* Upload files
* Anonymization
* Select series
* Nifti conversion

### Import processed dataset (NifTi) (TODO)

Steps of this import are:

* Upload files
* Describe dataset
* Select/create data processing
* Nifti conversion

### Import processed dataset (Special case for Nifti files) (TODO)

For imports started directly on the Details’ page of dataset/study

Steps of this import are:

* Upload files
* Describe dataset – without Study and Subject.
* Create data processing – new processing is automatically created - user chooses the type, date and comment

### Import from files (TODO)

Import from files allows to upload:

* Processed dataset (NifTi)
* Extra data
* Spectroscopy data

Steps to import processed dataset are:

* Upload files
* Describe dataset
* Select/create data processing
* Nifti conversion

Steps to import extra data are:

* Upload files
* Select clinical context

Steps to import spectroscopy data are:

* Upload files
* Select clinical context

## Steps

### Upload data

In this first step, depending on the chosen import type, the data to be uploaded could either from a file (see 2.2.1.1) or from the PACS (see 2.2.1.2).

#### Upload File

User selects a file to upload.

* For import from CD/DVD: zip with DICOM images
  + DICOMDIR is required here at the root of the zip. Otherwise, an error is displayed and the import is stopped here.
  + The modality of the datasets imported is read and displayed. Two modalities are available here: MRI or PET.
* Import Nifti: .nii or .nii.gz files
* Extradata (From files): any files

#### Query PACS

The query to the PACS is defined with the five fields below:

* Patient name
* Patient ID
* Patient birth date
* Study description
* Study date

Here are the rules of the query:

* At least one of the five fields should be present
* Two data fields are with the format “YYYYMMDD” (?To be changed)
* Only users with the Boolean “can import from PACS” could see this menu in nav-bar
* Json post should be sent to Import MS “/importer/query\_pacs” api

### Select series

A tree is displayed here to let users select series. Here is the structure of the tree:

* Patient
* Study
* Series
* Series
* Series
* …

The tree is an response from the import MS. The json object respond is described in the “**1. upload\_dicom.json**” file.

* For one patient, the information as patient ID, sex, name and birth date are displayed on click on the patient level.
* For one study, the study description is read from the dicom files. The value could be changed and overwritten in the dicom files.
* For each series, the information as series ID, protocol, series description, series date, series number, number of images, modality and number of non-imaged objects are displayed on click on the series level. Users could preview all the dicom images of each series here on click on the “load preview” button. The papaya viewer is called then.

Users could select here one or more series on click either on series level or on study level. If at least one series is selected, the “Next” button is available and we pass to the next step.

### Select clinical context

The clinical context is based on two responses from the studies MS. The json objects are described in the “**2. list\_for\_import.json**” and the “**2. centers.json**”file. For more details about the calculate of compatibility, see **3.3.1 Calculate of compatibility** part.

#### Select an existing research study

* User chooses from the list of studies available for him (studies he has access to, with right “**Can see, download and import datasets**”, “**Can see, download, import datasets and modify the study parameters**” or “**Is responsible for the research study**”)
* Once chosen, **Details** button appears on the right that shows details on the study in a new tab
* Once chosen, the list of centers is initialized with the study chosen.
* One chosen, the default Nifti Converter of the study will be displayed and used

#### Select a center

* User chooses from the list of centers associated to the study
* Once chosen, **Details** button appears on the right that shows details on the center in a new tab
* One chosen, the list of acquisition equipment is initialized with the study and center chosen.

#### Select an acquisition equipment

* User chooses from the list of acquisition equipment associated to the study and the center. On the list the name appears as: Manufacturer – Manufacturer model – (Magnetic fields if MR) – Modality – Serial Number - Center
* Once chosen, **Details** button appears on the right that shows details on the acquisition equipment in a new tab
* One chosen, the list of subjects is initialized with the study chosen. The list is an response from the studies MS. The json object respond is described in the “**3. allSubjects.json**” file.

For the parts 2.2.3.1 to 2.2.3.3, a flag “**compatible**” is calculated from two json responses of the studies MS, for each of the study, center and acquisition equipment level:

* One acquisition equipment is considered as compatible if the manufacturer, the manufacturer model and the serial number are exactly the same as these three tags read from the dicom files.
* One center is considered as compatible if at least one of its acquisition equipment is compatible.
* One study is considered as compatible if at least one of its centers with its acquisition equipment is compatible.

The compatible entities are displayed in green and the not-compatible ones are in red.

**To save the time, in case of one and only one compatible study, center or acquisition equipment, the compatible entities are automatically selected.**

If there is not any compatible center or acquisition equipment, users (which role: user, expert or only admin?) could create a new one on click on the add button, with the information read from the dicom files. A warning message is displayed as below:

*“This center may not be not compatible with the imported data !*

*Your archive contains this data : CHU PONTCHAILLOU, Henri le Guilloux 2,Rennes/95A5EC/,District,FR,35000.*

*If the center you need is not in the above list, you can create a new center from this data by clicking on the 'add' button.”*

On the creation page, users are also guided by a message as below:

“Your archive contains the following data :

CHU PONTCHAILLOU, Henri le Guilloux 2,Rennes/95A5EC/,District,FR,35000*”*

#### Subject

* User chooses either a subject that exists already in the research study or decided to create a new one.
* If existing subject: chooses one from the list of common names. On the side, there is a link to details page. On the bottom there is an information about the subject study relationship:
  + Common name (cannot be modified)
  + Subject identifier for the study
  + “Is physically involved” checkbox
  + Subject type (patient or healthy volunteer)
* If new subject: subject creation page with several information pre-filled:
  + Imaged object category: Living humain being
  + First name and last name are calculated from the patient name dicom tag
  + Birth date
  + Sex

A common name is required to finish the subject creation.

* After creation, this subject is automatically chosen
* The list of examination is initialized with the subject and the study chosen. The json file is described in the “**4. findExaminationsBySubjectAndStudy.json**” file.

#### Examination

* Use can select from an existing examination (if there is any for this subject in this study) or can create a new one
* If existing: list of examination appears in the form: Date + Comment/Name + (ID). Once chosen, Details link appears as well.
* If new: examination creation page with several information pre-filled and grey out:
  + Research study
  + Subject
  + Center
  + Examination Date which is a copy of the series date of the 2.2.2 part
  + Comment which is a copy of the study description of the 2.2.2 part
* After creation this examination is automatically chosen

### Finish import and go back to the first step

During this last step, anonymization (see 2.2.4.1) and nifti conversion (see 2.2.4.2) should be done. Users are redirected to the first step (2.2.1).

Another json file “**5. start\_import\_job.json**” is sent to import MS. This json is similar to the json file “**1. upload\_dicom.json**”. They have the same format but in this step, the “**5. start\_import\_job.json**” file is filled up with information got during the import, as the meta data for the dicom images, the “fromDicomZip” Boolean, the ids of study, acquisition equipment, examination and nifti converters used during the import.

#### Anonymization

This page shows all the fields that are going to be anonymized. The column in the table are as follow:

* DICOM field number
* DICOM field description (Signification)
* Initial value
* Final value

#### Nifti conversion

Nifti conversion is done after anonymization.

* For DICOM import: the default Nifti converter of the selected study is displayed and used for the conversion.
* For Nifti import: conversion only to one nifti file.

# Software/technical Architecture

## Clinical context

### Calculate of compatibility

Two json are got from the studies MS. The first one contains the list of studies which the user is allowed to import into. The second one contains the list of all centers.

These two lists are filtered to complete a new list of studies which contains the information about all the centers which are attached to each study of the list. Each centers contains information about all the acquisition equipments that each center owns. If one acquisition equipment’s serial number, manufacturer and manufacturer model are exactly the same with the information present in the dicom files, this acquisition equipment is then considered as compatible for the import. The center which contains this acquisition equipment is then considered as compatible, and also the study which contains this center.

### Subject (To be discussed)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Json 1  *subject* | Find y | Find y | ~~Find y,z~~ | ~~Find y,z~~ | ~~Find y,z~~ | Find ∅ |
| Json 3 | Find [X]  y ∉ [X] | Find [X]  y ∈ [X] | ~~Find [X]~~  ~~y ∈ [X]~~  ~~z ∉ [X]~~ | ~~Find [X]~~  ~~y,z ∈ [X]~~ | ~~Find [X]~~  ~~y,z ∉ [X]~~ | Find [X] |
| IHM neurinfo | Create subject  OR  add y to [X] list (preselected by default)  OR  Selection of subject (change preselection) | Create subject  OR  add y to [X] list (preselected by default)  OR  Selection of subject (change preselection) | ~~select y~~  ~~Put z as second choice in list with warning z ∉ [X]~~  ~~+ add [X] - y to list~~ | ~~select y~~  ~~Put z as second choice in list~~  ~~add [X] - y - z to list~~ | ~~select y~~  ~~Add warning : y ∉ [X]~~  ~~Put z as second choice in list with warning z ∉ [X]~~  ~~+ add [X] to list~~ | No preselection  Create subject  OR  Selection of subject |
| IHM ofsep | TBD |  |  |  |  |  |

**Changes in the table are justified because we can have an anonymized patient with the same name and birth date in the zip file for different real patients.**

## Anonymization

### Anonymized fields :

See anonymization.xlsx document

### Prerequisite

Zip files should be unzipped.

Check if zip files are unzipped during anonymization step or before.

### Steps

* Anonymization
* Send email
* Create folders

### Technical Specification of Shanoir NG Anonymization

DICOM tags to anonymize are listed in an excel document named anonymization.xlsx. This document is based on [*DICOM Supplement 142: Clinical Trial De-identication Profiles*](ftp://medical.nema.org/medical/dicom/final/sup142_ft.pdf). This document matches one tag to anonymize to a code. Depending on this code, an action applies to the tag. The following actions are used in the table:

* D – replace with a non-zero length value that may be a dummy value and consistent with the VR
* Z – replace with a zero length value, or a non-zero length value that may be a dummy value and consistent with the VR
* X – remove
* K – keep (unchanged for non-sequence attributes, cleaned for sequences)
* C – clean, that is replace with values of similar meaning known not to contain identifying information and consistent with the VR
* U – replace with a non-zero length UID that is internally consistent within a set of Instances
* Z/D – Z unless D is required to maintain IOD conformance (Type 2 versus Type 1)
* X/Z – X unless Z is required to maintain IOD conformance (Type 3 versus Type 2)
* X/D – X unless D is required to maintain IOD conformance (Type 3 versus Type 1)
* X/Z/D – X unless Z or D is required to maintain IOD conformance (Type 3 versus Type 2 versus Type 1)
* X/Z/U\* – X unless Z or replacement of contained instance UIDs (U) is required to maintain IOD conformance (Type 3 versus Type 2 versus Type 1 sequences containing UID references)

The document contains many profile. These profiles, based on a basic one, allow to different kinds of anonymization. A profile matches an action to a tag. For a tag, anonymization action can be different depending on profile.

Thus, the anonymization service parses the excel file, reads the list of tags and the operation to do to anonymize each tag based on the anonymization profile that should be applied (figure 2).

The algorithm reads the DICOM tags of DICOM image using DCM4CHE3 library and anonymizes each tag present in the DICOM file that are mentioned in the excel file. For so, it follows several steps:

* Treats public tags: public tags specified in the excel file have to be anonymized.

We distinguish some tags that should be anonymized differently to keep coherent dicom file or to be able to store the resulting anonymized file in DCM4CHEE PACS (see figure 1).

* Treats private tags: assume that a DICOM tag is represented by (gggg, eeee) form. All private tags recognized when the “gggg” part is odd should be anonymized (see figure 1).
* Treats “xx” tags: some tags belonging to the same interval have to be anonymized the same way. For instance, curve data recognized when the tag is beginning with “50” should be removed. Thus all tags (50xx,xxxx) should be removed after anonymization.

In this step, we treat three particular use cases which are: (50xx,xxxx), (60xx,4000) and (60xx,3000) tags (see figure 1).

* Treats Shanoir special tags: to map to Shanoir specifications, some tags should be differently anonymized:
  + SOPInstanceUID (0008,0018)
  + SeriesInstanceUID (0020,000e)
  + StudyInstanceUID (0020,000d)
  + PatientName (0010,0010)
  + PatientID (0010,0020)
  + PatientBirthDate (0010,0030)
  + ContrastBolusAgent (0018,0010)

The whole anonymization process is described by figure 1.

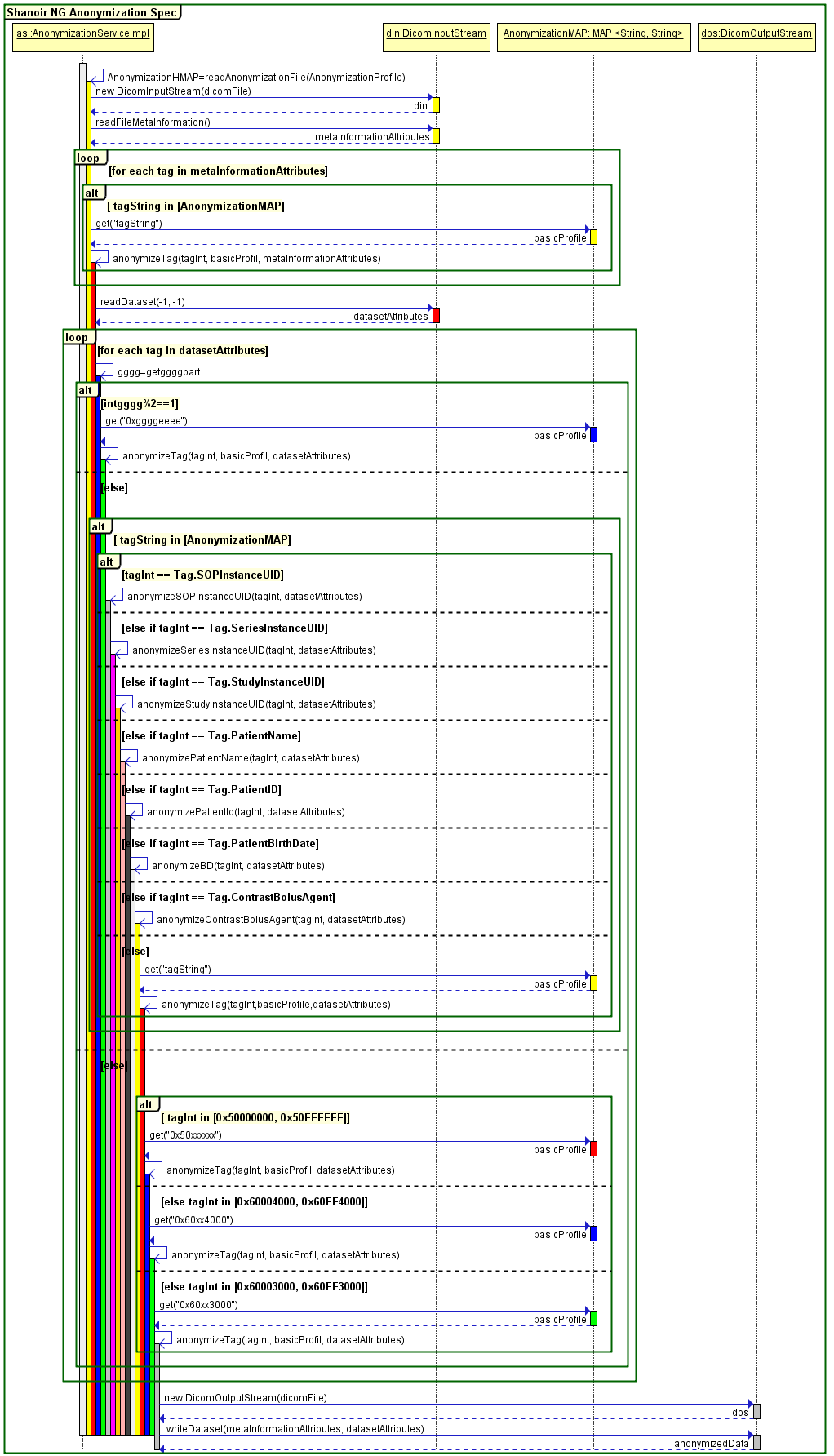


Figure 1. DICOM image anonymization process

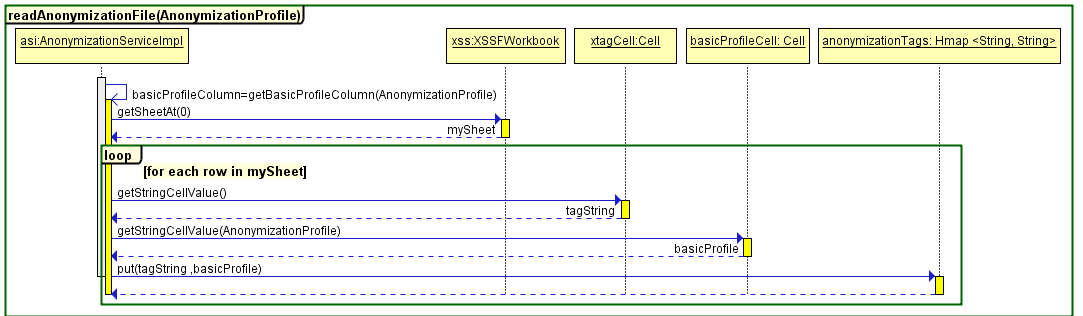


Figure 2. Excel anonymization tags file reading

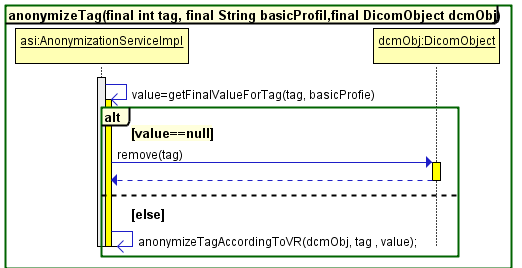


Figure 3. DICOM tag anonymization process

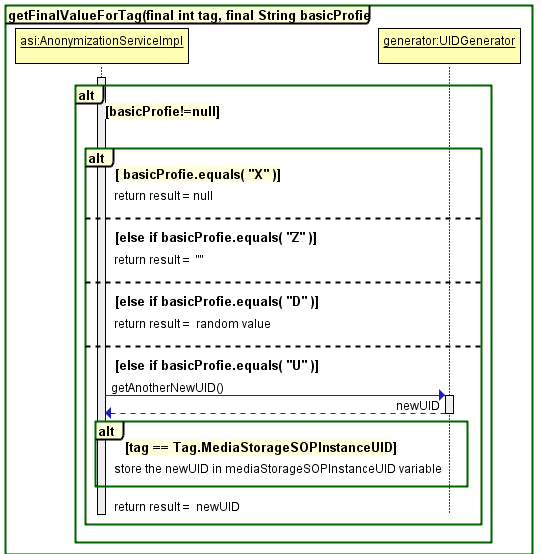


Figure 4. Anonymized tag value based on profile

#### UIDs anonymization

##### UIDs rules

To anonymize UIDs, they should be replaced by non-zero length UIDs that are internally consistent within a set of Instances.

There are several rules to be followed when making UIDs (<https://www.medicalconnections.co.uk/kb/UID_Rules>):

* Total length must be <= 64 characters, including the stops
* Must contain **only** digits 0-9 and full stops
* Each numeric "component" (between stops) must be a valid and unambiguous integer number, and so must **not** have a leading zero (unless the whole component is zero)
* Must be guaranteed to be unique - this means:
  + It must be derived from a proper official root under **your** sole control.
  + It must **not** be created by appending digits (however special you consider the combination!) to **someone else's** UID.
  + In particular, series UIDs for secondary capture images, KIN objects etc. must **not** be created as derivatives of the Study UID (unless you own that root!)
* Related to the above, there is **no** expectation or requirement that the Study UID, Series UID and Instance UID for images should be derived from the same root (though in practice, Series UID and Instance UID normally are, as both must be generated internally by the equipment which generates the images)
* Date and Time are useful for generating UIDs, but only if:
  + Each machine has a unique root (normally your company UID root + a machine specific suffix such as a serial number
  + If it is possible for UIDs to be generated at > 1 per second, then a sequential counter should also be used
  + If on a multi-threaded machine, then the thread ID or a properly interlocked counter are needed to prevent 2 applications or 2 threads in the same application from generating identical UIDs simultaneously.
  + Do not use time on its own - it is too easy to end up with a leading zero 0 - e.g. 20060724.093017 use instead 20060724093017

##### New UIDs generation

The anonymized UIDs are generated using the java source code of PixelMed Toolkit.

PixelMed toolkit is a stand-alone DICOM toolkit that implements code for reading and creating DICOM data, DICOM network and file support, a database of DICOM objects, support for display of directories, images, reports and spectra, and DICOM object validation.

The toolkit is a completely new implementation, which does not depend on any other DICOM tools, commercial or free. It does make use of other freely available pure Java tools for compression and XML and database support.

The toolkit is open source and made freely available for both non-commercial and commercial use, under [BSD License](http://www.opensource.org/licenses/bsd-license.php) terms.

##### DCM4CHEE constraints to correctly store DICOM data

###### DCM4CHEE constraints to accept to store data

SOPInstanceUID is a unique identifier (0008,0018) for the DICOM dataset. MediaStorageSOPInstanceUIDis thefirst value of the Media Storage SOP Instance UID (0002,0003).

**MediaStorageSOPInstanceUID should have the same value as in the SOPInstanceUID**. The only difference is that MediaStorageSOPInstanceUID is stored in the meta-information header of the DICOM file so that it becomes easy to read and fast to access while the SOPInstanceUID is stored in the dataset.

###### DCM4CHEE constraints to correctly visualize data

**Patient Identification**

If the issuer of patient id is null, dcm4chee attempts to identify the patient by name+id+dob[[1]](#footnote-1).

The Patient Matching code has problem if the patient name in the database is incomplete and is "truncated", e.g. the patient has no middle name and the patient name in the database is 'SURNAME^GIVEN\_NAME' rather than 'SURNAME^GIVEN\_NAME^^^'.

**Study/Serie** **Identification**

Study Instance UID Attribute is a unique identifier for study, so all images acquired for the same study should have the same Study Instance UID.

Series Instance UID Attribute is a unique identifier for the Series that is part of the Study identified in Study Instance UID (0020,000D). So all images acquired for the same series should have the same Series Instance UID.

### Email

Send email

Define email content and when it is sent.

### New folders

TODO

1. dob: date of birth [↑](#footnote-ref-1)