International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:5, No:10, 2011

Automatic Real-Patient Medical Data De-Identification for Research Purposes

Petr Vcelak, Jana Kleckova
Department of Computer Science and Engineering,
University of West Bohemia, Univerzitni 8, Pilsen, Czech Republic
vcelak@kiv.zcu.cz, kleckova@kiv.zcu.cz

Abstract—Our Medicine-oriented research is based on a medical data set of real patients. It is a security problem to share patient private data with peoples other than clinician or hospital staff. We have to remove person identification information from medical data. The medical data without private data are available after a de-identification process for any research purposes. In this paper, we introduce an universal automatic rule-based de-identification application to do all this stuff on an heterogeneous medical data. A patient private identification is replaced by an unique identification number, even in burned-in annotation in pixel data. The identical identification is used for all patient medical data, so it keeps relationships in a data. Hospital can take an advantage of a research feedback based on results.

Keywords—DASTA, De-identification, DICOM, Health Level Seven, Medical data, OCR, Personal data

I. Introduction

Our medicine-oriented research is based on a collection of real patient data. Primarily, our research team cooperates with a neurological clinic on the cerebrovascular diseases research. The cerebrovascular diseases and strokes itself are one the most common cause of death worldwide. It is the second most frequent cause of death in the Czech Republic. Cancer incidence and mortality are two to three times greater in the Czech Republic than in other developed countries in Europe. [1], [2] Other departments plans cooperation, too.

Purpose of our research is to make any vital contributions to the solution of complex cerebrovascular brain diseases problem. Medical research data is fundamental of medical information system developing. It support any data processing, evaluation (e.g. evaluate volume of necrotic tissue of whole brain and its location) or making statistics. Medical data can be used as a practical education and diagnostic tool.

Data about a human subject are suitable for research purposes, but not its identity. A raw medical data refer to medical history, clinical trials, imaging examinations, reports, or any other documents which contain patients personal data. The raw medical data can contain national patient identifier, name, social security number, date of birth, addresses, contacts or any other unique identifications. It can contain medical doctor identification or name and accounting data for an insurance company, too. It is even a security problem to share the raw medical data with peoples outside a hospital – other than clinician or hospital staff. We have to remove any personal information from the raw medical data to solve this security risk. Medical data without personal data are available after a de-identification process at the hospital site. Only, the anonymous medical data are transferred (from hospital/source site) to the research site. Research site does not have any patient private information.

We introduce the AnonMed, an universal automatic rule-based de-identification application to do all of this stuff on an heterogeneous raw medical data. In fact, there is a set of de-identification applications, anonymizers or cleaners, but none of them can remove personal data from more file formats, keep data relations and does it automatically. Primarily, we need DASTA and DICOM format support. The DASTA standard format is used instead of HL7 [3] for clinical documents interchange between hospitals and any medical systems in the Czech Republic. No tool support DASTA file format. That is why we will make HL7 rules and implement its format support, later. Our solution was inspired by The Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy and Security Rules and [4].

II. Medical Data

Heterogeneous medical data are used in our research. It is a patient's clinical data set in a DASTA [5] format and imaging examinations in a DICOM format [6] of patients affected by a stroke at the University hospital in Pilsen. We are interconnecting clinical and imaging data [7] together in an experimental database we build [8].

A. DASTA

The DASTA is Data Standard abbreviation [5]. It is a national electronic communication standard format of a public health service in the Czech Republic. The DASTA is the Health Level Seven (HL7) standard equivalent and was developed by the Czech Public Health Informatics and Scientific Information Organisation supported by the Minister of Health of the Czech Republic. It is based on the XML-based mark-up standard with XSD schema. National Laboratory Code-List is a part of the DASTA standard.

The code-list is used while making an investigation order to a medical laboratory or getting a formalised laboratory result.

The DASTA standard version 03.01.01 is used in a hospital information system in the University hospital in Pilsen. This version supports only a DTD Schema, not the XSD Schema.

B. DICOM

The Digital Imaging and Communications in Medicine (DICOM) standard has been developed with an emphasis on diagnostic medical imaging. It is applicable to a wide range of image and non-image related information exchanged in clinical and other medical environments. This standard is widely used for representing and communicating radiology images scans and reporting. The DICOM standard is supported by the National Electrical Manufacturers Association. [6]

C. HL7

Health Level Seven (HL7) is a non-profit organisation involved in the development of international health care standards. It is widely used for data interchange between hospitals and physician record systems and between electronic medical record systems and practise management systems. [9]

The HL7 Clinical Document Architecture (CDA) documents are used to communicate documents such as physician notes and other material [3]. The CDA is an XML-based mark-up standard intended to specify the encoding, structure and semantics of clinical documents. The CDA document consists of a mandatory textual part and optional structured parts. The mandatory textual part ensure human interpretation of the document contents. The structured part relies on coding systems from SNOMED and LOINC to represent concepts and it is primarily designed for software processing.

The de-identification application bargain for the HL7 documents support. It is designed to process HL7 format. We need the collaboration clinical centre with medical data in this format before we can create final set of rules.

III. De-Identification

De-identification process treats with couple of difficulties. Medical data formats are heterogeneous and contains structured, unstructured and pixel data.

A. Anonymous Identification

An anonymous identification generator provides a unique identification number (UIN) to each patient. It is the most difficult part of de-identification process. An original patient identification is replaced by UIN. Any other patient personal data can be removed. The same UIN in all patient's records guarantee us to keep relationships in the research data, too. We have a comprehensive information about a human medical history without harming patient.

There is a couple of methods how to generate the UIN, e. g. one-way hash, zero-knowledge protocol, encryption with public and private keys or a hash map at the hospital site.

One-way hash method is vulnerable to dictionary attack. The Czech Republic national patient identifier system is based on date of birth and gender. It is nine or ten numbers length. The first six numbers are date of birth in a format $YYMMDD$ (year YY, month MM, day DD). Gender GG is appended as a sum of the month and a) $GG = 0$ for male, b) $GG = 50$ for female. Next three numbers UUU makes the number uniqueness – different for the same date and gender. A control number Q is the last figure. This Q makes the difference of sum of even figures and sum of odd figures commensurable by 11. For example, a male born on 6/20/1973 results in 730620 base and a female with the same date of birth has 735620 base. The control number Q was introduced on January 1^{st}, 1954.

A problem with one-way hash, like SHA1, is that a structured part for patient identifier can allow only number with nine to ten figures, not alphabet letters. We expect a new UIN in a medical research data, so we ignore the option that any medical application can check validity of original number (if it is commensurable by 11), but file have to be valid by DTD or XSD schema.

The zero-knowledge protocol is a cryptographic protocol. It determines if two records belong to the same person by solving a question without learning anything about the subjects in question. An original patient's identifier is added to a random number generator that create a new UIN. The generated UIN is same for the two records of the same patient. [4]

Now, we use a simple solution based on a combination of random number generator and a hash table. Both are hidden in the demilitarized zone at the hospital. When identification occurs, a hash table look-up is done with a recent personal identification as a table key. If found, UIN value corresponding to the personal identification from hash table is used. A new patient has no record in the hash table, so we generate new UIN and store it together with a patient personal identification in the hash table. UIN is unique for each hospital/source site, because the site code is part of the UIN. At the hospital site, a different UIN for the same patient can be a possible weak point of our UIN. De-identified data are transferred out of the hospital site – into the research site. On a research site, there is a too narrow opportunity to bind medical data to concrete patient. Results can be reused and hospital can take an advantage of a research feedback based on the hash map at the hospital site.

B. Operations

Operations we have supported in de-identification process are (sorted alphabetically):

- APPEND_AFTER appends any text after the existing value.
- APPEND_BEFORE appends any text before the existing value.

- CHANGE replaces text to a new predefined value from a profile file.
- EMPTY clears any existing value and makes it empty. Tag, element or attribute will exists in an output, but it will be empty.
- EXTERNAL_APPLICATION executes an external application with a file as an argument. By this way you can do anything else you need to do.
- IDENTIFICATION operation replaces personal identification value by UIN. Details are written in III-A.
- KEEP does nothing until you use a strict mode. There will be only attributes, elements and tags in the output that was mentioned in any of *EMPTY*, *CHANGE*, *IDENTIFICATION* or *KEEP* operation. This is working only when the strict mode is enabled and supported for a file format.
- REMOVE operation delete attribute, element or tag.
- SPECIFIC operation is file format specific. It can be used even for a configuration purposes of the file format de-identifier. This is a right way to specify e. g. removing private tags in a DICOM file.
- NONE does absolutely nothing. It can be used as a comment if you do not want to erase a rule. It is similar to the KEEP operation, but NONE operation has even no effect in the strict mode.

C. Structured Textual Data

Structured parts processing is a simple job, because you know structure and the meaning of structured parts. You can decide on removing element, erasing value, or keep a value based on DTD or XSD schema.

Both, DASTA and HL7 format are in a XML file format. We do this based on a XPath expression that makes a rule together with operation (III-B) keyword.

D. Unstructured Textual Data

DASTA and HL7 formats have an element for unstructured textual data. The element can contain anything what was written as a report by medical doctor. It is a weak point, that medical doctor can do wrong and write patient name, or any other personal information into it. Both formats can have an attachment with formatted version of that text. In DASTA files, we use from the clinical centre, it is a RTF file attachment in Base64. It contains the same text as in the XML element, but it is formatted.

We make a dictionary based on a file content before de-identification and then we try to find out any text or subtext occurrence. These texts are removed. The same operation is done in a formatted version.

E. Pixel Data

Imaging examinations are stored in the DICOM format, only. DICOM has a structured tags and the situation is same as in the subsection III-C. Problem occurs when an imaging machine burn in annotation which contains patient personal data, like name or birth date.

We try to solve this automatically by an artificial intelligence methods for text recognition in a case we know the text what we look for. The *EXTERNAL_APPLICATION* operation is used for this job. You can see the result of pixel data de-identification of the figure 1. It is based on an optical character recognition (OCR) method. We based OCR on well-known text what we are looking for in a pixel data, e.g. patient name, identification, birth date, operator.

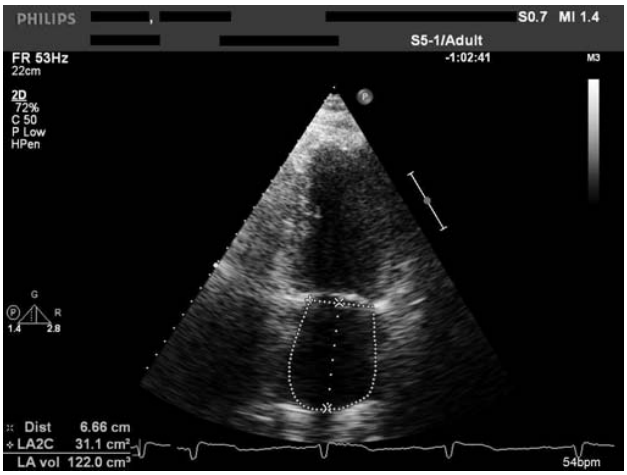

Figure 1. A DICOM file pixel data with automatically removed burned-in annotation with hospital, patient identification and staff data.

It is important when we find out DICOM file with burned in annotation, the output file is not in standard directory. It does not go to a production research database. It is used *uncertain* directory, instead. It is when the *(0028,0301)* tag has *YES* value. It is a problem when modalities does not record this tag into the file. That is why we have to process all files with not yet known equipment type.

F. Flow Chart

The de-identification application flow chart is on the figure 2. First, the application parse command line parameters and loads profile from a configuration file. Read input directory and list recursively all files in it. The input medical data processing can start at this time. A file type is determined for each file, not its extension. Data processing is chosen by the file type. All file type rules are applied on the file. The rule *IDENTIFICATION* of each file type uses unique identification number (UIN) generator (III-A). External application execution can be used by a rule with *EXTERNAL_APPLICATION* operation from any file type, too. The processing repeats until file list is empty. Each done rule with operation details is logged to a standard output and errors are logged to standard error output.

IV. Results and Tests

The application was written in the Java language with dcm4che library. It is open source and publicly available.

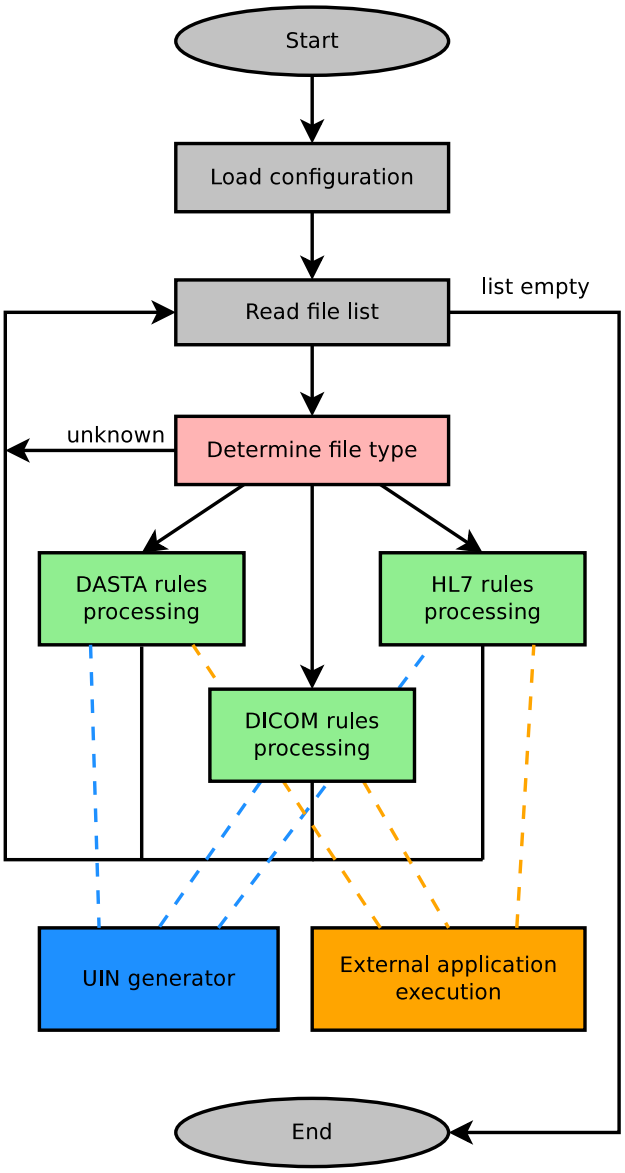

Figure 2. The de-identification AnonMed application flow chart.

It is primarily used on a de-identification server with GNU/Linux operating system, the stable Debian distribution (Squeeze).

Tests and benchmarking was done on the same machine, with CPU *Intel(R) Core(TM)2 6300 at 1.86 GHz*, RAM *1 GB DDR2*, chipset *Intel Corporation 82801H (ICH8 Family)* and SATA interface *Intel Corporation 82801H (ICH8 Family) 2 port SATA IDE Controller (rev 02)*.

The benchmark results are in the table I. The speed median is $8,479.70\ kB/s$. It is much faster than medical data transfer from the hospital information system through Samba server. On the other hand, we cannot compare the processing speed results with any other de-identification application because of missing DASTA support in other tools or batch mode for DICOM, e.g. DicomCleaner, DICOM support in the Oracle database.

We have already tested more than 400 DASTA and

Table I
De-Identification benchmark

Data amount [kB]	Time [s]	Speed [kB/s]
1,024,000	121.00	8,463
3,154,180	388.72	8,114
5,150,720	606.21	8,497
9,020,954	1,051.10	8,582
9,220,954	1,084.93	8,499
9,889,382	1,408.35	7,022
16,806,198	1,916.40	8,770
18,846,470	2,108.04	8,940
18,441,908	2,195.89	8,398
24,723,456	2,934.06	8,426

500,000 DICOM files of 360 patients. We do not have any HL7 file, because there is the DASTA format in the Czech Republic health care. DASTA files were de-identified with 100 % success. Computer tomography (CT) DICOM files was without a pixel data annotation. The rest of 1,000 DICOM files, e.g. diagnostic sonography (ultrasonography) or X-ray, were 85.4 % successfully de-identified, 7.3 % partially (but only the imaging date left in pixel data) and 7.3 % failed. Both, the failed and partially de-identified files were from the same equipment – the image was too noisy under the text or in the case of date had an unknown format.

V. Conclusion

In this paper, we present the AnonMed – an automatic rule-based de-identification application of heterogeneous medical data. The presented application fulfil requirements, provides expected results and processing speed. It allows medical data use for research purposes in our team. It is open source and public available at http://medical.kiv.zcu.cz/projects/anonmed/.

The AnonMed supports batch mode for automatic DASTA and DICOM de-identification. Strict mode adds opportunity to list elements or tags that are safe and can be present in the the output research data. Everything not listed in a keep list is removed. Medical data can be used as a practical education and diagnostic tool. More strict de-identification rules, like HIPPA, we plan when used for a wide-spread education purpose.

Optionally, the de-identification application is able to use a rule with an external application for a data processing. It is triggered by a specific value or any other condition you need. We use this rule e. g. when burned in annotation is indicated in a pixel data. It use optical character recognition (OCR) application for that image and black out a recognised area with any patient identifications (figure 1). It is in a production testing use. Results are in a special directory where we output an uncertainly de-identified data and all of them are checked before the transfer to the research site.

Finally, the success rate is almost 100 % in a whole data set. The subset with a burned-in annotation is rare, it is only 0.2 % of the whole medical data set we have. The success rate in this subset is 92.7 % when we ignore (and 85.4 % we do not ignore) examination date and time in images. In all these cases the recognition failed when in

the image was not implemented date and time format as it is used in pixel data.

In the future, there is a possibility to improve UIN (unique identification number) generator to assign the same UIN even for independent sites, e.g. [4]. We test de-identification of more modalities and equipment types medical data. National patient identifier system is based on an birth date and gender in the Czech Republic. National patient identifier system is the best candidate for unique code used in zero-knowledge protocol. The method we use is suitable until the research medical data are not public available and are used only by our research team.

Acknowledgment

The work presented in this paper was supported by the project Czech Science Foundation number 106/09/0740.

References

[1] CSU – Czech Statistical Office, "Český statistický úřad: Úmrtnostní tabulky (Death-rate Statistics)," Online, 2010-03-02. http://www.czso.cz/csu/redakce.nsf/i/umrtnostni_tabulky, 2010.

[2] MZCR – Ministry of Health of the Czech Republic, "Ministerstvo zdravotnictví České Republiky: Věstník č. 2/2010: Péče o pacienty s cerebrovaskulárním onemocněním v České republice," Online, 2010-03-01. http://legislativa.mzcr.cz/File.ashx?id=233&name=V%C4%9Bstn%C3%ADk_%20%C4%8D_02_2010.pdf, 2010.

[3] R. Dolin, L. Alschuler, S. Boyer, C. Beebe, F. Behlen, P. Biron, and A. Shabo Shvo, "HL7 clinical document architecture, release 2," *Journal of the American Medical Informatics Association*, vol. 13, no. 1, p. 30, 2006.

[4] Jules J. Bernman, "HHS Workshop on the HIPAA Privacy Rule's De-Identification Standard," HHS Workshop, March 8–9, 2010, Marriot at Metro Center, Washington, DC, March 8, 2010.

[5] Karlova univerzita v Praze – 2. lékařská fakulta v Praze (Charles University in Prague – 2nd Faculty of Medicine), "Data Standard (DASTA)," Online, 2011-03-02. http://dasta.lf2.cuni.cz/, 2011.

[6] National Institute of Neurological Disorders and Stroke, "Digital Imaging and Communications in Medicine (DICOM)," Online, 2010-03-02. http://medical.nema.org, Virginia, 2010.

[7] V. Rohan, P. Sevcik, J. Polivka, Z. Ambler, B. Kreuzberg, and J. Ferda, "Klinický pohled na výpočetní tomografii u akutní ischemie mozku (A clinical Approach to Computed Tomography in Acute Cerebral Ischemia)," *Česká a slovenská neurologie a neurochirurgie*, 2007.

[8] P. Vcelak, J. Kleckova, and V. Rohan, "Cerebrovascular diseases research based on heterogeneous medical data mining and knowledge base," in *2010 International Conference for Internet Technology and Secured Transactions (ICITST)*. London, United Kingdom: IEEE, Infonomics Society, 2010, pp. 345–350.

[9] Health Level Seven, Inc., "What is hl7?" Online, 2010-03-02. http://www.hl7.org/about/index.cfm, 2010.