Medical Informatics in a United and Healthy Europe
K.-P. Adlassnig et al. (Eds.)
IOS Press, 2009

doi:10.3233/978-1-60750-044-5-861

Reversible Anonymization of DICOM Images Using Automatically Generated Policies

Michael ONKEN [a,1], Jörg RIESMEIER [b], Marcel ENGEL [c], Adem YABANCI [c], Bernhard ZABEL [d], Stefan DESPRÉS [c,d]
[a] *OFFIS – Institute for Information Technology, Oldenburg, Germany*
[b] *ICSMED AG, Oldenburg, Germany*
[c] *M-SPEC GmbH, Mainz, Germany*
[d] *Center for Paediatrics Medicine, Freiburg University Hospital, Germany*

Abstract. Many real-world applications in the area of medical imaging like case study databases require separation of identifying (IDATA) and non-identifying (MDATA) data, specifically those offering Internet-based data access. These kinds of projects also must provide a role-based access system, controlling, how patient data must be organized and how it can be accessed. On DICOM image level, different image types support different kind of information, intermixing IDATA and MDATA in a single object. To separate them, it is possible to reversibly anonymize DICOM objects by substituting IDATA by a unique anonymous token. In case that later an authenticated user needs full access to an image, this token can be used for re-linking formerly separated IDATA and MDATA, thus resulting in a dynamically generated, exact copy of the original image. The approach described in this paper is based on the automatic generation of anonymization policies from the DICOM standard text, providing specific support for all kinds of DICOM images. The policies are executed by a newly developed framework based on the DICOM toolkit DCMTK and offer a reliable approach to reversible anonymization. The implementation is evaluated in a German BMBF-supported expert network in the area of skeletal dysplasias, SKELNET, but may generally be applicable to related projects, enormously improving quality and integrity of diagnostics in a field focused on images. It performs effectively and efficiently on real-world test images from the project and other kind of DICOM images.

Keywords. DICOM, reversible anonymization, SKELNET, rare diseases

1. Introduction

The SKELNET [1] project is an emerging German expert network that is designed to collect cases of skeletal dysplasia which are rare diseases including more than 350 different forms of genetically caused affections of bone development. Typically, medical imaging supports diagnosis, follow-up and clinical evaluation. Accordingly, the project provides – besides other services for data collection – a DICOM image archive (PACS). Images being stored to that PACS can be evaluated easily by the few specialists in the field; thus, the consultation process can be significantly shortened and

[1] Corresponding Author: Michael Onken, OFFIS – Institute for Information Technology, Escherweg 2, 26121 Oldenburg, Germany; E-mail: onken@offis.de.

formalized. Generally, DICOM images being sent to the SKELNET archive must pass through the Internet until they reach the SKELNET architecture. This requires encryption and early elimination of all identifying data (IDATA) resulting in an anonymous DICOM image still containing all relevant medical information (MDATA). Extracted IDATA and the remaining anonymized DICOM object are then marked by a "token" that allows for later re-assembling of the original object. After this process, herein called "reversible anonymization", the two DICOM data blocks are stored within two physically separated databases according to the data privacy rules in Germany. Clinically, authenticated SKELNET physicians potentially stating diagnoses will see patient-related data (including images) within a de-anonymized state. All decisions will, therefore, base on the re-assembled, original DICOM objects; all others can access anonymized DICOM images. In the described context, this paper presents an approach for the reversible anonymization on the DICOM object level.

2. Problem

DICOM images require special handling for physically separating IDATA and MDATA, because all this information is usually mixed-up in a single DICOM object. Thus, all IDATA must be extracted from the object with only MDATA remaining. Furthermore, for enabling full access to the original images as required in SKELNET, it must be possible to fully reverse the anonymization process by re-creating a semantically equivalent copy of the original image.

There are more than 70 different kinds of DICOM objects defined that can be stored to a PACS, ranging from images (e.g., CT, MR) to non-image information (like ECG). DICOM exactly specifies which information is required to be stored for a specific object type. This information is defined as a tree of so-called attributes, with each attribute (e.g., defining "Patient's Name" or "Modality" information) being of a prescribed DICOM data type and permitting a specific range of values. There are attributes that are mandatory for a particular image type, but only optional, conditional or even not permitted at all in others. As a consequence, DICOM objects (even of the same type) can differ substantially in the amount and type of information stored. Also, for a single DICOM object, different kinds of encodings are permitted and must be considered[2]. Furthermore, it is not feasible to just delete all known DICOM attributes that may contain IDATA because such an approach could result in invalid DICOM objects violating DICOM's attribute constraints described above.

Reversing anonymization of the archive's DICOM objects mainly requires the correct re-insertion of IDATA formerly extracted. Due to the possibilities offered by the DICOM Query/Retrieve protocol, it must also be considered that the encoding of the object may have changed intermittently.

3. Materials and Methods

The DICOM Standard [2] comprehensively describes on over 1,000 pages which attributes are mandatory and optional for each object type and the attribute values being

[2] A single DICOM object may be encoded varying in compression, storage of multi-byte value (endianness), data type encoding, etc. without changing semantically.

permitted. There are two general anonymization approaches: Either a *negative list* being common for all image types could be used, specifying which of the 2,500 existing DICOM attributes may contain IDATA information and therefore must be cleaned or deleted (respecting the constraints of the corresponding DICOM object type), or a *positive list* could be considered, listing for each object type which attributes may be safely taken over (the MDATA) into the anonymized object and which data (the IDATA) has to be removed or cleaned.

For the current project, a positive list[3] was chosen. This decision was mainly taken, because then the resulting, anonymized objects of a specific object type (e.g., all CT images) will contain a defined set of attributes which can be fully controlled by the anonymization rules and therefore, by the SKELNET project operators. This is an important advantage providing a consistently defined set of attributes ("feature set") for each image type in the archive to the SKELNET users. Furthermore, basic objects with a minimum set of attributes can be constructed during anonymization, enriched by a well-defined set of optional attributes. If a negative list was used for anonymization, it is only clear which attributes will not be part of the resulting object – but because of arbitrarily possible optional attributes in the original images (originating from different clinics), the anonymized objects would contain considerably different attributes[4]. However, it is a disadvantage of a simple positive list, that the de-identified images may not contain as much information as possible.

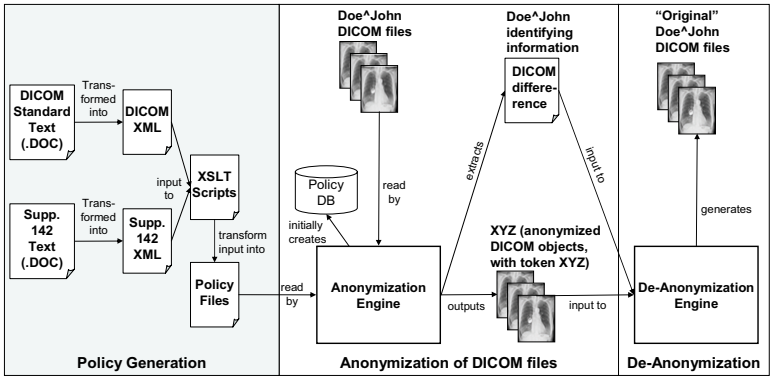

Figure 1. Process and data flow for reversible anonymization

Of course, building specific attribute lists (further called *policies*) for each object type is more challenging than providing a common huge list of attributes to be anonymized. This is because each attribute being required by the corresponding object type must be represented by a rule describing how to handle the attribute, e.g., keeping it unchanged, inserting a valid replacement value, or deleting it. Being tedious for one object type, for more than 70 object types a manual creation of such policies is not desirable. Therefore, an automatic generation of policies from the DICOM standard was put into effect. The complete process for anonymization and de-anonymization is shown in Figure 1 and is further described below.

[3] Of course, an existing positive list always could be converted into a negative list (and vice versa) by "inverting" it. Thus, the choice made more reflects how the problem was approached for implementation.

[4] This may also give attackers the chance to reveal by which device or clinic the image was created.

3.1. Policy Generation

So far, the DICOM standard is published in Microsoft Word and Adobe PDF format. Due to the hundreds of (partly nested) tables and quite informal textual descriptions that are distributed over different parts of the standard, it is extremely difficult to extract formal, machine-readable object descriptions (i.e., attribute lists) that are also necessary for generating anonymization policies. The approach taken was to generate an XML version[5] of the corresponding standard parts and to (partly manually) enhance it to meet the requirements of the project. Thus, for each DICOM object type, a machine-readable XML version was created, listing for each object type which attributes are mandatory, conditional and optional. These pre-requisites allow for selecting the attributes that are minimally needed for constructing a valid (anonymized) object. The next step is to consider which of the remaining attributes taken over into the anonymized object must be further investigated due to privacy concerns. DICOM Supplement 142 [3] is an extension to the standard currently being developed. It is mainly proposing a large list of attributes that may contain information sensitive to patient privacy. Besides a "Basic Profile", some optional Profiles are defined; e.g., the "Retain Longitudinal Option" mitigates the Basic Profile in keeping longitudinal (date/time information) attribute information. The core supplement text was also converted semi-automatically from Word to an XML format.

Based on these two XML inputs – DICOM standard and Supplement 142 – XSLT scripts were developed that automatically generate the required policy files, one for each type of DICOM object. Each policy file lists attribute rules for anonymizing a specific type of object, also permitting the use of specific Supplement Profiles by providing parameters to the XSLT scripts. Using a positive list approach, a policy file lists all attributes which should be taken over into the anonymized object as well as which attributes must be cleared or replaced by an appropriate value. All other attributes not explicitly listed are extracted (IDATA) and removed from the anonymized object.

The effectiveness of the anonymization firstly relies on the fact that only minimal attribute sets are taken over into the anonymized object. Secondly, the remaining attributes are de-identified as necessary based on the rules proposed by Supplement 142 which is being developed by Working Group 18 (Clinical Trials) of the DICOM Standard Committee. Therefore, the list of attributes and modifications being relevant for anonymization are based on the domain expert knowledge of Working Group 18.

3.2. Implementation

Based on the automatically generated policy files, a framework was implemented that reads those files into an SQLite database and anonymizes a given set of DICOM files. For later requests to "original" data it is necessary to also store any altered attribute information. This is realized by storing all those attributes in a DICOM-encoded "difference" file containing a binary copy of all attributes (and their position in the DICOM attribute tree) that are necessary for reversing the anonymization. For each patient, a unique token is inserted into the attribute *Patient ID* linking together all

[5] There is ongoing work for converting parts of the standard to a quite structured Docbook/XML format, but at this time there is no official XML-based version.

anonymized files that belong to the same patient[6]. There are some attributes that need special attention; e.g., the framework also inserts attributes shortly describing the de-identification approach. Also, all vendor-specific (so-called "private" attributes) are removed from the original DICOM object to assure that no "hidden" IDATA remains in the file. The anonymized DICOM file together with the difference file, both created by the framework, are sufficient for reversing the anonymization. Therefore, a quite simple tool was developed which "applies" the difference to the anonymized object, thus restoring the original object state. During that process, the actual encoding of the difference file and the anonymized object is harmonized in terms that the resulting object has a consistent DICOM encoding and is semantically equivalent to the original.

The framework described was implemented on top of the OFFIS DICOM Toolkit DCMTK [4] in C++ programming language. Currently, various Unix-like as well as 32-bit Windows operating systems are supported.

4. Results and Conclusions

The automatically generated pseudonymization policies and the corresponding framework have been tested on real test data originating from experts in the area of skeletal dysplasias and other sources. It turned out that the approach can be successfully used for efficient anonymization of DICOM images with effectively performing de-identification of DICOM objects by extracting identifying information and also re-applying that information to re-assemble the original DICOM data. By manually altering the policy files, fine-grained control over the anonymization process is provided to the SKELNET operators, allowing different anonymization policies for different clinics and even for different modalities. Further research could be done in that area, e.g., in the reversible anonymization of DICOM Structured Report (SR) documents which are currently not addressed by the presented approach. Also, it is expected that during SKELNET operation further challenges have to be met for unexpected (e.g., corrupted) data, burnt-in IDATA in the pixel data and so on.

Surely, the described approach can be used for similar projects that show a need for reversible anonymization of DICOM images and makes an important contribution to the quality and integrity of diagnostics in the field of medical images.

References

[1] Després, S., Engel, M.W., Zabel, B. (2007) Skeletal dysplasias. The network SKELNET. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz* 50(12):1548–1555.

[2] NEMA Standards Publication (2008) *Digital Imaging and Communications in Medicine (DICOM)*, National Electrical Manufacturers Association, Washington.

[3] NEMA Standards Publication (2008) *Digital Imaging and Communications in Medicine (DICOM) Supplement 142: Clinical Trial De-Identification Profiles*, Version 3, National Electrical Manufacturers Association, Washington.

[4] OFFIS – Institute for Information Technology. (2009) *DCMTK – DICOM Toolkit*. OFFIS, Oldenburg, http://dicom.offis.de/dcmtk.

[6] Issues like how the token is generated and kept consistently are beyond the scope of this paper.