Understanding Challenges in Preserving and Reconstructing Computer-Assisted Medical Decision Processes

Sang-Chul Lee and Peter Bajcsy

Abstract— **This paper addresses the problem of understanding preservation and reconstruction requirements for computer-aided medical decision-making. With an increasing number of computer-aided decisions having a large impact on our society, the motivation for our work is not only to document these decision processes semi-automatically but also to understand the preservation cost and related computational requirements. Our objective is to support computer-assisted creation of medical records, to guarantee authenticity of records, as well as to allow managers of electronic medical records (EMR), archivists and other users to explore and evaluate computational costs (e.g., storage and processing time) depending on several key characteristics of appraised records. Our approach to this problem is based on designing an exploratory simulation framework for investigating preservation tradeoffs and assisting in appraisals of electronic records.**

We have a prototype simulation framework called Image Provenance To Learn (IP2Learn) to support computer-aided medical decisions based on visual image inspection. The current software enables to explore some of the tradeoffs related to (1) information granularity (category and level of detail), (2) representation of provenance information, (3) compression, (4) encryption, (5) watermarking and steganography, (6) information gathering mechanism, and (7) final medical report content (level of detail) and its format. We illustrate the novelty of IP2Learn by performing example studies and the results of tradeoff analyses for a specific image inspection task.

Index Terms— **Biomedical decision support systems, -Medical informatics, -Electronic medical record (EMR) data mining**

I. Introduction

There is a vast amount of electronic records in medicine that cannot be utilized, mined and learned from because the records have not been preserved properly. Human or machine learning will be impossible tomorrow if we cannot overcome our lack of understanding how to preserve and reconstruct medical data and decision processes taking place every day. For example, it is critical to compare patients' records acquired today with the patients' records from 5, 10, 50, or 70 years (short term comparisons) in order to assess functional, structural or low level biological changes due to diseases, treatments and/or aging. It is conceivable that future genealogy studies would compare data sets over several hundreds and thousands of years (long term comparisons). Thus, our goal is to understand how to appraise medical electronic records for short and long term preservation and reconstruction purposes in order to enable comparative studies, and human and machine learning.

The motivation for our work comes from the fact that managing electronic medical records (EMR) requires large financial investments with significant ramifications on preservation and reconstruction of medical records. Thus, there is a need to provide institutions managing and appraising EMR with tools to better understand the tradeoffs between information value and computational costs. The overarching motivation of our effort is to provide a simulation environment for optimizing large investments into EMR management and preservation, and making the EMR systems economical and of high information value.

The challenges of preserving medical electronic records have been discussed at several forums, such as at the National Library of Medicine (NLM) [6], at workshops organized by the Research Information Network (RIN) or Medical Research Council (MRC) in UK [7], [8], [9] or in a general context at the National Archives [4]. The first goal in the NLM's vision is "Seamless, Uninterrupted Access to Expanding Collections of Biomedical Data, Medical Knowledge, and Health Information" that includes Recommendation 1.1. "Ensure adequate space and storage conditions for NLMs current and future collections to guarantee long term access to information and efficient service delivery." and Recommendation 1.2. "Preserve NLMs collections in highly usable forms and contribute to comprehensive strategies for preservation of biomedical information in the U.S. and worldwide." Similarly, RIN defines one of the key principles preservation and sustainability (Principle 5) to be concerned with. The preservation challenges include (1) growing amounts of medical data, (2) increasing number of computer assisted medical decisions, and (3) rapid change of storage media and computer technologies. Novel and improved techniques and instruments for sensing and monitoring of human health conditions (new cDNA microarray techniques [2], new imaging modalities like Optical Coherence Tomography [1], or improved spatial resolution of microscopy imaging lead to massive amounts of raw data to preserve and retrieve information. Computer assisted medical decisions enable gathering provenance trails about the decisions [3] that lead to

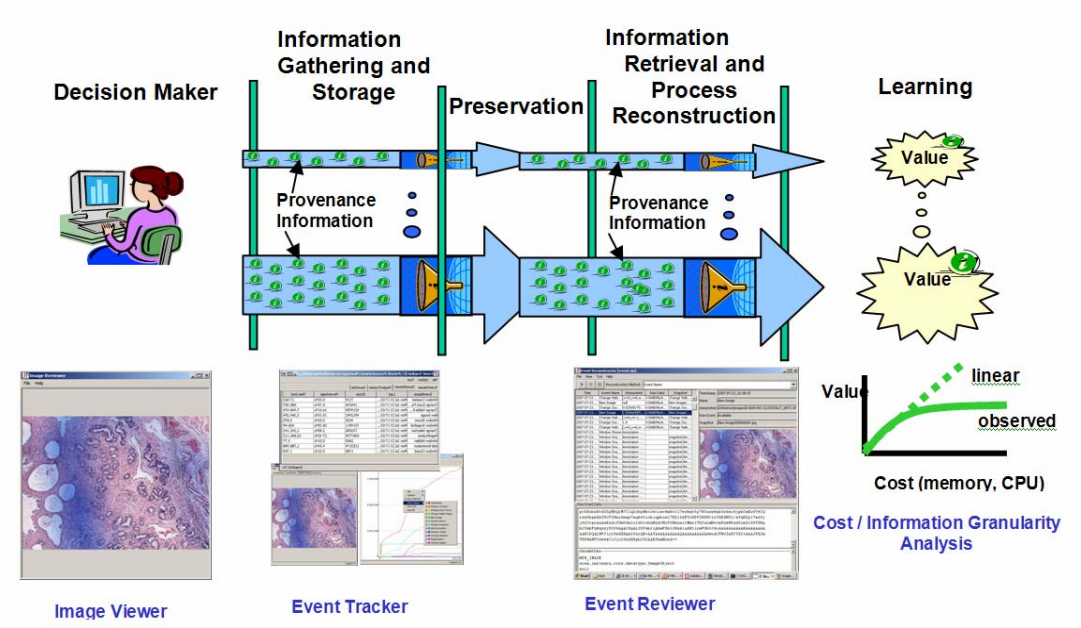


Figure 1. The overall architecture of a simulation environment.

a large volume of structured or unstructured metadata. Finally, rapid changes of media and technologies together with many paper-based records moving to electronic records are forcing us to think of preservation of hardware, operating systems and software, as well as the issues of security and authenticity [5].

Our approach to the above challenges is to design a simulation environment for optimizing institutional decisions about EMR management and preservation. For this purpose, we have prototyped a simulation framework called "Image Provenance To Learn" (IP2Learn) that is built for a class of medical decisions based on image inspections. The choice of this class of medical decisions was based on image properties (language independent, omnipresent, multi-spectral/multi-dimensional and frequent in many decisions). The current prototype enables to explore some of the EMR management and preservation tradeoffs related to (1) information granularity (category and level of detail), (2) representation of provenance information, (3) compression, (4) encryption, (5) watermarking and steganography, (6) information gathering mechanism, and (7) final medical report content (level of detail) and its format. The novelty of the simulation environment is that based on our knowledge there has not been developed such a decision support system for optimizing institutional decisions and appraising electronic records.

This paper presents the simulation framework architecture and the list of available simulation variables. The framework and the simulation variables enable understanding of computational requirements associated with preservation and reconstruction, as well as learning about activities during medical decision processes for education and automation purposes. In the experimental section, we illustrate how to analyze image inspection and annotation processes and would be the outcome of such analyses.

II. Simulation Framework Architecture

The simulation prototype consists of Image Viewer for visual inspection of images, Event Tracker for information gathering about the image manipulation operations, and Event Reviewer for information retrieval and decision reconstruction. The architecture of the simulation environment is shown in Figure 1.

Image Viewer captures supports the decision process and contains Image Frame and Image Panel, and provides all image manipulation functionality. Event Tracker tracks events in Image Viewer, allows setting preferences for information gathering and storage, lists event activities, summarizes event activities, displays inspected area and displays computational requirements. It also enables to generate reports with protected authentic information about decision processes, speeds up reporting and provides a foundation for tracking versions of reports. Event Reviewer retrieves gathered and stored information by Event Tracker, reconstructs processes with selected information granularity, and displays hierarchy of events and replays image inspection events according to the available information granularity. Event Reviewer serves as a tool for assessing the value of preserved and reconstructed information and for learning about spatial and temporal characteristics of activities during visual inspections.

III. Simulation Variables

We provide a list of simulation variables that are typically of interest when it comes to preservation. First, it is the information granularity (categories of information and the level of detail) of observable and measurable variables. In our system, there are three categories: (1) Interpreted– what the programmer encoded as a textual description (interpretation) of image operation, (2) Raw – what the computer work with when image operation was recorded, and (3) Snapshots – what was rendered on the computer screen at the time of image operation. For example, the word "zoom" belongs to the category of interpreted, Java event message reporting zoom is of type Raw, and a snapshot of zoomed image is of type

Snapshot. The level of detail corresponds to the description used for reporting, e.g., raw at the user program level or operating system level or hardware level. The current system reports information at the user program level.

Second, it is the information gathering mechanisms that defines how to gather information about computer system activities ranging from hardware to user program levels. In our system, there are mechanisms, such as (1) triggered by logging functions (checkpoint execution anywhere and everywhere), (2) triggered by events (consumer & producer model) and (3) triggered by Mouse/Keyboard Inputs (human-computer interface (HCI)).

Third, it is the information organization and compression. The current simulation software allows comparisons of meta-data organized as key pairs or following the resource description framework (RDF), and compressed as zip files or uncompressed.

Fourth, the simulation variable is the authentication of information by (a) encryption with AES (Advanced Encryption Standard) allowing only users with the private key to view files, (b) watermarking or (c) steganography. Watermarking allows to label files with visible textual signatures (e.g., for copyright purposes) while steganography enables hiding secret text of image into a file as a future authentication proof.

Fifth, it is the storage format of gathered information and of summary reports. The gathered textual information is stored following either xml or plain ASCII format and gathered images are stored either in tiff or jpg format. The summary reports can be stored in HTML (Hypertext Markup Language) or PDF (Portable Document Format) format and could be edited freely or with encrypted information that should not be tampered with during editing. The editing is important for simulating report versioning and how the content (level of report detail) impacts computational requirements.

Sixth, the simulation variable is the remote or local information retrieval. One can experiment with distributed or centralized location of gathered information and the associated cost during retrieval.

Seventh, it is the decision process reconstruction methods. In the current system, reconstruction methods of a decision process are presented by (a) displaying static time instances of gathered textual and image information, (b) dynamic replay of the activities during a decision process, or (c) visualizing all gathered information following RDF or key pair organization. The static and dynamic reconstructions allow inspections of not only gathered information but also reconstructed status of Image Viewer at any time instance according to the appropriate information granularity.

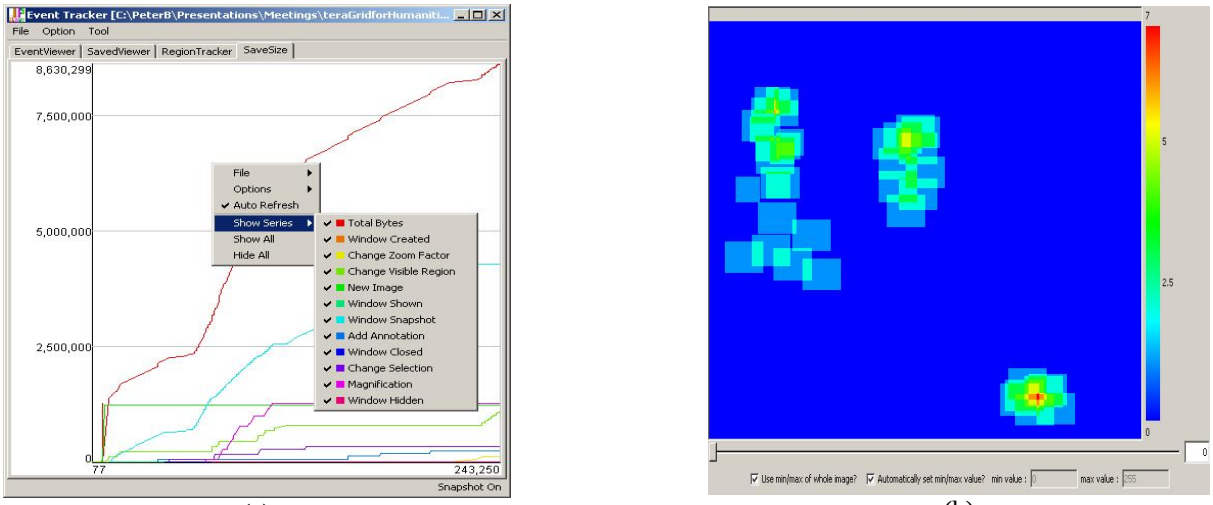
IV. Example Analyses

We have explored the case of annotating microscopy images of prostate cancer biopsies stained with H&E stain and imaged with a bright field microscope. We view the annotation process as a decision process in which the areas of interest are annotated and many other areas are inspected without annotating them. The images came from UIC patients[1] with recurring and non-recurring cases of prostate cancer. Therefore, it is imperative that the images be compared and analyzed over time either by humans or by machine learning algorithms.

In this study, we focused on assessing required computational resources required for preservation and reconstruction of annotated images, summary reports about patient's prostate biopsies and the annotation (image inspection) process, and hence helping with appraisal of the electronic records. Figure 2 shows one of the annotated images. Figure 4 summarizes the computational requirements of preserving the information about image inspection and annotation process. This summary is viewable directly in Event Tracker or inserted into automatically generated reports about the process. Based on the summary, the highest storage cost is for preserving snapshots (~30% - what was rendered on the screen), followed by the high spatial resolution image sub-areas viewed using the magnification operation (~29%) and by the New Image operation (~20.5% - loaded new image). Figure 3 (left) shows a graph of the storage requirements per image manipulation operation. Additional learning is enabled by providing the spatial distribution of activities during the inspection and annotation of the prostate cancer image in Figure 2 as shown in Figure 3 (right). The spatial or temporal aggregation of activities is conducted automatically and could be used for educational purposes as well.

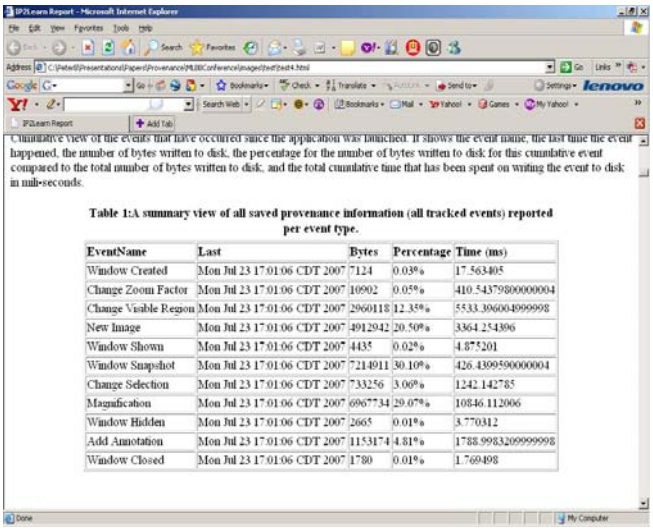
Figure 2: The non-recurring case of prostate cancer that was annotated using the IP2Learn simulation framework.

[1] The images used in our study are courtesy of Dr. Andre Balla from Pathology Department at UIC.



(a) (b)

Figure 3. (a) Visualization of storage requirements (vertical axis) over time (horizontal axis) in Event Tracker. Each colored line corresponds to one image manipulation operation. (b) visualization of the spatial distribution of activities during the inspection and annotation of the prostate cancer image in Figure 2.



in milli-seconds.

Table 1:A summary view of all saved provenance information (all tracked events) reported per event type.

EventName	Last	Bytes	Percentage	Time (ms)
Window Created	Mon Jul 23 17:01:06 CDT 2007	7124	0.03%	17.563405
Change Zoom Factor	Mon Jul 23 17:01:06 CDT 2007	10902	0.05%	410.54379800000004
Change Visible Region	Mon Jul 23 17:01:06 CDT 2007	2960118	12.35%	5533.396004999998
New Image	Mon Jul 23 17:01:06 CDT 2007	4912942	20.50%	3364.254396
Window Shown	Mon Jul 23 17:01:06 CDT 2007	4435	0.02%	4.875201
Window Snapshot	Mon Jul 23 17:01:06 CDT 2007	7214911	30.10%	426.4399590000004
Change Selection	Mon Jul 23 17:01:06 CDT 2007	733256	3.06%	1242.142785
Magnification	Mon Jul 23 17:01:06 CDT 2007	6967734	29.07%	10846.112006
Window Hidden	Mon Jul 23 17:01:06 CDT 2007	2665	0.01%	3.770312
Add Annotation	Mon Jul 23 17:01:06 CDT 2007	1153174	4.81%	1788.9983209999998
Window Closed	Mon Jul 23 17:01:06 CDT 2007	1780	0.01%	1.769498

Figure 4. The summary report in HTML format that was automatically generated for the image inspection and annotation process of the prostate cancer image in Figure 2. It shows the computational requirements for preserving information about the process.

We have conducted several tradeoff studies with the simulation variables listed in the previous section for simple inspection and annotation tasks. The representative results are shown below in Figure 5, Figure 6 and Figure 7. Based on the results, one could conclude that RDF representation requires more storage than key pair representation (as expected) and quantify the difference in file size (proportional to storage cost). In general, the information retrieval time (noted as "Average Response Time" in Figure 5(b)) is proportional to the "Saved Size" due to the disk I/O access speed. Based on the graphs in Figure 6, we concluded that the encryption time and file size are proportional to the size of input files. The file size change due to encryption/decryption was negligible. In terms of compression efficiency (compression ratio), compressing RDF representation was slightly more efficient than compressing key pair representation because the RDF representation includes more redundant information than key pair representation.

V. Summary

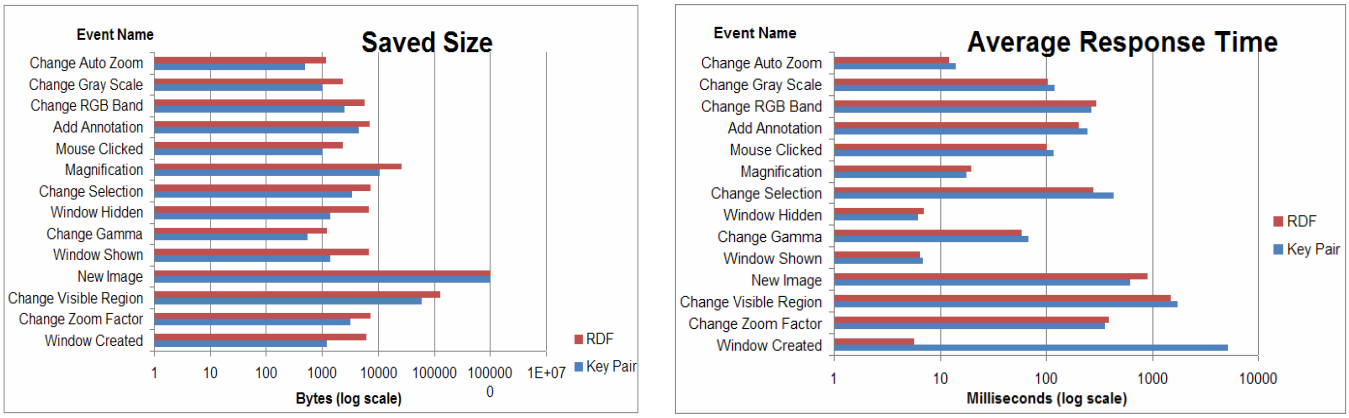
We presented a simulation framework for understanding preservation and reconstruction requirements for computer-aided medical decision-making using visual inspection and annotation. According to our knowledge, this is a first simulation framework for managers of electronic medical records (EMR), archivists and other users to simulate computational costs (e.g., storage and processing time) depending on several key characteristics of appraised records. Our effort led to a prototype called "Image Provenance To Learn" (IP2Learn) that is freely available for downloading at http://isda.ncsa.uiuc.edu/downloads. The ultimate goal of our research is to understand the cost of long term preservation of medical electronic records using the cutting edge technologies, high performance computing and novel computer architectures. In addition, the current framework will enable us to address several machine learning problems over a larger time period in the future.

VI. Acknowledgement

This research was partially supported by a National Archive and Records Administration (NARA) supplement to NSF PACI cooperative agreement CA #SCI-9619019. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the National Science Foundation, the National Archive and Records Administration, or the U.S. government.

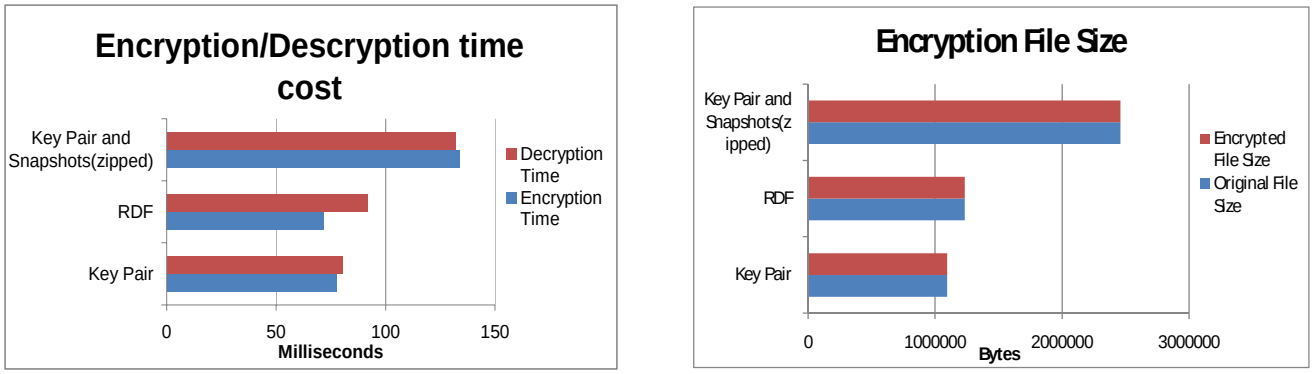
VII. References

[1] Boppart SA, Brezinski ME, Fujimoto JG. "Optical Coherence Tomography Imaging in Developmental Biology." In Methods in



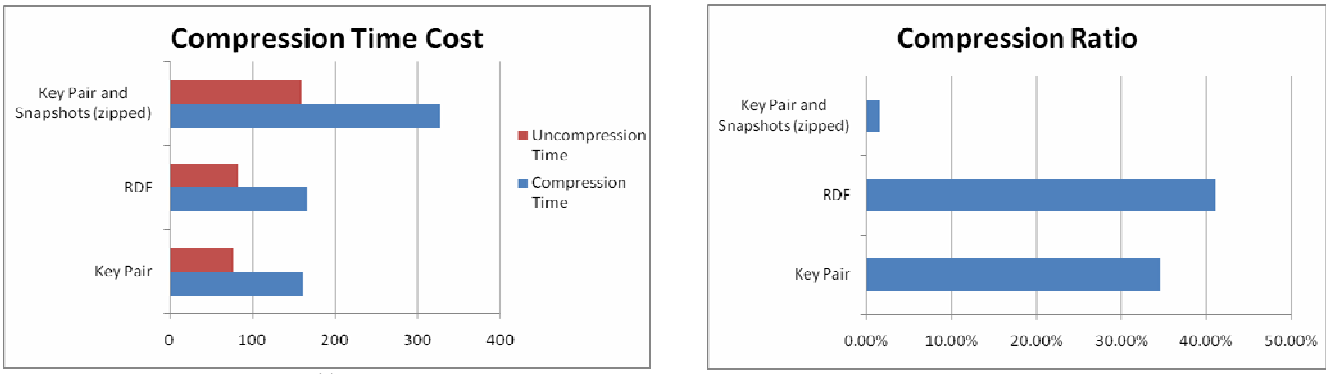
(a) (b)

Figure 5: Comparison between key pairs and RDF format for event preservation in terms of (a) storage requirement and (b) average information retrieval speed.



(a) (b)

Figure 6: Encryption/decryption cost analysis. (a) shows the time cost and (b) refers to storage cost.



(a) (b)

Figure 7: Compression cost analysis. (a) shows the time cost of compression and (b) can be mapped to storage cost.

Molecular Biology, Vol. 135: Developmental Biology Protocols, Volume 1. Tuan RS, Lo CW, Eds., Humana Press, Inc., Totowa, NJ, 2000.

[2] Bajcsy P., J. Han, L. Liu and J. Young, "Survey of Bio-Data Analysis from Data Mining Perspective," Chapter 2 of Jason T. L. Wang, Mohammed J. Zaki, Hannu T. T. Toivonen, and Dennis Shasha (eds.), Data Mining in Bioinformatics, Springer Verlag, 2004, pp.9-39.

[3] Lee Y-J. and P. Bajcsy, "An Information Gathering System For Medical Image Inspection," Proceedings of SPIE Conference on Medical Imaging, Vol. 5748-48, 12-17 February 2005, San Diego, CA.

[4] R. W. Moore, J. F. Jaja, R. Chadduck Mitigating Risk of Data Loss in Preservation Environments, Proc. of the 22nd IEEE/13th NASA Goddard Conference on Mass Storage Systems and Technologies (MSST 2005)

[5] Abby Smith, Authenticity in Perspective, Council on Library and Information Resources (CLIR) meeting on January 24, 2000, p 69-75, ISBN 1-887334-77-7 http://www.clir.org/pubs/reports/pub92/smith.html

[6] Charting a course for the 21st century – National Medical Library's long range plan 2006-2016 http://www.nlm.nih.gov/pubs/plan/lrp06/report/executivesummary.html

[7] David Shotton, "The nature of Biomedical Research Data," Research Information Network Workshop, on December 5, 2006, Royal institute of Public Health, London, URL: HTTP://WWW.RIN.AC.UK/FILES/SHOTTON - NATURE OF BIOMEDICAL RESEARCH DATA.PDF (last visited: July 22, 2007)

[8] STEWARDSHIP OF DIGITAL RESEARCH DATA - PRINCIPLES AND GUIDELINES, RESEARCH INFORMATION NETWORK, APRIL 2007, URL: http://www.rin.ac.uk/files/Research Data Principles and Guidelines – published draft for consultation.pdf (LAST VISITED: JULY 22, 2007)

[9] Medical Research Council (MRC) Data Sharing and Preservation Initiative, URL: http://www.mrc.ac.uk/Utilities/Documentrecord/index.htm?d=MRC003346