IET Software

Special Issue: Advances in Knowledge and Information Software Management

IET Journals
The Institution of Engineering and Technology

Intelligent generator of big data medical imaging repositories

ISSN 1751-8806
Received on 31st July 2016
Revised 30th January 2017
Accepted on 4th February 2017
E-First on 25th May 2017
doi: 10.1049/iet-sen.2016.0191
www.ietdl.org

Tiago Marques Godinho[1] ✉, *Carlos Costa*[1], *José Luís Oliveira*[1]
[1]*DETI/IEETA, University of Aveiro, Aveiro, Portugal*
✉ *E-mail: tmgodinho@ua.pt*

Abstract: The production of medical imaging data has grown tremendously in the last decades. Nowadays, even small institutions produce a considerable amount of studies. Furthermore, the general trend in new imaging modalities is to produce more data per examination. As a result, the design and implementation of tomorrow's storage and communication systems must deal with big data issues. The research on technologies to cope with big data issues in large scale medical imaging environments is still in its early stages. This is mostly due to the difficulty of implementing and validating new technological approaches in real environments, without interfering with clinical practice. Therefore, it is crucial to create test bed environments for research purposes. This study proposes a methodology for creating simulated medical imaging repositories, based on the indexing of model datasets, extraction of patterns and modelling of study production. The system creates a model from a real-world repository's representative time window and expands it according to on-going research needs. In addition, the solution provides distinct approaches to reducing the size of the generated datasets. The proposed system has already been used by other research projects in validation processes that aim to assess the performance and scalability of developed systems.

1 Introduction

In the last decade, medical imaging studies have become one of the most important means of diagnosis [1]. The introduction of digital modalities has lowered the exploitation costs and increased the quality and usefulness of medical images [2]. For these reasons, nowadays, even small medical institutions are capable of producing large quantities of medical imaging studies [3].

Picture archive and communications systems (PACS) is the common designation of information systems that manage medical imaging data and associated workflows [4]. They encompass not only storage infrastructures that allow the storing of images for later use but also communication infrastructures, which support sharing and remote access to data across distinct institutions and applications. A typical PACS environment includes three major groups of applications: Image repositories, acquisition devices, and viewer applications [4]. To integrate these heterogeneous components, PACS rely on the DICOM Standard [5], which defines the format for storing medical imaging data and the network communication protocol.

Currently, the community is particularly interested in medical imaging content discovery for supporting clinical and technological research activities, including features extracted from pixel data (image) and metadata [6]. In fact, DICOM objects are composed of a metadata header along with the actual pixel data. The metadata section includes information related to imaging procedures such as patient identification, equipment acquisition parameters, diagnosis report or radiation dose exposure [6]. However, the heterogeneous nature of the DICOM metadata overburdens general purpose technologies. In DICOM, different modalities and clinical protocols produce different sets of metadata, making it difficult to use strict data models such as relational databases [7]. Furthermore, the massive volume of data generated by medical imaging procedures is itself a huge problem [8], which raises several big data related issues in their exploitation [9]. At the same time, there is also a considerable research effort in healthcare services federation [10] and Cloud outsourcing of medical imaging repositories. Managing communication between multiple geo-distributed locations is still challenging due to the huge volume of data that must be readily available for medical practice.

Most of the traditional technologies are not prepared to handle the requirements raised by those big data environments. Recent research efforts have been made with the adoption of non-conventional database technologies, such as document stores and free text indexes, in PACS [7, 11, 12, 13]. Concerning networks, several distributed architectures have been proposed, contemplating optimised communication processes, cache, and pre-fetch mechanisms to support those specific scenarios [3, 14].

These technologies promise to increase search performance and the flexibility of repository databases and reduce data access latency. However, they generally have poor validation processes, due to the lack of real-world data sets and workflows able to reproduce big data environments. Some of the reasons for this are: First, privacy reasons, which make it more difficult to perform tests using real-world data. In addition, tests could be hazardous since there is a considerable probability of jeopardising the regular operation of the healthcare institution. Another reason is that a particular institution dataset may not cover all the specificity required for validation purposes. Finally, it is crucial to analyse how new contributions will perform in the long-term.

This paper proposes and describes a new system able to generate big datasets using only representative portions of data extracted from real environments, causing no impact on production processes. The generated repositories are excellent elements for supporting research, development and validation of PACS technologies.

2 Background

The DICOM Standard [15] was introduced in 1993 by National Electrical Manufacturers Association, with the main purpose of normalising data structures and communications to promote interoperability between different PACS components. Before that, every manufacturer had its own protocol, which promoted vendor locking situations and poor exploitation of resources. Nowadays, it is a de facto standard, universally adopted by distinct medical imaging modalities and manufacturers. This standard normalises the content of medical imaging objects, how they should be formatted and broadcast among the different PACS applications [16]. For simplicity purposes, this article will not deal with the

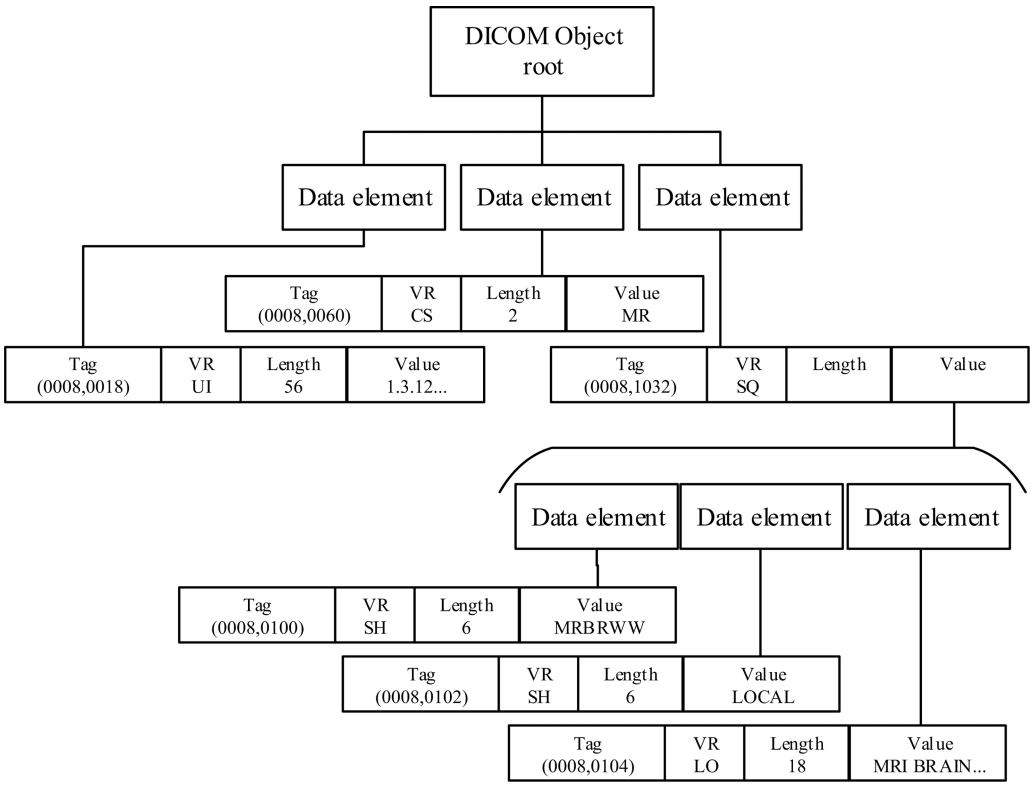

Fig. 1 *Example of a DICOM Object's structure*

communication part of the standard, since the proposed methods only demand knowledge of DICOM data structures.

A DICOM object aggregates metadata with images' pixel data [16]. The header contains meaningful information associated with the examination such as patient identification, equipment details, and radiation exposure information. These elements have a unique identifier [17] denominated as tag. For instance, the PatientName attribute is identified by the tag (0010,0010). These attributes are organised in semantic groups following the DICOM Information Model, a hierarchical database that captures real-world organisation: patient\study\series\images. A patient has multiple studies that may have multiple series that include one or more images. DICOM supports 27 data types; the attributes are matched to their data type using a dictionary [5], alternately, each attribute may declare its data type explicitly. Although the DICOM standard specifies a wide range of attributes, new ones can be declared via private dictionaries. Therefore, the standard is extensible and manufacturers can include their latest equipment metadata. Finally, a file object may contain one or more images (i.e. cine-loop) [18]. Fig. 1 provides an example of a DICOM file's metadata and its hierarchical organisation. It is perceivable that the root object holds multiple data elements. These elements can be composite elements holding further members in their hierarchy.

Regarding the problem stated in this article, the literature reports several studies that use state-of-the-art technologies for storage, search and retrieval of medical imaging data, but where the lack of representative datasets is evident in the validation process. For instance, Savaris *et al.* [7] propose a decomposed storage model, an approach similar to document-oriented databases, and present its comparison with the relational model. However, the dataset used in the trials has only 67 studies with a combined size of around 4 GB. In fact, this dataset is very small when compared to real medical imaging repositories, and does not validate the system's long-term performance. In [14], a regional PACS archive is proposed and the validation scenario includes two small diagnostic centres that handle an average of 3000 monthly examinations, with a combined volume of around 60 GB. In this scenario, a trial with a 4 GB dataset is not sufficient to analyse even a week of the PACS operation, nor the long-term scalability of the proposed system.

Multiple document-oriented databases for PACS archives have been put through a comparative laboratory trial [13]. The authors considered the performance of the databases in the short, medium and long-term by using three different datasets. However, their largest dataset might not be representative enough of the current scenario, since it contains only around 50 thousand images, approximately 4 times the daily CT production of our dataset C (see Section 4). More evidence of this issue is found in [11, 19, 20, 21].

3 Method

This section presents our methodology for generating big data medical imaging repositories to be used in controlled laboratory trials. This methodology was instantiated as a software system which orchestrates multiple components to (i) capture the production statistics of an existing repository over a period of time; (ii) create a representative regression model of the collected productivity statistics; (iii) and finally, generate datasets of DICOM files with similar productivity to the model, however possibly encompassing larger time periods. The main challenge is to capture as accurately as possible how productivity has evolved over time in the model repository. In this regard, we have also included a results section intended to assess this accuracy, and validate our method.

The Dicoogle PACS arises as a crucial component in our method for generating medical imaging datasets. It is an open-source PACS archive that offers a plugin-based architecture and a software development kit [12, 22]. It is intended to support the research and development of new PACS components and applications as plugins. These plugins may leverage the existing capabilities of Dicoogle, such as storage, indexing of DICOM metadata or even other functionalities. Therefore, it is possible to orchestrate services, creating complex workflows between multiple plugins, as demonstrated further. The components of the proposed system were developed as Dicoogle plugins, namely the DICOM Statistics Analyser and the DICOM Exporter.

Our method can be broken down into three stages: indexing, statistical analysis and exportation. Next, we provide a description of each stage, as well as the roles of the components presented above, as illustrated in Fig. 2.

To generate an accurate dataset, our method starts by indexing a real-world repository that will be used as a model. The study

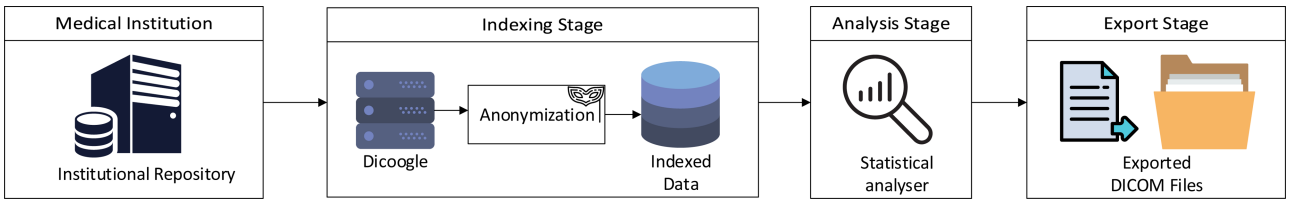

Fig. 2 *Architecture and information flow of the proposed methods*

patterns discovered in this model repository will be translated into the synthesised datasets. This indexing stage is a crucial task and requires full access to DICOM metadata of the model repository. Dicoogle's data interfaces and indexation capabilities support this task. It can import images from the institutional PACS repository using the DICOM query/retrieve services. Alternatively, Dicoogle can detect and automatically index DICOM files stored in the local file systems or network folders. This can be used to speed-up the indexation process, but it has more limited applicability. Once Dicoogle gains access to a DICOM image, it parses the metadata within the object and stores it in an indexing plugin. Although Dicoogle supports multiple indexing plugins, we have been using the Lucene plugin supplied with its distribution.

In this stage, Dicoogle also captures sensitive information related to patients and medical professionals such as names, identification numbers and birthdates. This information is not required by our method, and it may even constitute a security hazard, hindering the generalised usage of our method. As a result, this stage also includes an anonymisation mechanism that removes those elements to ensure data privacy. It involves the obfuscation of several fields that allow identification of the different subjects in the study, including PatientName, InstituionName, PhysicianNames etc. The fields that require anonymisation are configurable, allowing different levels of privacy. The algorithm is based on term replacement, where attributes' values are replaced by an anonymous counterpart. We used a one-to-one relation keeping the coherence of the indexed data. For instance, a given patient name will produce the same anonymised value, keeping the relations between the different image objects intact.

Our statistical analyser is responsible for generating a statistical model out of the model dataset. This model will enable the generation of more imaging studies with similar production characteristics. A model is generated for each medical imaging modality separately. Preliminary experiments show that this option produces more accurate results. Moreover, it also allows us to customise the exported datasets since other modalities can be included or excluded according to user preferences. For each modality, the study production is modelled using an autoregressive moving average model (ARMA) regression [23]. This regression analysis tool is well suited to time series data because it models the series' general underlying trend, as well as some extrinsic variation factor. In our particular case, this configuration allows the generation of a dataset that better represents the long-term production evolution. In fact, the inclusion of the ARMA regression model was a fundamental improvement in our method. Previously, our model was composed of simple statistical functions, which were unable to capture the trend in the study production series or other fluctuations. The quality of the model produced by the ARMA regression is heavily influenced by the timespan of the index collected from the real-world repository. Longer timespans allow for a more accurate model, which will perform better at generating datasets which extend the collected timespan. A more in-depth description of the ARMA regression is outside the scope of this work and can be found in [23].

As soon as statistics collection finishes, the DICOM Exporter module is able to start the process of generating new datasets. The synthesised dataset replicates the patterns discovered in the model repository but many options can be configured according to researchers' requirements. The volume of the generated dataset can be controlled by either directly specifying the dataset's time window in number of weeks, or indirectly by defining the maximum dataset size. The number of exported modalities can also be configured. For instance, it is possible to generate only CT studies even if the model dataset contains other modalities.

As expressed, the generated DICOM images have a similar content to the model dataset images. This is accomplished by using several models for each modality. The DICOM Export Plugin starts by selecting a model that matches the modality of the exported study, then replaces the content of several fields to create a unique image instance. Finally, the actual DICOM image file is written to the output storage medium.

The DICOM export plugin can operate in two modes regarding its behaviour with pixel data. It can use real images, collected from the model dataset. Despite producing genuine data, this mode may raise some privacy considerations related to the use of real images. Alternatively, the plugin can generate pixel data with 'noise', i.e. random byte streams or black images, according to users' requirements.

The generation of a dataset of several years' radiology practice will require several terabytes of storage space to accommodate it. This can be a serious problem in many experimental environments. Despite seeming a major drawback, pixel data is rarely needed in many technology trials. Moreover, laboratory trials requiring access to image pixel data also usually need to have datasets that are manually curated and controlled. Taking this into consideration, the proposed system also allows the reduction of pixel data size, making the generation of huge datasets possible. The strategy to deal with this issue involves stripping the pixel data from the generated images. In experiments performed in the Dicoogle environment, these DICOM objects can be completed with pixel data when requested because the developed system includes a third module (plugin) that generates it on the fly.

4 Results

This section demonstrates the feasibility of the proposed system in generating Big Data repositories to support medical imaging informatics research. In the scope of other research projects, involving clinical partners, Dicoogle was used to collect real data from distinct types of healthcare institutions. One of these datasets, from a medium-sized facility, covers roughly four years of medical imaging practice, enclosing 300,000 studies from four modalities (CT, CR, DX, US). It is a comprehensive dataset that shows the enormous amount of data produced in medical imaging laboratories nowadays.

The distribution of the ultrasound weekly study production is shown in Fig. 3 (blue), as captured by Dicoogle and our statistical analyser. Over the years, roughly 31,000 ultrasound studies were produced. We have chosen this modality because early in 2012 it had a significant increase in the number of studies produced, due to the inclusion of much equipment in the institutional PACS. This allows us to better demonstrate the pertinence of the proposed method.

The system was used to generate an ultrasound modality dataset two years longer than the model dataset. The production distribution of the generated dataset is shown in Fig. 3 (orange). It can be observed that the study production closely follows the model dataset. Moreover, comparing with the previous version of our method Fig. 3 (green), which was uniquely based on descriptive statistics rather than on stochastic processes, it models the trend in the production distribution more effectively. This is specially felt when the model is used in forecasting mode, i.e. used to generate a dataset larger than the model.

The same pattern is observed when analysing the statistical markers of both datasets over time, as shown in Table 1. The table

Table 1 Productivity statistics across the different datasets

Datasets	1st year	2nd year	3rd year	4th year	5th year	Overall
model	μ: 25.9	μ: 70.8	μ: 114.9	μ: 149.1		μ: 149.1
	σ: 12.2	σ: 65.3	σ: 86.5	σ: 98.6		σ: 98.6
	Σ: 1323	Σ: 7292	Σ: 17923	Σ: 30861		Σ: 30861
generated B	μ: 65.9	μ: 112.3	μ: 145.8	μ: 175.5	μ: 196.7	μ: 208.1
	σ: 19.4	σ: 50.2	σ: 61.9	σ: 74.1	σ: 77.5	σ: 79.4
	Σ: 2043	Σ: 9324	Σ: 19826	Σ: 32988	Σ: 47214	Σ: 56599
generated C	μ: 150.0	μ: 150.3	μ: 150.0	μ: 149.6	μ: 149.6	μ: 149.3
	σ: 10.4	σ: 9.9	σ: 9.6	σ: 10.0	σ: 9.7	σ: 9.7
	Σ: 7650	Σ: 15476	Σ: 23400	Σ: 31112	Σ: 38902	Σ: 43582

shows the cumulative mean (μ), standard deviation (σ), and total number of studies (Σ) in the first N year period. It is perceivable that in the first year the model dataset has a relatively small weekly study production, and that it increases steadily over time. This leads to overestimation of the weekly study production by the old model (Generated C). This effect is severely limited by the capacity of the ARMA regression to capture the rising trend of study production (Generated B).

In relation to the volume of data produced, the proposed system is able to provide a clear reduction of storage space for a dataset generated without pixel data. For a dataset consisting of 4,295,069 images, which required 870,955.6 MB on the storage medium, the exported dataset ended up requiring less than 2%, 17,161 MB, of the dataset stored along with the pixel data.

5 Discussion

The proposed methods are well suited to any scenario that requires simulated DICOM metadata. The results in the previous section reinforce that the quality of resulting datasets makes possible their usage to test multiple PACS services in load environments, such as storage, query and retrieve. A major aspect of the proposed system is its immediate utility and the Dicoogle project is a good example of this statement. Dicoogle has been offering search and retrieval services over all metadata contained in the DICOM images of its repository. The development of its indexing engine was subject to a continuous validation process to assess its response to increasing load, and how the various patches developed influenced the system behaviour. This process required strong datasets to validate the solution. For instance, the dataset used to address the reduction in storage space only produced an index with 50 GB. Despite seeming a considerable size, it is not representative of several years of data produced in a medium-sized institution. Using datasets that are not representative of actual real-world institutions, we can never be sure about the long-term performance of new technological approaches until they are actually deployed in a real production environment.

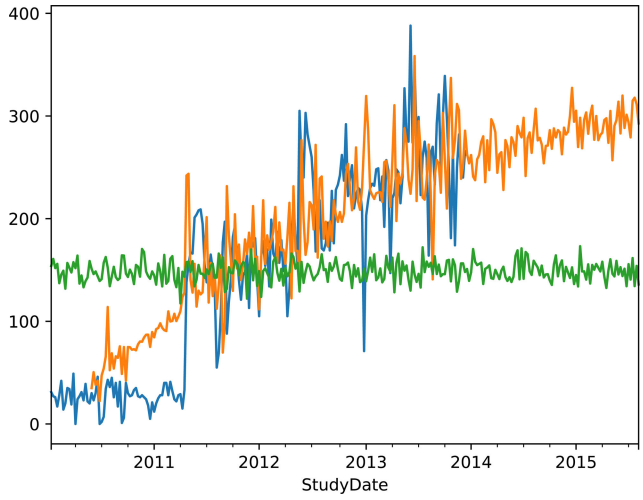

Fig. 3 *Distribution of weekly ultrasound studies over time on the different datasets. The model dataset in (blue), generated dataset with the presented method (orange); a generated dataset with an alternative statistical model (green). Best viewed in colour online*

Currently, there are two other projects using datasets generated by our system, aiming to test database technologies in medical imaging big data scenarios, namely architectures based on relational and document-oriented models [24].

Finally, our tool falls short in scenarios that require the actual pixel data. To overcome this limitation, the system has the option of inserting the pixel data gathered from actual images. This trick is useful in situations such as testing the storage services of the PACS. Nonetheless, in scenarios such as content-based imaging retrieval, this tool is not applicable as it cannot generate meaningful pixel data.

6 Conclusions

Nowadays, researchers on medical imaging networks and storage systems face a major issue when validating their contributions. The lack of big datasets that perfectly mimic real institutional scenarios is a serious constraint regarding the quality of their contributions. Many of those proposals have been validated with datasets that are simply not large enough to stress them, i.e. push the system towards its operational limit to determine its robustness, availability, and error handling under heavy load as real laboratories would. Drawing conclusions based on poorly designed trials may induce errors and hinder future research efforts. Moreover, most developers do not have a clear picture of how their solutions will perform in the long-term. Due to this lack of knowledge, unpredicted episodes may take place during the PACS lifetime, compromising the healthcare service provision. The proposed methods address these issues by enabling the creation of artificial datasets with the same characteristics as a model dataset, for instance, a real-world institutional repository. Moreover, these methods enable the creation of larger datasets to be used in long-term and big data exploration scenarios. For this purpose, the approach used to reduce the size of the generated datasets seems vital.

7 Acknowledgments

This work was financed by the ERDF – European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation – COMPETE 2020 Programme, and by National Funds through the FCT – Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project «CMUP-ERI/TIC/0028/2014». Tiago Marques Godinho is funded by FCT under grant agreement SFRH/BD/104647/2014.

8 References

[1] Viana-Ferreira, C., Costa, C.: 'Challenges of using cloud computing in medical imaging', in Ramachandran, M. (Ed.): '*Advances in Cloud Computing Research*' (Nova Publisher, 2014)

[2] Tavakol, P., Labruto, F., Bergstrand, L., *et al.*: 'Effects of outsourcing magnetic resonance examinations from a public university hospital to a private agent', *Acta Radiol.*, 2011, **52**, (1), pp. 81–85

[3] Silva, L.B., Costa, C., Oliveira, J.: 'A PACS archive architecture supported on cloud services', *Int. J. Comput. Assist. Radiol. Surg.*, 2012, **7**, (3), pp. 349–358

[4] Huang, H.K.: '*PACS and imaging informatics: basic principles and applications*' (Wiley, 2010, 2nd edn.)
[5] Pianykh, O.S.: '*Digital imaging and communications in medicine (DICOM): A practical introduction and survival guide*' (Springer, 2008)
[6] Santos, M., de Francesco, S., Silva, L.A.B., *et al.*: 'Multi vendor DICOM metadata access a multi site hospital approach using Dicoogle'. Information Systems and Technologies (CISTI), 2013 8th Iberian Conf. on, 2013
[7] Savaris, A., Härder, T., von Wangenheim, A.: 'DCMDSM: a DICOM decomposed storage model', *J. Am. Med. Inf. Assoc.*, 2014, **21**, (5), pp. 917–924
[8] Jin, Z., Chen, Y.: 'Telemedicine in the Cloud Era: prospects and challenges', *IEEE Pervasive Comput.*, 2015, **14**, (1), pp. 54–61
[9] Hui, Y., Kundakcioglu, E., Jing, L., *et al.*: 'Healthcare intelligence: turning data into knowledge', *IEEE Intell. Syst.*, 2014, **29**, (3), pp. 54–68
[10] Sernadela, P., Lopes, P., Oliveira, J.L.: 'A knowledge federation architecture for rare disease patient registries and biobanks', *J. Inf. Syst. Eng. Manage.*, 2016, **1**, pp. 83–90
[11] Prado, T., de Macedo, D.J., Dantas, M.A.R., *et al.*: 'Optimization of PACS data persistency using indexed hierarchical data', *J. Digit. Imaging*, 2014, **27**, (3), pp. 297–308
[12] Costa, C., Ferreira, C., Bastião, L., *et al.*: 'Dicoogle - an open source peer-to-peer PACS', *J. Digit. Imaging*, 2011, **24**, (5), pp. 848–856
[13] Bastiao Silva, L.A., Beroud, L., Costa, C., *et al.*: 'Medical imaging archiving: A comparison between several NoSQL solutions'. Biomedical and Health Informatics (BHI), 2014 IEEE-EMBS Int. Conf. on, 2014
[14] Godinho, T.M., Viana-Ferreira, C., Silva, L.A.B., *et al.*: 'A routing mechanism for cloud outsourcing of medical imaging repositories', *IEEE J. Biomed. Health Inf.*, 2016, **20**, (1), pp. 367–375
[15] National Electrical Manufacturers Association, Digital imaging and communications in medicine (DICOM), (NEMA, 2009) available at http://dicom.nema.org/standard.html
[16] Larobina, M., Murino, L.: 'Medical image file formats', *J. Digit. Imaging*, 2014, **27**, (2), pp. 200–206
[17] Pianykh, O.S.: '*Digital imaging and communications in medicine (DICOM)*' (Springer, 2011)
[18] Ismail, M., Philbin, J.: 'Multi-series DICOM: an extension of DICOM that stores a whole study in a single object', *J. Digit. Imaging*, 2013, **26**, (4), pp. 691–697
[19] Savaris, A., Härder, T., Wangenheim, A.v.: 'Evaluating a row-store data model for full-content DICOM management'. IEEE 27th Int. Symposium on Computer-Based Medical Systems, 2014
[20] Rascovsky, S.J., Delgado, J.A., Sanz, A., *et al.*: 'Informatics in radiology: use of CouchDB for document-based storage of DICOM objects', *Radiographics*, 2012, **32**, (3), pp. 913–927
[21] Macedo, D.D.J.D., Wangenheim, A.V., Dantas, M.A.R., *et al.*: 'An architecture for DICOM medical images storage and retrieval adopting distributed file systems', *Int. J. High Perform. Syst. Archit.*, 2009, **2**, (2), pp. 99–106
[22] Valente, F., Silva, L.B., Godinho, T., *et al.*: 'Anatomy of an extensible open source PACS', *J. Digit. Imaging*, **2015**, pp. 1–13
[23] Brockwell, P.J., Davis, R.A.: '*Time series: theory and methods*' (Springer-Verlag, New York, 1991, 2nd edn.)
[24] Alves, A.P., Godinho, T.M., Costa, C.: 'Assessing the relational database model for optimization of content discovery services in medical imaging repositories'. IEEE 18th Int. Conf. on e-Health Networking, Applications and Services (Healthcom), 2016