

CAP De-Identification Protocol

v. 1.0 (December 2013)



Definitions

Acronyms

HIPAA	Health Insurance Portability and Accountability Act
DICOM	Digital Imaging and Communications in Medicine
LONI	Laboratory Of Neuro Imaging
IDA	Image Data Archive
CAP	Cardiac Atlas Project
PHI	Protected Health Information
IRB	Institutional Review Board
PACS	Picture Archiving and Communication System

Protected Health Information

A primary goal of the Cardiac Atlas Project¹, a NIH funded research project, is to establish a database of clinical images of the heart. Providing access to medical images involves a high degree of attention to the protection of patient information. The HIPAA Privacy Rule and Public Health² from the U.S. Department of Health and Human Services provide guidelines for protected health information (PHI). It describes what type of information is regarded as PHI, principles for the use and disclosure of the data, and ways to de-identify information.

LONI Debabeler

The LONI Debabeler³ is a software package developed at the UCLA for de-identification and transfer of medical image data. The UCLA IRB has approved the Debabeler as HIPAA compliant. The Debabeler is configurable using mapping files, thus allowing the users to provide missing information and correct erroneous metadata values, as well as expand abbreviations, decipher

¹ <http://www.cardiacatlas.org>

² The entire Privacy Rule may be found at: <http://www.hhs.gov/ocr/hipaa>

³ Scott C. Neu, Daniel J. Valentino and, Arthur W. Toga, The LONI Debabeler: a mediator for neuroimaging software, NeuroImage, Volume 24, Issue 4, 15 February 2005, Pages 1170-1179, ISSN 1053-8119, DOI: 10.1016/j.neuroimage.2004.10.035.

enumerations, and convert between units. To ensure compliance to the HIPAA, we have identified PHI in the data provided for this project and implemented a de-identification mapping⁴ for the LONI Debabeler.

Protected DICOM Attributes

An implementation claiming conformance to the Basic Application Level Confidentiality Profile as a de-identifier in the scope of a health science project shall protect all instances of the Attributes listed below. In addition, an approval of the MESA steering committee allows an extended limited data set for the CAP, which includes: age (years), gender (M/F), height (cm), weight (kg), systolic and diastolic blood pressure (mmHg), hypertension (y/n), heart rate (bpm), race/ethnicity (class), diabetes (class), smoking (y/n), alcohol (y/n), ECG (class). Therefore, the tags to be protected (de-identified) are as follows⁵:

- (0008,0014) Instance Creator UID
- (0008,0018) SOP Instance UID
- (0008,0012) Instance Creation Date
- (0008,0020) Study Date
- (0008,0021) Series Date
- (0008,0022) Acquisition Date
- (0008,0023) Content Date
- (0008,0050) Accession Number
- (0008,0080) Institution Name
- (0008,0081) Institution Address
- (0008,0090) Referring Physician's Name
- (0008,0092) Referring Physician's Address
- (0008,0094) Referring Physician's Telephone Numbers
- (0008,1010) Station Name
- (0008,1030) Study Description
- (0008,103E) Series Description
- (0008,1040) Institutional Department Name
- (0008,1048) Physician(s) of Record
- (0008,1050) Performing Physicians' Name
- (0008,1060) Name of Physician(s) Reading Study
- (0008,1070) Operators' Name
- (0008,1080) Admitting Diagnoses Description
- (0008,1155) Referenced SOP Instance UID
- (0008,2111) Derivation Description
- (0010,0010) Patient's Name
- (0010,0020) Patient ID

⁴ SVN repository: <http://sourceforge.net/projects/cardiacatlas/>

⁵cp. DICOM Base Standard PS 3.15-2008 page 35 ff

- (0010,0030) Patient Birth Date
- (0010,1000) Other Patient Ids
- (0010,1001) Other Patient Names
- (0010,1090) Medical Record Locator
- (0010,2180) Occupation
- (0010,21B0) Additional Patient's History
- (0010,4000) Patient Comments
- (0018,1000) Device Serial Number
- (0020,000D) Study Instance UID
- (0020,000E) Series Instance UID
- (0020,0010) Study ID
- (0020,0052) Frame of Reference UID
- (0020,0200) Synchronization Frame of Reference UID
- (0020,4000) Image Comments
- (0040,0275) Request Attributes Sequence
- (0040,A124) UID
- (0040,A730) Content Sequence
- (0088,0140) Storage Media File-set UID
- (3006,0024) ReferencedFrameofReferenceUID
- (3006,00C2) Related Frame of Reference UID

De-Identification Process

The LONI Debabeler acts as a mediator between imaging software packages by automatically using an appropriate file translation to convert files between each pair of linked packages. These translations are built and edited using the Debabeler graphical interface and compensate for package-dependent variations that result in interpackage incompatibilities. The Debabeler gives imaging environments a configurable automaton for file translation and provides users a flexible application for developing robust solutions to translation problems. Solutions to translation problems can be developed within the Debabeler's visual environment. Its Java graphical user interface provides access to libraries of functional modules that can be graphically connected together between input and output trees of image data and metadata. This configurability enables users to provide missing information and correct erroneous metadata values, as well as expand abbreviations, decipher enumerations, and convert between units. With the appropriate processors, proprietary metadata may be decoded and conversions can be applied between different pixel data types. The individual rules for image conversions are stored as *Debabeler mappings*. This allows the use of appropriate mappings on a case-by-case basis.

DICOM Value Representation

For each public attribute, the DICOM manual associates one of the following VR:

- AE Application Entity
- AS Age String
- AT Attribute Tag
- CS Code String
- DA Date
- DS Decimal String
- DT Date/Time
- FL Floating Point Single
- FD Floating Point Double
- IS Integer String
- LO Long String
- LT Long Text
- OB Other Byte
- OF Other Float
- OW Other Word
- PN Person Name
- SH Short String
- SL Signed Long
- SQ Sequence of Items
- SS Signed Short
- ST Short Text
- TM Time
- UI Unique Identifier
- UL Unsigned Long
- UN Unknown
- US Unsigned Short
- UT Unlimited Text

DICOM uses three different Data Element encoding schemes. With Explicit VR Data Elements, for VR's that are not OB, OW, OF, SQ, UT, or UN, the format for each Data Element is:

- GROUP (2 bytes),
- ELEMENT (2 bytes),
- VR (2 bytes),
- LengthInByte (2 bytes),
- Data (variable length).

General De-Identification

The Debabeler can decode all DICOM tags translated using the public DICOM dictionary. There are however private (manufacturer-defined) tags that can be decoded but have no public definition.

All DICOM tags have an associated data type (value representation = VR) and tag number. Typically, DICOM tags with the following VR's: (AE, AS, AT, CS, DS, FD, FL, IS, SL, SS, TM,

UI, UL, US) are not removed or replaced within the CAP de-identification. These are considered “safe” data types that should not contain patient information. E.g., because the 0008,0008 tag has VR=CS, it will not be removed nor replaced in the de-identified DICOM file. Typically all DICOM tags with the VR=UI are encrypted and hashed into a long integer. This long integer is appended to UCLA’s purchased DICOM prefix (2.16.124.113543.6006.99.), which is under the mandatory 64-character limit. This preserves uniqueness of the images while masking identity. DICOM tags with VR’s that may contain patient information (DA, LO, OB, OW, PN) are automatically be removed from the de-identified DICOM file unless purposefully kept. For example, the study/series dates and descriptions are kept, because the Image Data Archive (IDA) requires that information for organizational purposes.

CAP Specific De-Identification

Using the CAP specific mapping file, the DICOM de-identification works in the following way:

(1) “Split by tag”: DICOM tags that have allowed tag numbers are left unchanged. These include tags that contain unique identifiers that are not to be encrypted, such as trusted dates, trusted binaries, and trusted strings. Specified tags for the CAP are:

[0002,0001; 0002,0002; 0002,0010; 0002,0013; 0008,0012; 0008,0016; 0008,0020; 0008,0021; 0008,0022; 0008,0023; 0008,0050; 0008,0070; 0008,0080; 0008,0100; 0008,0102; 0008,0104; 0008,1030; 0008,103E; 0008,1080; 0008,1090; 0009,0010; 0009,1002; 0009,1004; 0010,0030; 0010,2160; 0010,2180; 0011,0010; 0018,0024; 0018,0031; 0018,0085; 0018,1020; 0018,1030; 0018,1085; 0018,1250; 0018,1251; 0019,0010; 0019,1018; 0019,101A; 0019,109C; 0019,109D; 0019,109E; 0019,10D3; 0020,0010; 0020,1040; 0021,0010; 0023,0010; 0025,0010; 0025,101A; 0025,101B; 0027,0010; 0027,1030; 0029,0010; 0043,0010; 0043,1028; 0043,1029; 0043,102A; 0043,102D; 0043,102E; 0043,1062; 0043,106F]

(2) “Split by VR”: DICOM tags not selected in step (1) are left unchanged if they have allowed VR’s. Numbers, times, and small code strings are typically left unchanged. Strings and binaries are not. Specified valid VR’s for the CAP are:

[AE; AS; AT; CS; DS; FD; FL; IS; SL; SS; TM; UI; UL; US]

(3) “Encrypt”: DICOM tags not selected in step (2) are either encrypted or discarded. Specified for the CAP:

[discard].

(4) “Replace”: Out of all the DICOM tags in steps (1), (2), and (3), the values of specified tags are replaced. Physician names are replaced with empty strings as well as the patient name, and the patient identifier is replaced with a user-specified one. Default values for missing tags are also set. Specification for the CAP:

- UID’s except [0002,0002; 0002,0010; 0008,0016] : replaced with encrypted value
- Dates : rounded to year
- Patient’s Name : blanked

- Patient ID : replaced with CAP identifier specified on execution
- Patient's Birth Date : rounded to year
- Device Serial Number : replaced with encrypted value

Note that:

- The month/day of each date is removed.
- Within the limited data set approved for the CAP, certain DICOM tags such as the Age (0010,1010) are preserved.
- The Study ID (0020,0010) is preserved, since it is a numerical Study identifier generated by the equipment (usually a small integer: 1, 2, etc.).
- Acquisition comments (0018,4000) are removed because they are long free-form text that can potentially contain patient information.

Using LONI Debabeler

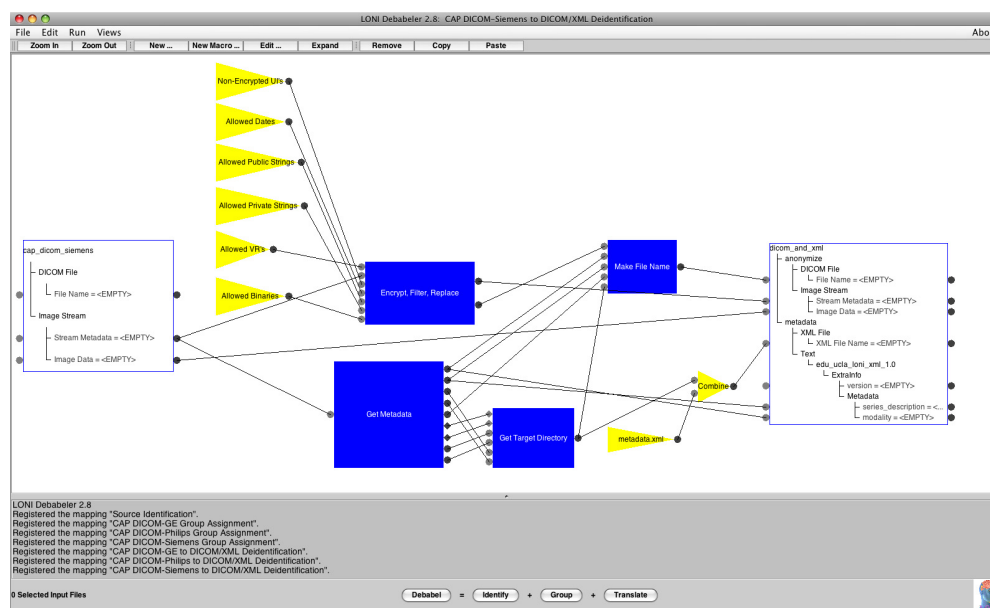


Figure 1: Screenshot of the LONI Debabeler using the CAP specific mapping file.

LONI Debabeler is a graphical tool to translate medical imaging format (not only just DICOMs) by using a mapping XML file. CAP has created a specific mapping file (CAPDeIdentification.xml), which anonymize metadata according to the previously mentioned rules. Editing specific de-identification customization can be performed through this graphical representation. Please refer the LONI Debabeler manual for a complete documentation on how to edit a mapping file.

The LONI Debabeler can de-identify a set of image files directly through its GUI, but it may take a lot of time, particularly for large image data. Fortunately, LONI Debabeler has a feature to

save the whole de-identification process that has been attached with the mapping XML file into an executable JAR file (menu: File -> Save as Executable JAR File).

The CAP's executable JAR file for de-identification is: `CapDicomDeidentifications.jar`

Running the CAP-specific JAR for de-identification

The Debabeler⁶ has been tested and successfully run on Windows, Solaris, Linux, IRIX, and OS X platforms. To process the large amount of Image data for the CAP, we use an executable jar file based on the Debabeler in conjunction with the CAP specific mapping, which can be included in a script for automatic processing of a large number of files. This jar file can be executed using Java 1.4 or higher on the command line with the following usage:

```
java -jar CapDicomDeidentifications.jar -input <args> -args <args> -  
target <args> -suppress -recursive
```

where:

input	⇒	Input files and/or directories
args	⇒	Cmd line arguments (e.g. output directory)
target	⇒	What type of processing should occur
suppress	⇒	Use to suppress verbose messages
recursive	⇒	Use to recursively search input directories

To execute the CAP specific jar file, the command line arguments are:

```
java -Xmx800m -jar CapDicomDeidentifications.jar -input <src_dirs> -  
target anonymize metadata -args <output dir> <log file> <project_name>  
<new_patient_id_<=_10_chars>
```

Subdirectories will be created for later use in the upload process. Suggested Patient ID's for the images are study name in conjunction with incremental numbering for each case, e.g. MES0000001, MES0000002, etc. The Debabeler mapping for the CAP includes separate de-identifications for GE, SIEMENS, and Philips scanners. Based upon our experience, we expect that they will work well for images from these scanners. Note that capturing the command line output allows the subsequent identification of source data and the according de-identified data.

⁶ <http://www.loni.usc.edu/Software/Debabeler>

Validation

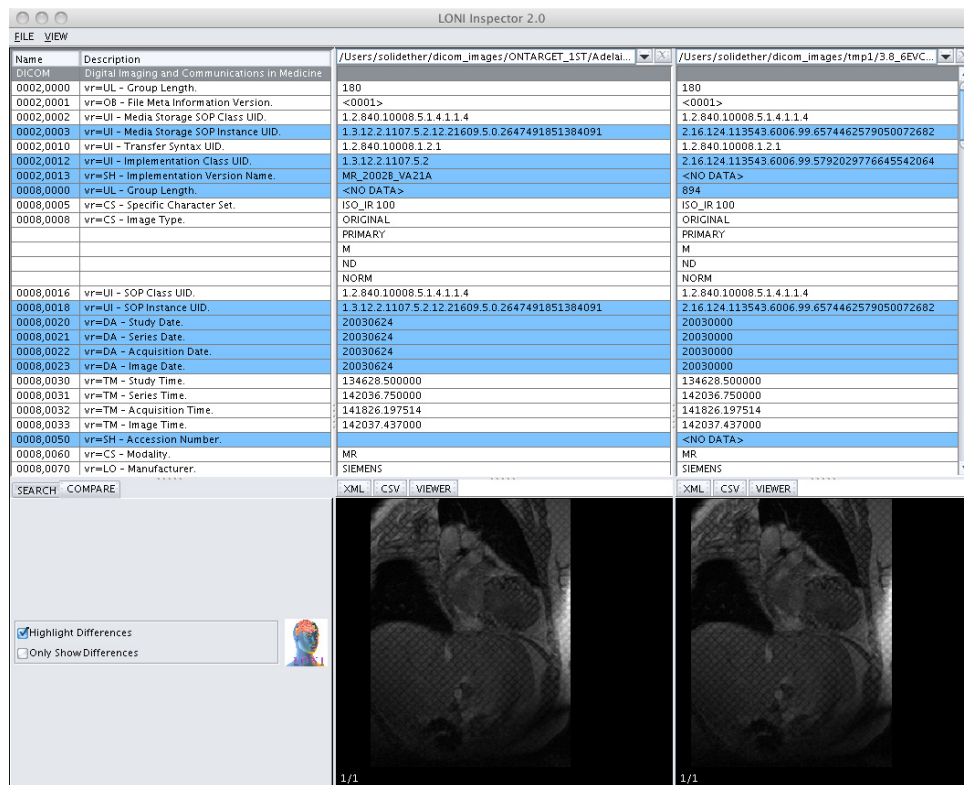


Figure 2: The LONI Inspector comparing a de-identified image with the original source, showing highlighted differences between the image data

Using LONI Inspector

Once the data are de-identified, they can be compared with the original data using the *LONI Inspector*⁷. The Inspector provides a single interface for viewing both metadata and image data in many common medical image file formats (AFNI, ANALYZE, DICOM, ECAT, GE, Interfile, MINC, NIFTI), searching files for keywords, comparing files for differences, and exporting metadata into XML and CSV files. The Inspector has been tested on and successfully run on Windows, Linux, and OS X platforms. Files can be loaded automatically into the Inspector by specifying them as command line arguments:

```
java -jar -Xmx1024m Inspector2_22Jun2007.jar <file1> <file2> <file3> ...
```

After execution, the Inspector will present an interface from where an arbitrary number of images can be imported using the menu option "File -> Add/Remove Files". After loading the images, the Inspector will list tag names, value range descriptions and the images specific values in a table form. If more than one image was opened, the Inspector will automatically display the values of two images side by side, which intuitively allows comparing the tag values.

⁷ <http://www.loni.usc.edu/Software/LONI-Inspector>

Further, the COMPARE tab besides the SEARCH tab allows to “*Highlight Differences*” or “*Only Show Differences*” in the selected files.

Using Comparison Script

CAP has created a simple Python-based script, called DicomDiff⁸, to compare the de-identification results with the original DICOM images. This script requires GDCM library⁹ to dump DICOM headers. Comparing a de-identified DICOM image (processed using the methods developed for the CAP) with the original image produces the following result (note that values containing [xxx] have been blanked):

```
=====
MISSING (17)
=====
0002,0013 : [xxx] : Implementation Version Name
0008,0050 : (no value) : Accession Number
0008,0081 : [xxx] : Institution Address
0008,0090 : (no value) : Referring Physician's Name
0008,1010 : [xxx] : Station Name
0008,1070 : [xxx] : Operators' Name
0010,4000 : [xxx] : Patient Comments
0018,1200 : [xxx] : Date of Last Calibration
0028,1055 : [xxx] : Window Center & Width Explanation
0029,0010 : [xxx] : Private Creator
0029,0011 : [xxx] : Private Creator
0029,1009 : [xxx] : CSA Image Header Version
0029,1019 : [xxx] : CSA Series Header Version
0029,1131 : [xxx] : PMTF Information 1
0032,1060 : [xxx] : Requested Procedure Description
0040,0244 : [xxx] : Performed Procedure Step Start Date
0040,0253 : [xxx] : Performed Procedure Step ID
=====
DIFFERENT (15)
=====
0002,0003 : xxx : Media Storage SOP Instance UID
==> [2.16.124.113543.6006.99.114451580699916 <==
0002,0012 : [xxx] : Implementation Class UID
==> [2.16.124.113543.6006.99.579202977664554 <==
0008,0018 : xxx : SOP Instance UID
==> [2.16.124.113543.6006.99.114451580699916 <==
0008,0020 : xxx : Study Date
==> [20030000] <==
0008,0021 : xxx : Series Date
==> [20030000] <==
0008,0022 : xxx : Acquisition Date
==> [20030000] <==
0008,0023 : xxx : Content Date
==> [20030000] <==
0010,0010 : xxx : Patient's Name
==> (no value) <==
0010,0020 : xxx : Patient ID
==> (no value) <==
```

⁸ <http://cardiacatlas.svn.sourceforge.net/viewvc/cardiacatlas/trunk/DicomDiff/>

⁹ <http://sourceforge.net/projects/gdcm/>

```

0010,0030 :                xxx      : Patient's Birth Date
==>                [19370000] <==
0018,1000 :                xxx      : Device Serial Number
==>                [4zmCh4K.0Uw2] <==
0020,000d :                xxx      : Study Instance UID
==> [2.16.124.113543.6006.99.367216069017295 <==
0020,000e :                xxx      : Series Instance UID
==> [2.16.124.113543.6006.99.634714607537772 <==
0020,0052 :                xxx      : Frame of Reference UID
==> [2.16.124.113543.6006.99.026625471348764 <==
0088,0140 :                xxx      : Storage Media File-set UID
==> [2.16.124.113543.6006.99.327761635876218 <==
=====
      ADDED (10)
=====
0008,0000 :                884 : Generic Group Length
0010,0000 :                66 : Generic Group Length
0018,0000 :                452 : Generic Group Length
0020,0000 :                332 : Generic Group Length
0028,0000 :                162 : Generic Group Length
0029,0000 :               34974 : Generic Group Length
0040,0000 :                22 : Generic Group Length
0088,0000 :                52 : Generic Group Length
5200,0000 :                20 : Generic Group Length
5200,9230 :      (Sequence with defined length) : Per-frame Functional Groups Sequence
=====
      SAME (85)
=====
0002,0000 :                180 : File Meta Information Group Length
0002,0001 :                00\01 : File Meta Information Version
0002,0002 :      [1.2.840.10008.5.1.4.1.1.4] : Media Storage SOP Class UID
0002,0010 :      [1.2.840.10008.1.2.1] : Transfer Syntax UID
0008,0005 :      [ISO_IR 100] : Specific Character Set
0008,0008 :      [ORIGINAL\PRIMARY\M\ND ] : Image Type
0008,0016 :      [1.2.840.10008.5.1.4.1.1.4] : SOP Class UID
0008,0030 :      [134628.500000 ] : Study Time
0008,0031 :      [140206.625000 ] : Series Time
0008,0032 :      [140010.149989 ] : Acquisition Time
0008,0033 :      [140206.859000 ] : Content Time
0008,0060 :      [MR] : Modality
0008,0070 :      [SIEMENS ] : Manufacturer
0008,0080 :      [ADELAIDE CARDIAC IMAGING] : Institution Name
0008,1030 :      [PERRETT - HEART^Routine ] : Study Description
0008,103e :      [cine_short_axis ] : Series Description
0008,1090 :      [Sonata] : Manufacturer's Model Name
0008,1111 : (Sequence with defined length):Referenced Performed Procedure Step Sequence
0008,1140 :      (Sequence with defined length) : Referenced Image Sequence
0010,0040 :      [F ] : Patient's Sex
0010,1010 :      [066Y] : Patient's Age
0010,1030 :      [120 ] : Patient's Weight
0018,0020 :      [SE] : Scanning Sequence
0018,0021 :      [SK\OSP] : Sequence Variant
0018,0022 :      [CT] : Scan Options
0018,0023 :      [2D] : MR Acquisition Type
0018,0024 :      [tfi2d1_20 ] : Sequence Name
0018,0025 :      [N ] : Angio Flag
0018,0050 :      [6 ] : Slice Thickness
0018,0080 :      [31] : Repetition Time
0018,0081 :      [1.55] : Echo Time
0018,0083 :      [1 ] : Number of Averages

```

```

0018,0084 : [63.649481 ] : Imaging Frequency
0018,0085 : [1H] : Imaged Nucleus
0018,0086 : [0 ] : Echo Number(s)
0018,0087 : [1.494 ] : Magnetic Field Strength
0018,0089 : [140 ] : Number of Phase Encoding Steps
0018,0091 : [1 ] : Echo Train Length
0018,0093 : [60] : Percent Sampling
0018,0094 : [81.25 ] : Percent Phase Field of View
0018,0095 : [930 ] : Pixel Bandwidth
0018,1020 : [syngo MR 2002B 4VA21A ] : Software Version(s)
0018,1030 : [cine_short_axis ] : Protocol Name
0018,1060 : [93] : Trigger Time
0018,1062 : [867 ] : Nominal Interval
0018,1090 : [22] : Cardiac Number of Images
0018,1201 : [155147.000000 ] : Time of Last Calibration
0018,1251 : [Body] : Transmit Coil Name
0018,1310 : 0\256\125\0 : Acquisition Matrix
0018,1312 : [ROW ] : In-plane Phase Encoding Direction
0018,1314 : [65] : Flip Angle
0018,1315 : [N ] : Variable Flip Angle Flag
0018,1316 : [1.3025874 ] : SAR
0018,1318 : [0 ] : dB/dt
0018,5100 : [HFS ] : Patient Position
0020,0010 : [1 ] : Study ID
0020,0011 : [8 ] : Series Number
0020,0012 : [1 ] : Acquisition Number
0020,0013 : [3 ] : Instance Number
0020,0032 : [-11.96977\ -198.92244\130.72878] : Image Position (Patient)
0020,0037 : [0.75927131\0.65077421\3.0784194e-009\ -0 : Image Orientation (Patient)
0020,1040 : (no value) : Position Reference Indicator
0020,1041 : [-78 ] : Slice Location
0028,0002 : 1 : Samples per Pixel
0028,0004 : [MONOCHROME2 ] : Photometric Interpretation
0028,0010 : 256 : Rows
0028,0011 : 208 : Columns
0028,0030 : [1.3671875\1.3671875 ] : Pixel Spacing
0028,0100 : 16 : Bits Allocated
0028,0101 : 12 : Bits Stored
0028,0102 : 11 : High Bit
0028,0103 : 0 : Pixel Representation
0028,0106 : 1 : Smallest Image Pixel Value
0028,0107 : 563 : Largest Image Pixel Value
0028,1050 : [255 ] : Window Center
0028,1051 : [619 ] : Window Width
0029,1008 : [IMAGE NUM 4 ] : CSA Image Header Type
0029,1010 : 53\56\31\30\04\03\02\01\29\00\00\00\4d\0 : CSA Image Header Info
0029,1018 : [MR] : CSA Series Header Type
0029,1020 : 53\56\31\30\04\03\02\01\27\00\00\00\4d\0 : CSA Series Header Info
0029,1132 : 106496 : PMTF Information 2
0029,1133 : 0 : PMTF Information 3
0029,1134 : [DB TO DICOM ] : PMTF Information 4
0040,0245 : [134628.500000 ] : Performed Procedure Step Start Time
7fe0,0010 : 00\00\00\00\00\00\00\00\00\00\00\00\00\00\0 : Pixel Data

```

```

=====
Tags in File A : 117
=====

```

```

=====
Tags in File B : 110
=====

```