A DICOM Document-Oriented Approach to PACS Infrastructure

Fred Behlen

The need for long-term storage requires the future migration of image data from a PACS to its successor system. This paper considers the cost of such migration. It is proposed that storage of data as "documents" in DICOM Part 10 formats on industry-standard media could reduce the time and cost of data migration relative to on-line DICOM transfer. The relation to present efforts in developing document-oriented electronic patient records is discussed. DICOM Part 10 files are found to be a sufficient representation of image documents, but additional software tools will be needed to reach its full potential. There is a significant cost benefit of the document storage method, but it is one of many factors which must be balanced in the selection of a PACS.

KEY WORDS: PACS, Archive, Data Migration, Document-oriented, SGML

A PACS, like any patient records system, must store data for a long time, much longer than the expected useful life of any of the computer components on which it is built. The only viable long term strategy is to plan for the migration of data to successor systems in the future.[1] In the best case, this migration would be done automatically from the operating old system to the new, but during that transfer time, both systems would have to be operational. In the worst case, the old system could be impaired and significant operator or programming effort would be required to transfer data from a vendor's proprietary media format.

If the data is stored in information objects which map directly to the clinical events that produced them, and if the format of those information objects is standardized and contains the information needed to relate them to each other, then the entire database can be reconstructed from the storage media. This is happily the case if DICOM Information Entities are stored in Part 10 files on non-proprietary media in what could be called a document-oriented PACS.

DOCUMENT-ORIENTED PATIENT RECORDS

The term "document oriented" is taken from related work in electronic patient records. Such work is exemplified in the activities of the Health Level Seven SGML/XML Special Interest Group (HL7 SGML/XML SIG).[2] The SGML/XML SIG aims to develop standards for the storage and interchange of electronic health records as applications of Standard Generalized Markup Language (SGML), a powerful document processing environment employed in publishing.[3] The SIG's efforts are focused on XML, a subset of SGML which is widely viewed as a successor to HTML for Internet World Wide Web implementations.

The motivation for this document-oriented approach lies not only in the desire to leverage powerful mass market document management tools. The power of SGML also lies in the capability of markup to provide access to tagged fields while accommodating semantic diversity of varied source information. In this regard the document approach avoids the escalating complexity of database schemas which occurs when patient records are stored in relational databases. SGML/XML also provides for a written format of the document file which can be validated against a Document Type Definition (DTD) and transferred easily between systems. That same capability of transferring data from one system to another also enables transfer from an old system to its successor.

However, the move toward document-oriented patient records also represents a paradigm shift from viewing patient records systems as databases which communicate by messages to viewing them as collections of documents which communicate by document interchange. The shift could appropriately be viewed as a return to a traditional outlook, as paper-based patient records have always been considered collections of documents, in the classical definition of "an original or official paper relied on as the basis, proof, or support of something."[4] We had come to view the electronic systems as databases because such systems grew out of hospital information systems originally implemented to manage the dynamic data in the healthcare setting, while permanent records were retained in paper form.

From the University of Chicago, Department of Radiology.
Address reprint requests to Fred M. Behlen, Assistant Professor of Radiology, The University of Chicago, MC 2026, 5841 South Maryland Avenue, Chicago, IL 60637.

0897-1889/98/1103-1012$8.00/0

Both documents and messages are packaged data, but several characteristics distinguish information entities that comprise a repository of medical records:

- Identification. Each document must contain information linking it to a particular patient.
- Attestation. The information in a medical records document must be attributed to a source who attests to its validity.
- Versioning and revision history. A document once posted cannot be changed, but may be superseded by a revised document. Information in the document must enable the identification of prior versions.
- Explicit context. Documents, unlike messages, do not operate in real time where the context of prior messages can be assumed. Documents must explicitly define or reference the context of the information they contain.
- Self assembly. The documents should contain all the information required to define the relationships among documents. Ideally, it should be possible for a system to input a collection of documents in any order for any number of patients, and assemble them into a coherent patient record for each patient.

This is not intended to be a definitive or irreducible set of requirements for "document-orientation," but rather a list of illustrative features.

Nothing within the document-based view dictates how a system actually stores the document data. The documents could be stored verbatim, or at the other extreme could be completely parsed and stored in a fully normalized database. What is required is that the system be able to deliver the complete set of documents on demand. It is expected that many systems will employ a hybrid approach in which documents are stored in their original form and some elements will be parsed out and stored in databases as indexes or searchable fields. Rather than intrinsic to the document-based view, *it is a system management decision that copies of documents should be stored on non-proprietary media.* Document orientation promotes standards which allow this policy to be enacted.

DICOM "DOCUMENTS"

Fortuitously, DICOM Composite Information Objects stored as files defined in DICOM Part 10 possess most of the document characteristics outlined above.

- Identification. The Patient Name, ID, Date of Birth and the Accession Number comprise a basic set of identifying information, although additional information would be are desirable.
- Attestation. The information source device is identified.
- Versioning and revision history. Certain instances are supported, but DICOM's support in this respect is not complete.
- Explicit context. The negotiated message context is stored in the Part 10 header.
- Self assembly. The use of Unique Identifiers in the Relational elements (such as study instance UID) illustrate how self assembly can be performed.

As a result, at our institution we concluded that a written collection of DICOM Part 10 information objects would be a sufficient representation of the Radiology image data. Transporting the stored data to another system could then be done on off-line media, without the requirement that the original system be available.

VALUE OF DICOM DOCUMENT STORAGE

In the selection of a supplier for the second phase of our PACS program, we desired a document-oriented system as described above, but recognized that it is one of many factors that must be considered. Thus, we undertook an analysis to estimate the value of a document-oriented system relative to other offerings.

Radiology Information System records and procedure volume statistics were analyzed to compute the total data load needed to be copied to a successor system at five years in the future. The results are outlined in Table 1. Procedure volume was assumed to grow 4% annually from projected figures for 1998. Storage requirements were computed from procedure volume broken down by

Table 1. Estimated Procedure Volumes and Storage Requirements Using Lossless Compression

	1998	1999	2000	2001	2002	2003
Procedures/year	179,500	186,680	194,147	201,913	209,990	218,389
Terabytes/year (lossless compression)	1.69	1.76	1.83	1.90	1.98	2.06
Terabytes stored	1.7	3.4	5.3	7.2	9.2	11.2

modality and subspecialty into 24 groups. For each group, the number of images and lossless compression factor was estimated, which together with the known procedure volume and image size produced an estimated storage requirement. These subspecialty storage requirements sum to 1.69 Terabytes for 1998, and are assumed to grow proportionally to overall procedure volume thereafter. At the end of the fifth year following, the total stored data would be 11.2 Terabytes.

We are planning to acquire the new PACS archive in 1998 and backload it with approximately one year of digital data already stored in other systems. If the PACS archive is replaced at the end of five years, nearly all of this data will have to be transferred to the successor system. The strict requirement in Illinois is to keep each image for five years for adults. Thus, only the 1998 data (15% of the total stored) would be a candidate for deletion at that point. Furthermore, our hospital policy has been to retain complete folders for any patient who has had a radiology procedure within the last five years. Also, the stored data will include images from mammographic and pediatric exams which must be retained for longer periods. Thus, we assume that at least 90% of the 11.2 Terabytes of compressed image data will have to be migrated.

The recommendation of most manufacturers is to transfer the old data to the new system using DICOM. A limit today is how fast the systems can accept DICOM data, parse it and populate their databases, and store the image data. There are substantial additional costs for providing DICOM input bandwidth beyond what is required for accepting current clinical images. Thus it is assumed that the input capacity of the new archive will be scaled to accept the then-current clinical data load, and that the transfer of old data would have to use capacity available during off-peak hours. We will assume that the new archive is scaled with enough capacity so that it can accept all current clinical data in four peak hours of each business day (20 hours per week), and that the time available for transfer loading would be 8 hours each weeknight and 16 hours per day on weekends. Thus, there would be 72 hours per week available for the transfer and it would take approximately 20/72 of a year to transfer the data collected in a year. During the transfer process, the old archive will continue to occupy space, and staff will be required to administer the process and deal with exception conditions. We estimate the need for 25% of a programmer's time during the transfer process. Finally, the future cost must be discounted to present value to compare it against present costs in the selection of an archive. Table 2 details the calculation.

DISCUSSION AND CONCLUSIONS

As Table 2 shows, the cost of transfer of stored data to a successor system through the DICOM interface is expected to be significant, with a net present value of $228,564. Furthermore, the assumptions going into that calculation are far from the most conservative that could be made. For example, the limitations of the transfer rate were assumed to be the new archive's ability to accept DICOM data, whereas the old archive's ability to deliver data may easily be the limiting factor and at a far lower level. The assumed ability of the archive to load 2.06 Terabytes per year during 20 hours per week translates to 549 kilobytes per second, a rate the old archive may not be able to deliver for sustained periods.

The cost of transfer of a document-oriented image archive is likely to be significantly lower, but is more difficult to estimate because the performance of transfer tools is unknown. The transfer of image data using file transfer utilities would still take 4.5 months at 1 megabyte per second. This assumes that the transfer could be done in all but peak hours, as the DICOM interface and database resources would not be used. Additional time

Table 2. Computation of Cost of Transfer of Stored Data From PACS Archive to the Successor System, Five Years in the Future

Hours/week for primary storage	20
Hours/week for transfer storage	72
New study loading capacity, TB/year, during available hours	2.06
Transfer capacity, TB/year	7.40
Total stored data in old archive, Terabytes	11.2
Percent of archive to be transferred	90%
Terabytes to be transferred	10.09
Years required for transfer	1.36
Old archive maintenance cost, annual	$200,000
Old archive space cost (100 s.f. @ 30), annual	$3,000
Annual staff cost for transfer operation (0.25 FTE)	$13,725
Annual cost during transfer operation	$216,725
Cost during transfer operation	$295,387
Years to discount to Present Value	5
Annual discount rate	5%
Net Present Value of cost during transfer operation	$228,564

would be required to update the image database on the new archive to include the transferred studies. "Lightweight" database organizations which do not store entries for each image, but rather for each study, will be needed as these archives grow.

In summary, we have concluded that the cost of future migration of image data will be a significant burden. The document-oriented approach is expected to significantly reduce this cost, but optimized tools will be needed which are not available today and it is impossible to estimate their effectiveness. Therefore, at this time we arbitrarily estimate that the amount saved by the document-oriented storage methods is only about half the cost of migrating the data conventionally. While this savings, with a present value of $114,000, is still large, it remains a relatively small portion of the 5-year lifecycle costs of the PACS.

The document approach should be a consideration in purchasing, although it is not an overriding factor. Its benefit is significant, however, and it should be a feature in new designs. These considerations are being used in the selection of a vendor for the next phase of our PACS program.

REFERENCES

1. Behlen FM, Hoffmann KR, MacMahon H: Long term strategies for PACS archive design. *Proc SPIE* 2711:2-8, 1996

2. Extensive documentation is at the HL7 SGML/XML Web Site, a link to which can be found at http://www.hl7.org.

3. Bowns IR: Standard Generalized Markup Language: Should It Have a Role in Clinical Records? Top Health Inform Manage 18(2):24-31, 1997

4. Merriam Webster Dictionary. http://www.m-w.com/cgi-bin/mweb.