Quality of Life through Quality of Information
J. Mantas et al. (Eds.)
IOS Press, 2012

doi:10.3233/978-1-61499-101-4-539

A DICOM Architecture for Clinicians and Researchers

Michael ONKEN[a,1]
[a] *OFFIS – Institute for Information Technology*

Abstract. Over the last years there has been a strong trend of publishing health data in anonymized format in order to make it available for research. This is also true for medical imaging where the DICOM standard is the predominant data format and network protocol. This paper proposes an extension to any DICOM networking infrastructure that permits sharing of medical images in an anonymized way. Standard DICOM software is utilized on client and server side. While offering researchers access to all images in anonymous format, the architecture enables authorized clinicians to access the same images including their original patient information (name, institution, etc.). Identifying parts and anonymous parts of the image data are stored to geologically different databases. Together with sophisticated network protocols, patient privacy is fully preserved.

Keywords. DICOM, research, anonymization, pseudonymization, re-identification

Introduction

Data sharing for research has become a major trend over the past years. In the health care environment, there are many projects where real world health data is anonymized and shared for medical research purposes. Medical standardization bodies like DICOM and IHE started working on this topic, too (e.g. [1]).

Sharing medical data for research offers several advantages for researchers, clinicians and patients: Since medical research often requires large amounts of data, collected research data is usually not sufficient if only taken from a single site. This is especially true in the area of rare diseases. Furthermore, research collections of medical data offer great help for teaching purposes, permitting students to learn from real world data. One example for an active project that works within this area is the NBIA (National Biomedical Imaging Archive) belonging to the National Cancer Institute which has acquired a huge amount of more than 30 million medical images by 2011 [1].

Not only future patients but also current patients can benefit from findings in research data, if their identity can be determined from research data in a privacy-preserving, legally permitted way. This is usually achieved by pseudonymizing the data after anonymization, i.e. inserting a pseudonym into the de-identified data in order to mark images as being linked to a single (anonymous) person. If the mapping between real patient identities and their pseudonyms is maintained, anonymization and pseudonymization can be reversed, a procedure sometimes referred to as "re-identification".

[1] Corresponding Author. Michael Onken, Escherweg 2, 26121 Oldenburg, Germany, onken@offis.de

In order to permit seamless integration of sharing and accessing medical images into the clinical workflow, physician's should be able to use their every-day tools. In the medical imaging domain this adds up to the use of standard DICOM software for publishing and accessing medical images, on client and server side. This software is already installed and used at all sites that need to work with medical images.

As a summary, an architecture for sharing and accessing medical images is required permitting researchers to access anonymized (and pseudonymized) data, providing the possibility to re-identify patients, letting clinicians see their patients in a fully re-identified way, utilizing existing DICOM software on client and server side, and preserving privacy of patients in a secure and legally permitted way. Since there is no solution so far that fulfills these requirements, this paper proposes a dedicated architecture.

1. Methods

Some of the identified problems have already been covered by researchers in depth: Anonymization and pseudonymization has been widely discussed in literature, also for medical imaging. For anonymization, several approaches like k-anonymity [2] (and successors) and others have been proposed. DICOM has its own "practical" approach introduced with Supplement 142 [3], which can be combined with general methods like k-anonymity. Thus, the anonymization problem is mostly solved since there are tools available, e.g. from the author's research group [4], and huge projects like the NBIA that are already performing well. The challenge of generating robust pseudonyms has also been addressed in different projects, including a tool provided by members of the TMF itself. Furthermore, storing anonymized images to the PACS and querying for them using DICOM protocols is mostly straightforward and can be done with off-the-shelf software. Existing solutions need some adaptations in order to make the methods work within a specific project or domain.

However, the missing piece for medical imaging projects is an architecture and data flow which enables standard DICOM clients and a standard DICOM server to be used for anonymized research sharing and access and for re-identified clinical access at the same time. In Germany, the first step to design such an architecture, as for any "public" medical research database, is to follow the rules of the TMF (Technologie- und Methodenplattform für die vernetzte medizinische Forschung e.V.). Those rules are mandatory in order to have the project accepted by legal authorities. TMF already offers two blueprint solutions for research database architectures that are combined in this project into a single one and then extended as necessary for medical imaging.

2. Results

2.1. Architecture

An architecture has been designed that ultimately facilitates and extends the proposals of the TMF for medical imaging. Figure 1 gives an overview of the intended data flow regarding a clinical (re-identified) access mode query which is the most sophisticated one. The architecture is built from a few components. First of all, there is a central DICOM PACS (Picture Archiving and Communication System) that holds anonymized

DICOM images, and there are one or more participating clinics that share or access those images. Further, there is a component that stores and manages IDATA, i.e. identifying patient information (the "Patient List") as well as the pseudonymization, and another one performing an additional encryption of the pseudonym (Pseudonymization Service). Medical images consist of identifying information (IDATA) that should be kept secret from non-authorized people, and medical data portions (MDATA) that can safely be published without privacy concerns (including the pixel data itself). IDATA is stored within the Patient List, MDATA within the Research Database. These four components and the IDATA and MDATA concept are roughly defined by the original TMF proposal. However, it is not defined how they could work with medical image data and protocols (DICOM). Also, the above architecture already combines two separate concepts of TMF which is necessary to ease clinical access mode.

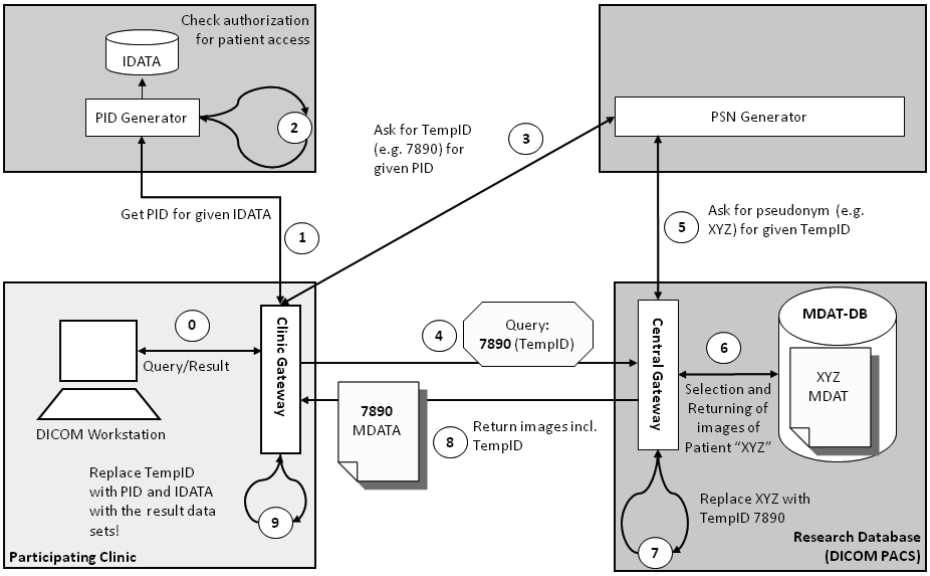

Figure 1. Proposed architecture showing data flow for clinicians (accessing fully identified data)

The architecture in figure 1 should ensure that IDATA is only revealed to authorized users (workflow shown is described below). Also, the Patient List operators should never be able to link IDATA to the corresponding medical images (MDATA). This is ensured by separating this component from the Pseudonymization Service. The Patient List only has a Patient Identifier (PID) for each patient, and the Pseudonymization Service uses that to compute a pseudonym (simply by using symmetric encryption) that is inserted into the medical data. At the same time, only Pseudonymization Service and Patient List can re-identify a medical image if they would collaborate. Thus both components have to be run by different, neutral parties.

In order to run the standard DICOM protocol from the workstation's and DICOM PACS' point of view, two components are added: a Clinic Gateway that resides in every participating clinic and a Central Gateway that is run by the same party running the DICOM PACS.

2.2. Data publishing and anonymous query

As stated, publishing data (anonymization, pseudonymization and DICOM storage of the resulting objects) have already been implemented by different projects. Within the above architecture, images may be sent from a Workstation or Modality via DICOM storage to the Clinic Gateway which performs the anonymization, sends the extracted IDATA to the Patient List and in exchange receives a PID (Pseudonym). The PID is sent to the Pseudonymization Service which returns a temporary ID (TempID) which then is inserted into the image instead of the PID, which is then sent to the Central Gateway. The Central Gateway asks the Pseudonymization Services for an encrypted pseudonym. The latter is finally inserted into the image that is then forwarded to the DICOM PACS. The procedure looks a little complicated, but offers important advantages: The research database never knows the PID of a patient or the IDATA itself. The pseudonymization service never sees patient data, either. Those systems cannot do any re-identification without collaboration with the other parties.
For a query in research mode, researchers can access the PACS from their DICOM Workstation by querying their Clinic Gateway which simply forwards the query to the Central Gateway, further to the DICOM PACS which then returns all results the same way back.

2.3. Clinical query and retrieve (re-identification)

The most difficult challenge is to enable query/retrieve in a re-identified way. Figure 1 illustrates the required data flow which is described within this section. The numbers refer to the steps in figure 1.

DICOM's Query/Retrieve Service permits searching for patient information like a patient's name, birth date or sex or for facts like study date and type of modality (CT, Ultrasound, etc). The problem is that the personal information is not available any more in the research database but was extracted as part of the IDATA. In order to query for such information anyway, the following can be done: A user selects the Clinic Gateway as query target, enters the query and sends it (0). The Gateway receives it and checks which user is in front of the system (see next section regarding authorization). Now, from the query all identifying information is extracted and sent with the credentials to the Patient List (1). The Patient List identifies the patient(s) that are covered by the user's authorization (2) and returns a PID for each. Each PID is exchanged by a TempID (3) by asking the Pseudonymization Service, and then inserted into the query. The query is transmitted (4) to the Central Gateway that inserts the Pseudonyms it gets (5) from the Pseudonymization Service in return for the TempIDs. The query is then forwarded (6) to the PACS, and the responses take the same way back, i.e. re-inserting (6) TempIDs for the pseudonyms, and at the Clinic Gateway re-inserting IDATA for the TempIDs (9). Finally, the results are returned to the workstation. Downloading images via DICOM Retrieve should work analogue to query.

2.4. Authorization

One problem is the authorization procedure of the person in front of the DICOM client system. DICOM optionally permits sending of user credentials like a password or Kerberos ticket. However, it is very rarely implemented and new systems should be avoided. Two other possibilities could be proposed: The user could type in login and

password into the query field usually used for entering the patient name, for example. This then could be checked by the Clinic Gateway/the Patient List and if login is successful, the user could be authorized for the next 30 minutes or until he enters a logout command (in the same way). Alternatively, one could use cryptographic one time pass codes which can be generated using devices in form of key rings and the like.

3. Discussion

The architecture described above could only be described in simple terms and lacks detail regarding DICOM and the surrounding TMF infrastructure. There are several challenges with above approach: For example, DICOM permits searching for wildcards, e.g. "A*" to find all patient names starting with the letter "A". However, it is viable to perform all the wildcard expansion on the Patient List and return the matching PIDs for further processing. Also there are more issues regarding the DICOM data and protocol specifications, e.g. how non-image data like DICOM Structured Reports can be anonymized. Future implementation of the system will show whether the remaining challenges could be solved and whether the system works as expected.

4. Conclusion

The proposed architecture allows easy sharing and accessing DICOM images for research and by offering re-identification of patients if required at the same time, in a legally permitted way. Privacy is ensured by utilizing TMF designs, adapted to work with standard DICOM clients and image archives in order to seamlessly integrate into clinical and research workflows. Parts of the system are already implemented in a software prototype. Completing this work and evaluating it in a real world scenario is planned for the next months.

References

[1] IHE's work on Quality, Research and Public Health: http://wiki.ihe.net/index.php?title=Quality,_Research_and_Public_Health (last access 9.5.2012)

[2] Freymann JB, Kirby JS, Perry JH, Clunie DA, Jaffe CC.Image data sharing for biomedical research--meeting HIPAA requirements for De-identification.J Digit Imaging. 2012;25(1):14-24. P. Samarati. Protecting respondents' identities in microdata release. In IEEE Transactions on Knowledge and Data Engineering. Volume 13 Issue 6. 2001

[3] NEMA Standards Publication, Digital Imaging and Communications in Medicine (DICOM) Supplement 142: Clinical Trial De-Identification Profiles, Final Text, National Electrical Manufacturers Association: Washington; 2009

[4] Onken M, Riesmeier J, Engel M, Yabanci A, Zabel B, Després S.Reversible anonymization of DICOM images using automatically generated policies.Stud Health Technol Inform. 2009;150:861-5.