Medical data management working group

Johan Montagnat

October 13, 2005

1 Purpose of this document

This is an internal document of the EGEE project[1]. The intent is to set up a working group on a grid service for medical data management. It describes the medical data format and the expected functionalities for such a service. It aims at explaining the constraints related to medical data and servers and organizing the development of the service.

2 DICOM: Digital Image and COmmunication in Medicine

DICOM[2] is the most established standard for medical data management. The DICOM standard covers **both** a medical image data format **and** an image communication protocol. The DICOM standard was thought and developed for answering needs of medical practice: mainly transferring images between image acquisition devices and visualization/post-processing consoles. It was not designed for image archiving nor image processing needs.

2.1 DICOM data format

A DICOM image usually contains one slice (a 2D image) acquired using any medical imaging modality (MRI, CT-scan, PET, SPECT, ultrasound, X-ray...). A DICOM image may contain a multi-slice data set but this is rarely encountered. A DICOM image contains both the image data itself and a set of metadata related to the image, the patient, the acquisition parameters and the radiology department. DICOM metadata are stored in fields. Each field is identified by a unique tag defined in the DICOM standard. A given field may be present or absent depending on the imager that produced the image. The standard is open and image device manufacturers tend to use their own fields for various information. A couple of fields (such as image size) are mandatory but experience proved that surprised should be expected. The image itself is usually stored as raw data. Most imaging devices produce an intensity value per image pixel, coded in a 12 bits format. Other format may be encountered such as 16 bits data or lossless JPEG.

[1] http://www.eu-egee.org/
[2] http://www.nema.org/

2.2 DICOM protocol

Most (reasonably modern) medical image acquisition device are DICOM clients. DICOM servers are computers with on-disk and/or tape back-ends able to store and retrieve DICOM images. The DICOM protocol defines the communication protocol between DICOM servers and clients.

2.3 DICOM storage

There is no standardization on DICOM storage. DICOM servers are implementing their own policy of data storage. One should not see DICOM datasets as a set of files. As stated above, a single DICOM image usually contains only one image slice. In practice, during a medical examination (a *study* in the medical world), a radiologist acquires several 3D images, representing up to hundreds to thousands of slices. A study is composed by one or several *examination*. An *examination* is composed by one or several slices. Their are 3 DICOM fields in each slice identifying the study, exam, and slice number. Note that their is no notion of 3D image: an exam may contain a set of slices composing several 3D images. The way a DICOM server stores these data sets on disk is irrelevant just like the way a database stores its table is usually not known: the medical user is never exposed to the DICOM storage and does not need to know if different files are used for each DICOM slice, exam, etc. Metadata that are included in DICOM image headers are difficult to manipulate. A DICOM server will often extract these metadata and store it in a database to ease data search.

2.4 DICOM security

The DICOM security model is rather weak. Security is often implemented in hospitals by isolating the images network from the outside world. DICOM files are unencrypted and transported unencrypted. Files contain patient data. The DICOM server security model is based on a per-application basis: all users having access to some DICOM client application can access to the information that the server returns to this specific application. DICOM servers are using random file names without any connection to the patient information and a proprietary data storage policy.

3 Medical data usage

3.1 Medical image sources

Each image acquisition device is a potential DICOM compliant medical image source. In a radiological department, one or several DICOM servers can be set up to centralize data acquired in this site. Medical data are naturally distributed over the different acquisition sites.

3.2 Access to medical images

In clinical practice, physicians do not access directly to image files. They identify data by associated metadata such as patient name, acquisition date, radiologist name, etc. The data are transferred mainly for visualization purposes. The physician quickly scans the slices stack in the DICOM study and focuses on the slices he is interested in.

In the medical image analysis community, the needs are quite different. One often need to identify images through metadata too, although the search are not

necessarily for nominative data but often related to the acquisition type or body area. 3D images are exported to disk files for post-processing and ease of use. Various 3D medical image format may be used to stack different DICOM slices into a single image volume (the most common being the *analyze* file format).

3.3 Medical data and security

All medical data should be considered as sensitive to preserve patient privacy. Nominative medical data are of course the most critical data and therefore, no binding between nominative data and images should be possible for non accredited users. In clinical practice, this result is often obtained by isolation of the image network. Only physicians taking part of one patient healthcare should have access to the data of this patient.

On a grid, the distribution of data make security problem very sensitive. To ensure patient privacy, the header of all DICOM images sent by a DICOM server should be wiped out, at least partially, to ensure anonymity. All images that are stored out of the source center should be encrypted to ensure that non accredited users cannot read the image content.

4 Medical data management service

Our objective is to develop a medical data management grid service able to:

- Register data produce by DICOM sources in the grid DMS.
- Register associated metadata.
- Retrieve medical images on UI/WN. Image search can be LFN-based or metadata-based.
- Register processed images and log processing traces.
- Offer an abstraction layer to expose the user to 3D images rather than DICOM slices.
- Preserve data privacy.

The system includes an interface to DICOM, a metadata management tool, a SRM interface for grid compliance, and grid data catalogs. A functional overview of the system is given in figure 1. Ideally, the metadata are distributed over the various acquisition sites.

4.1 Expected functionalities

The Medical Data Manager (MDM) exposes an SRM interface to the grid (and therefore can be seen as any SE for compatibility). As input files should only coming from inside the hospital, the SRM provides read-only file access to the DICOM server. Files written to the disk should be stored on a secondary storage but should not alter the hospital DICOM server.

In addition, the MDM interface receives DICOM push requests directly from imagers or from the DICOM servers and registers the files in the grid data catalogs (GUID and LFN creation, file registration).

The MDM interface returns LFNs or images corresponding to a selection based on metadata search.

The MDM implements an abstraction layer able to analyze set of DICOM slices to determine 3D image from DICOM slices.

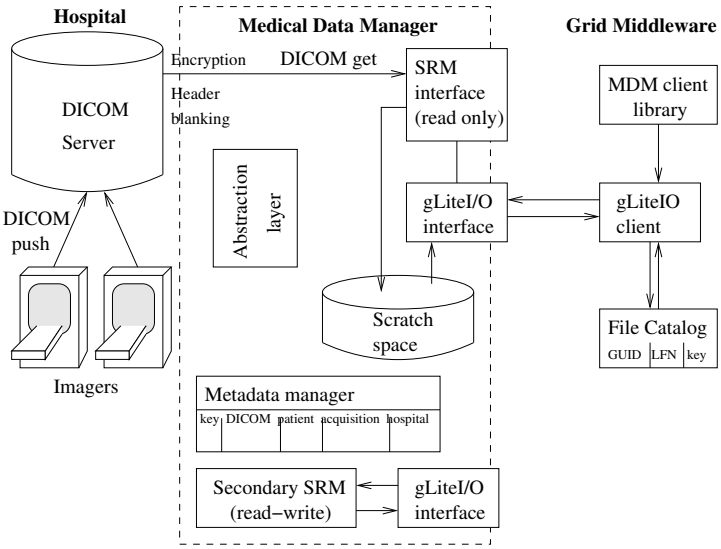

Figure 1: Overview of the medical data manager

Image files sent outside are encrypted. Request for metadata are controlled at an individual level.

Optionally, the MDM ensures traceability by logging image file transformations.

4.2 First step simplifications

For simplification, the MDM is split in two SRMs: a read-only DICOM-SRM and a secondary classical SRM used for writing files coming from the grid. Files are written in the DICOM-SRM from hospital imagers only. A trigger mechanism ensures that incoming files are properly registered.

The abstraction layer is delayed: a grid file will be registered for each incoming DICOM slice. Later, slices will be grouped in 3D images and a single grid file per 3D image will be registered. For each incoming slice, a SURL will be created from the DICOM SOPInsUID, SERInsUID, STUInsUID fields (and possibly the slice number).

Files sent outside the DICOM server will be converted into a 3D file format (such as INR), thus removing all metadata from the header.

4.3 Flow diagrams

The two main missing components are the triggering mechanism (registering a new DICOM image into the server) and the MDM library (providing read access to the server on the application side). They are illustrated in figures 2 and 3 respectively. Daniel is in charge of the trigger and Christophe is in charge of the MDM client library. Modifications will be needed on the SRM-DICOM side to ensure file format conversion and encryption. Encryption components are identified in blue in figure 3 and may be delayed.

5 Medical metadata schema

The following metadata schema may be implemented. It will probably needs refinement in the future. The imageid might probably be the GUID.

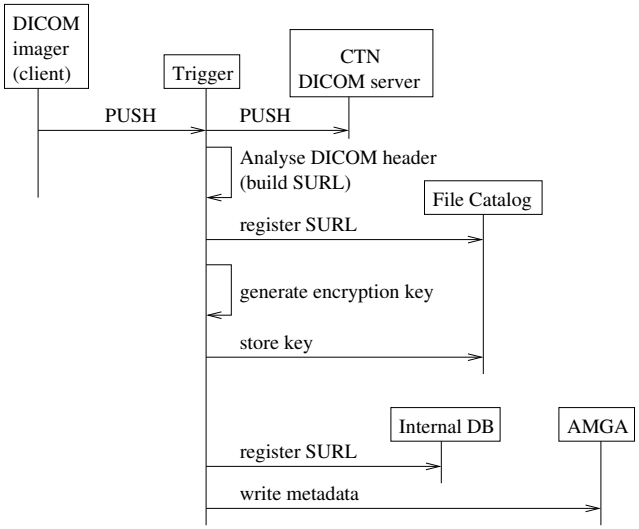

Figure 2: Triggered action at image creation

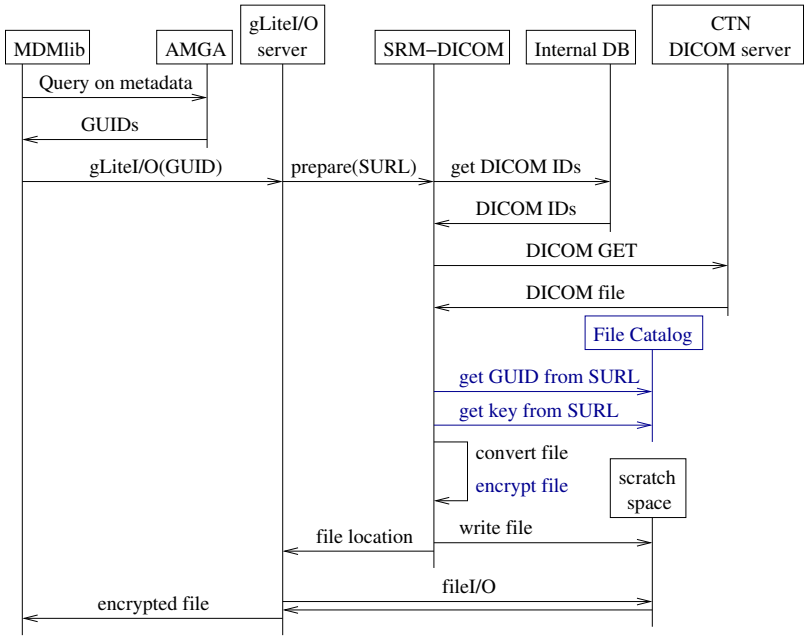

Figure 3: Accessing DICOM images

```
CREATE TABLE Image (imageid VARCHAR(64) NOT NULL, patientid
VARCHAR(64), kind
ENUM('DICOM','INR','MINC','ANALYZE','JPG','PNG','TIFF','GIF','OTHER'),
server VARCHAR(48), protocol VARCHAR(128), nx INT UNSIGNED, ny INT
UNSIGNED, nz INT UNSIGNED, nt INT UNSIGNED, sx FLOAT, sy FLOAT, sz
FLOAT, st FLOAT, vdim INT UNSIGNED, type ENUM('8bits', 'u8bits',
'16bits', 'u16bits', '32bits', 'u32bits', 'float', 'double', 'other'),
filesize INT UNSIGNED, PRIMARY KEY(imageid))

CREATE TABLE Protocol (imageid VARCHAR(64) NOT NULL, tablename
VARCHAR(255))

CREATE TABLE Medical (imageid VARCHAR(64) NOT NULL, modality
ENUM('MR','US','CT','PET','SPECT','US','other'), region
ENUM('head','thorax','abdomen','legs','arms','other'), hospital
VARCHAR(128), departement VARCHAR(128), radiologist VARCHAR(128),
acqdate DATE, acqtime TIME, imager VARCHAR(128), parameters
VARCHAR(255), annotation BLOB, PRIMARY KEY(imageid))

CREATE TABLE Patient (patientid VARCHAR(64) NOT NULL, name
VARCHAR(128), sexe ENUM('M', 'F'), dob DATE, PRIMARY KEY(patientid))

CREATE TABLE DICOM (imageid VARCHAR(64) NOT NULL, SOPInsUID
VARCHAR(64), SERInsUID VARCHAR(64), STUInsUID VARCHAR(64), seqnb INT
UNSIGNED)

CREATE TABLE NonDICOM (imageid VARCHAR(64) NOT NULL, filename
VARCHAR(255))

CREATE TABLE Log (imageid VARCHAR(64) NOT NULL, source VARCHAR(64),
algorithm VARCHAR(64))
```

The DICOM (resp. NonDICOM) tables associate DICOM slices (resp. files) with a GUID. The Image table contains image-related metadata. The Patient table contains patient-related metadata. Each Patient entry is NOT directly related to a file. The Medical table contains acquisition-related metadata. The Protocol table is an association between each image and a set of related table. It is intended to offer an extensibility for parameters that are specific to a given medical acquisition protocol (*e.g.* for all images acquired with the MRI turbo-spin sequence, provide additional metadata related to this modality). The Log table intends to enable traceability. As the ENUM data type does not exist in AMGA, it will be replaced by INT.