Separation of metadata and bulkdata to speed DICOM tag morphing

Mahmoud Ismail[1], Yu Ning[2], James Philbin[2]
[1]Department of Computer Science, [2]Department of Radiology
Johns Hopkins University

ABSTRACT

Most medical images are archived and transmitted using the DICOM format. The DICOM information model combines image pixel data and associated metadata into a single object. It is not possible to access the metadata separately from the pixel data. However, there are important use cases that only need access to metadata, and the DICOM format increases the running time of those use cases. Tag morphing is an example of one such use case. Tag or attribute morphing includes insertion, deletion, or modification of one or more of the metadata attributes in a study. It is typically used for order reconciliation on study acquisition or to localize the Issuer of Patient ID and the Patient ID attributes when data from one Medical Record Number (MRN) domain is transferred to or displayed in a different domain.

This work uses the Multi-Series DICOM (MSD) format to reduce the time required for tag morphing. The MSD format separates metadata from pixel data, and at the same time eliminates duplicate attributes. MSD stores studies using two files rather than in many single frame files typical of DICOM. The first file contains the de-duplicated study metadata, and the second contains pixel data and other bulkdata. A set of experiments were performed where metadata updates were applied to a set of DICOM studies stored in both the traditional Single Frame DICOM (SFD) format and the MSD format. The time required to perform the updates was recorded for each format. The results show that tag morphing is, on average, more than eight times faster in MSD format.

Keywords:
DICOM, Multi-Series DICOM, tag morphing, attribute coercion, patient information reconciliation, order reconciliation, study localization, PACS, VNA

1. INTRODUCTION

With the rapid development of 3D imaging modalities such as CT and MRI, the size of DICOM studies is growing. DICOM studies often contain hundreds of images. The standard DICOM format (see Figure 1a) typically stores each image in a separate object, called an instance, which includes metadata and pixel data. The metadata contains information about the patient, the physician, the imaging modality, scan parameters, image orientation, etc. Some of the parameters are related to the study and do not vary with the images, e.g. the patient name, while some others are image specific, e.g. the slice location. The size of the metadata is, in general, small compared to that of the pixel data. While coupling metadata and pixel data is useful for displaying individual images with related information, it creates unnecessary data redundancy, because the study and series level attributes are repeated with each DICOM instance; in addition, often many of the instance level attributes have the same value across the containing series. Furthermore, combining pixel data and metadata in a single object increases the time required to update the metadata, because modifying the metadata also requires reading and writing the pixel data, which is much larger than the metadata. Fast, easy access to metadata is desired for multiple real-world use cases, such as order reconciliation, search indexing, de-identification, and study transfer across Medical Record Number (MRN) domains. These applications rely on tag morphing, which is defined as adding, deleting or modifying the attributes of a DICOM study so that the study is accurate and usable in the target domain.

Examples of tag morphing include order reconciliation and study localization. Order reconciliation occurs when a PACS updates a study's attributes as it is received from a modality. This is done in order to ensure that certain attributes such as Patient Name, Patient ID, Issuer of Accession Number and Accession Number have the same values and format as the study order, which is typically created on a RIS, and not those created by the modality, which may be incorrectly formatted or missing altogether. Study localization must be done when a study is being transmitted to or displayed at a MRN domain that is different from the domain in which the study was created. Attributes such as Patient ID and Issuer

Medical Imaging 2014: PACS and Imaging Informatics: Next Generation and Innovations,
edited by Maria Y. Law, Tessa S. Cook, Proc. of SPIE Vol. 9039, 903905 · © 2014 SPIE
CCC code: 1605-7422/14/$18 · doi: 10.1117/12.2043933

of Patient ID must be updated to those of the new domain; other attributes such as Issuer of Accession Number and Accession Number might be removed or modified,[1, 2] and the facility receiving the study might insert facility specific attributes.[3] Tag morphing for a large study stored in the traditional DICOM format requires many I/O operations (network and/or file system) to access potentially hundreds of SFD instances, in order to update the metadata for every instance.

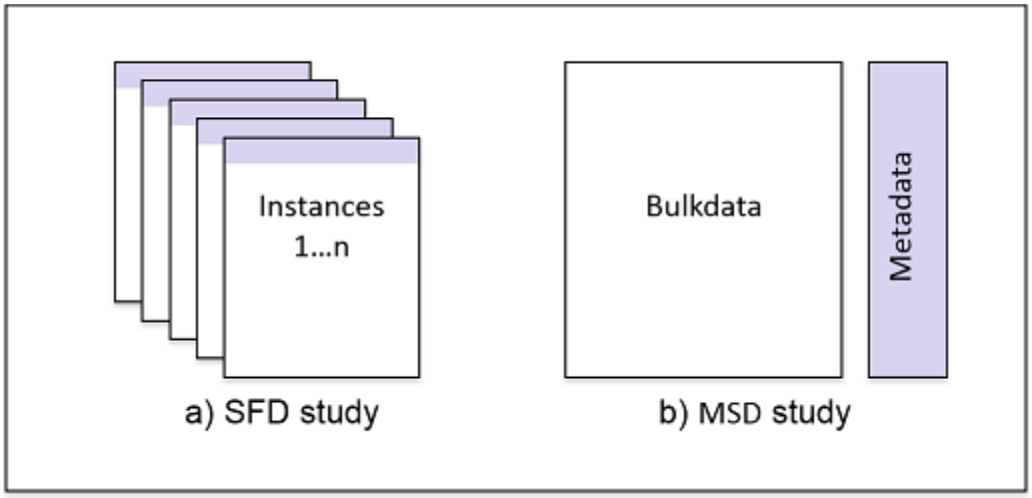

Figure 1: Structure of a Single Frame DICOM (SFD) study versus that of a Multi-Series DICOM (MSD) study.

1.1 The Multi-Series DICOM (MSD) format

The MSD format was developed with three goals in mind: 1) aggregating all the instances in a study into one data structure, 2) separating the metadata from the *bulkdata*, i.e., large values such as pixel, overlay and lookup table (LUT) data, and 3) eliminating duplicate attributes. MSD is an extension of Multi-Frame DICOM (MFD), which combines all the images contained in a series into a single DICOM instance.[5] MFD uses an attribute called Per-frame Functional Groups Sequence that contains a sequence of nested data sets, where each holds the attributes associated with a frame in the series. The advantage of MFD over traditional Single Frame DICOM (SFD) is that it does not repeat study and series level attributes within each frame in the series.

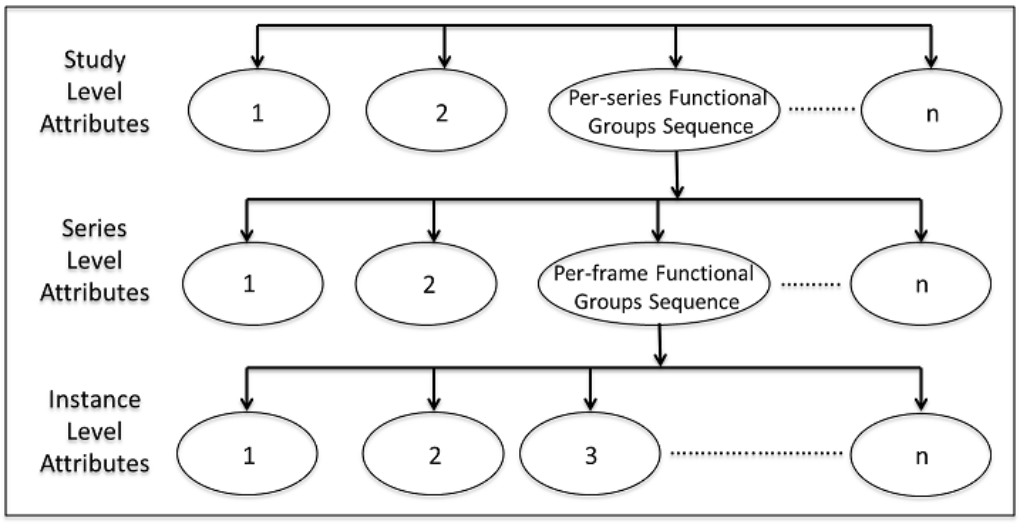

Figure 2: The nested structure of the MSD metadata.

MSD extends this idea by aggregating all of the series in a study into two objects, metadata and bulkdata. MSD uses an attribute called Per-series Functional Groups Sequence to aggregate all the series level attributes into a single Study object. This attribute contains a sequence of nested data sets that contain the attributes associated with all of the series in the study, which results in removing replicated series attributes within the study. These nested data sets have a similar structure to MFD. Figure 2 shows the nested structure of MSD. The One-Pass De-Duplication Algorithm[4] is used to perform this aggregation by efficiently finding and removing repeated attributes. The One-Pass De-Duplication Algorithm actually reduces the overhead of parsing input studies.

Figure 1 shows the difference between the traditional SFD and the MSD formats. For SFD, the DICOM study is represented using multiple instances, with each instance containing all the relevant study, series and instance level attributes. Most if not all of the study and series attributes are duplicated in each instance. MSD combines all instances into two objects, metadata and bulkdata. The metadata eliminates all unnecessary duplication of attributes. All large attribute values (really value fields) are moved to the bulkdata object, so that the metadata can be retrieved quickly.

1.2 Separating metadata from bulkdata

The MSD format was developed to allow bulkdata, i.e. large attribute values such as pixel data, to be stored in separate objects. For this paper bulkdata is defined to be any attribute values larger than 256 bytes. When a large value is moved into a bulkdata object, the original attribute value is replaced by a Bulkdata Reference. Data types in DICOM are known as value representations (VRs). For the MSD format a new VR, with a symbol 'BD', was created for Bulkdata References. A Bulkdata Reference contains 14 bytes that are structured as follows: The first two bytes contain the original VR of the attribute. The next four bytes hold the index of the attribute value in the bulkdata object, followed by four bytes that store the offset to the first byte of the attribute value within the bulkdata object, and the last four bytes hold the size of the value.

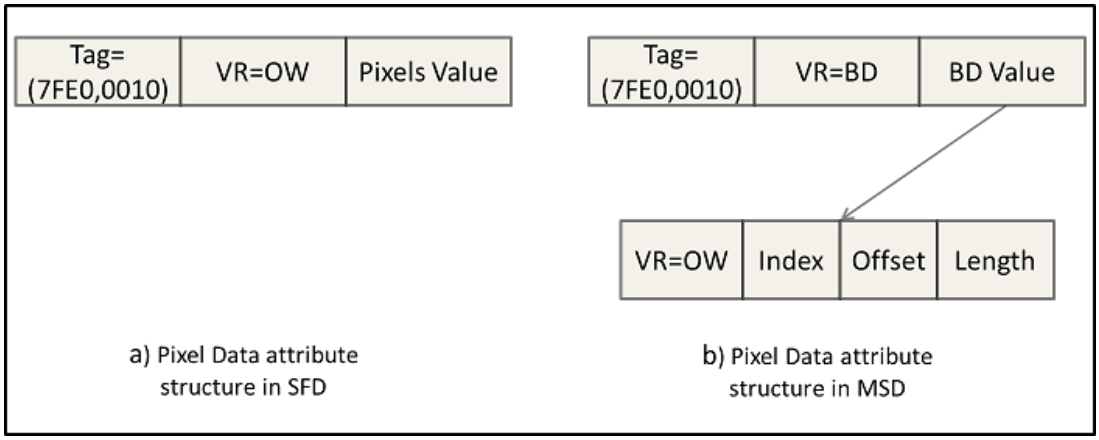

Figure 3: A Pixel Data attribute in traditional DICOM format and its counterpart in MSD format.

In summary, the MSD format is composed of one object that contains the metadata and another that contains bulkdata, i.e. large attribute values. The metadata contains Bulkdata References to bulkdata values. Figure 3 shows the representation of the Pixel Data attribute in the SFD format and its corresponding representation in the MSD format. MSD also addresses the limitation of current PACS and VNA implementations that store the metadata in a database and apply metadata changes to the database only (see Related Work).

This work investigates the performance improvement achieved by using the MSD format rather than the traditional SFD format. Performing tag morphing on a study stored in the MSD format should be more efficient than performing the same modifications on the same study stored in the SFD format. There are two reasons for this: First, the metadata in MSD format is smaller in size than its counterpart in the traditional format, and it has been demonstrated that retrieving studies in MSD format is faster than retrieving the same studies in SFD format. Second, the metadata of the whole study is separated from the bulkdata and they are contained in separate files. Consequently, the metadata can be accessed from one relatively small file, rather than accessing hundreds of SFD files.[4, 6]

2. RELATED WORK

Tag morphing has long been employed as an important tool that enables the exchange of digital medical images between modalities and PACS products of diverse manufacture, since even before the emergence of the DICOM standard.[3] Its application in patient information reconciliation is recognized and standardized by the Integrating the Healthcare Enterprise (IHE) initiative.[1] This has resulted in more and more hospitals implementing new workflows, including the use of supplemental software packages, to properly import external DICOM studies into their own PACS, usually from CD/DVDs. These workflows require human intervention for coercion of DICOM metadata, because the new values must be either manually input or at least confirmed by an operator.[1]

In recent years, the high demand for medical image sharing over the Internet has led to the evolution of Vendor Neutral Archives, which facilitate the large-scale aggregation of medical images. While most PACS do order reconciliation, one of the features that distinguish VNAs from PACSs is automatic study localization, i.e. manipulation of patient/study identifiers and possibly other DICOM attributes for outgoing studies, so as to ensure that studies originated from one institution can be correctly viewed or imported by another's PACS or VNA.[3]

In order to speed up tag morphing, most PACS and VNAs, such as the DCM4CHEE Archive,[7] keep a copy of selected metadata attributes in a database. When these attributes need to be modified, only the database entry is updated, but not the files that contain the study. Then, when the study files are retrieved, tag morphing is performed dynamically on each file in the study. This approach just delays the overhead of tag morphing until the study is retrieved, at which time the metadata in each file must be modified. It also makes the study metadata stored in the database inconsistent with that in the study files. This is undesirable because the database will become a bottleneck for study retrieval.

We are not aware of any literature that investigates the efficiency of DICOM tag morphing and how to improve it, which is the topic of this paper.

3. METHODS

This experiment takes advantage of the Java implementation of the MSD API[4] introduced previously, which was built on top of the dcm4che2 toolkit.[8] It supports reading and writing studies in both SFD and MSD formats. In this experiment, it is used to convert a set of DICOM studies to the MSD format in advance. The experiment was designed to compare the times required to tag morph each study in both SFD and MSD formats. The MSD Toolkit was modified to include procedures for inserting, deleting, or modifying study attributes.*

Study Name	# Series	# Images	Bulkdata Size (KB)	Metadata Size				
				SFD		MSD		SFD/MSD
				(KB)	%	(KB)	%	
SMALLMR	9	277	71,488	969	1.3%	158	0.22%	6.1
SMALLCT	5	338	173,088	920	0.5%	174	0.10%	5.3
TESTMR	17	1,116	213,352	4,278	2.0%	594	0.27%	7.2
TESTCT	7	1,018	613,926	3,310	0.5%	1,193	0.19%	2.8
TESTCTA	13	2,524	1,366,321	8,052	0.6%	2,845	0.21%	2.8
BREASTMR	22	2,362	1,499,589	8,605	0.6%	1,285	0.09%	6.7
Average	**12**	**1,273**	**656,294**	**4,356**	**0.7%**	**1,041**	**0.16%**	**4.2**

Table 1: Input study properties

The input dataset is composed of six different DICOM studies, three MRIs and three CTs, in two different formats (SFD and MSD) for a total of twelve studies in the dataset. The study sizes range between 70 MB and 1.5 GB. The dataset

* dcm4che2 already provides similar functions for manipulating SFD metadata.

properties are shown in Table 1. The metadata is a small percentage of the overall study size. On average, the SFD metadata is 0.7% and MSD is 0.16% of the original study size.

The experiment was carried out on each of the twelve studies in the following steps: 1) read each study from the file system into memory, 2) update the values of a predefined set of attributes in the study with dummy values, 3) save the updated study from memory to the file system in its original format (SFD or MSD), and 4) record the time for steps 1 to 3. The attributes chosen to be updated are Issuer of Patient ID, Patient ID, and Accession Number. They were selected because it is mandatory to update them when a study is transferred across MRN domains. The experiments were performed on a quad core 2.27 GHz x86 processor with 48GB of physical memory and 8GB of allocated heap memory.

Study Name	SFD Time (ms)	MSD Time (ms)	Speedup (%)
SMALLMR	519	150	346
SMALLCT	791	162	488
TESTMR	1518	288	527
TESTCT	2282	291	784
TESTCTA	4695	418	1123
BREASTMR	5251	390	1346
Average	**15056**	**1699**	**886**

Table 2: Tag morphing performance of SFD versus MSD

4. RESULTS

The results in Table 2 show that the processing time for tag morphing the studies stored in MSD format is, on average, more than eight times faster than applying the same changes to those in the standard SFD format. There are two reasons for this. First, the MSD metadata is de-duplicated, which makes it four times smaller, on average, than the metadata in SFD format (see Table 1), which reduces the parsing time. Second, and more importantly, the MSD format does not need to read the bulkdata into memory or write it back to the file system, which reduces I/O time and memory footprint.

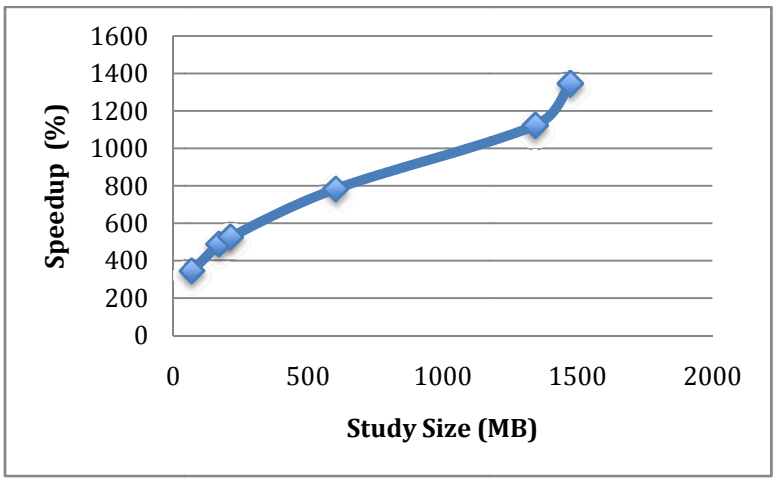

Figure 4: Tag morphing speedup versus study size.

Finally, it should be noted that performance of MSD improves as the study size increases. Figure 4 shows that the speedup is proportional to the study size. This performance difference is especially important for large hospitals that process hundreds of imaging studies daily. It could be even more important for regional Health Information Exchanges that will eventually manage billions of images for millions of people.

5. CONCLUSION AND FUTURE WORK

The MSD format is novel in that it separates metadata from bulkdata. The results show a significant improvement in processing time for tag morphing. Moreover, MSD addresses the limitation of current PACS and VNA implementations that store the metadata in a database and apply metadata changes only to the database. MSD is efficient for tag morphing for two reasons. First, there is no need to read or transmit the pixel data to access the metadata. Second, the size of metadata is significantly smaller than that of the traditional SFD metadata. To our knowledge, this is the first work to demonstrate the benefits of processing metadata separately from bulkdata for tag morphing.

Future work may include an evaluation of the end to end scenario that models the common tag morphing use cases in clinical facilities. The framework used in this experiment reads and writes the studies from the file system while, in reality, the studies are stored in an archive and transferred to a remote workstation on demand. Accessing studies from the archive introduces transmission delays. The end to end scenario where a study is retrieved from an archive, updated and transmitted over the network to a remote workstation will show the performance of MSD versus SFD in a full clinical workflow.

REFERENCES

[1] Van Ooijen P. M. A., Guignot J., Mevel G. and Oudkerk M., "Incorporating Out-Patient Data from CD-R into the Local PACS using DICOM Worklist Features," J. Digital Imaging 18(3), 196-202 (2005).

[2] "Digital Imaging and Communications in Medicine (DICOM) Part 6: Data Dictionary," National Electrical Manufacturers Association, 2011. http://medical.nema.org/Dicom/2011/11_06pu.pdf. (Accessed 12 February 2014).

[3] DeJarnette W. T., "Context Management and Tag Morphing in the Real World," DeJarnette Research Systems, Inc. White Paper Series, 4 January 2010. http://www.dejarnette.com/downloads/get.aspx?i=45048 (Accessed 12 February 2014).

[4] Ismail M. and Philbin J., "Multi-series DICOM: an Extension of DICOM that Stores a Whole Study in a Single Object," J. Digital Imaging 6(4), 691-697 (2013).

[5] "Digital Imaging and Communications in Medicine (DICOM) Part 3: Information Object Definitions," National Electrical Manufacturers Association, 2011. http://medical.nema.org/Dicom/2011/11_03pu.pdf. (Accessed 12 February 2014).

[6] Ismail M. and Philbin J., "Fast, Storage Efficient De-identification of Medical Studies," The DICOM International Conference and Seminar, March 2013.

[7] Zeilinger G., "DCM4CHEE Archive 4.x," 16 September 2013. https://github.com/dcm4che/dcm4chee-arc (Accessed 12 February 2014).

[8] Evans D., "dcm4che2 DICOM Toolkit," 25 July 2012. http://www.dcm4che.org/confluence/display/d2/dcm4che2+DICOM+Toolkit (Accessed 12 February 2014).