Rendering an archive in three dimensions

David A. Leiman[a], Claire Twose[b], Teresa Y.H. Lee[c], Alex Fletcher[d], Terry S. Yoo[e]
[a]Johns Hopkins University, 3400 North Charles Street, Baltimore, MD, USA 21218 [b]Johns Hopkins University School of Medicine, 98 North Broadway, Baltimore, MD USA 21231 [c]University of British Columbia, 2329 West Mall, Vancouver, BC Canada V6T 1Z4 [d]Howard University, 2400 Sixth Street, NW, Washington, DC USA 20059[e]National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD USA 20894

ABSTRACT

We examine the requirements for a publicly accessible, online collection of three-dimensional biomedical image data, including those yielded by radiological processes such as MRI, ultrasound and others. Intended as a repository and distribution mechanism for such medical data, we created the National Online Volumetric Archive (NOVA) as a case study aimed at identifying the multiple issues involved in realizing a large-scale digital archive. In the paper we discuss such factors as the current legal and health information privacy policy affecting the collection of human medical images, retrieval and management of information and technical implementation. This project culminated in the launching of a website that includes downloadable datasets and a prototype data submission system.

Keywords: Three-dimensional Volume Image Database, National Online Volumetric Archive

1. INTRODUCTION

Among the advances made possible by the growth of the Internet, accessible online databases have drawn increasing interest. This has been encouraged by a larger emphasis on data distribution, exemplified by the number of submissions to data archives like the National Center for Biotechnology Information's (NCBI) Genbank[1] and the recent release of the National Institutes of Health's (NIH) draft policy on data sharing[2]. Image databases, in particular, have been at the focus of this trend. Previously, however, these databases have dealt mainly with two-dimensional images. With the abundance of less expensive and faster hardware-oriented volume-rendering tools, three-dimensional images are becoming more feasible to use in practice. A need to store and distribute these data has therefore developed.

The primary users of a volumetric medical data archive include researchers (imaging specialists, computer scientists, anatomists, etc.), clinicians and educators. A diverse collection would allow comparative study of biomedical specimens and subjects. A broadly based biomedical archive could promote cross-disciplinary research and collaboration among scientific teams with a variety of expertise.

The goal of this project was to examine the requirements of creating a publicly accessible, indexed online collection of three-dimensional biomedical image data. We created NOVA as a case study in digital volume archives. Among the factors considered were: the features of current legal and health information privacy policy affecting the collection of human medical images; retrieval and issues in management of information; and, requirements for technical implementation.

2. BACKGROUND

Current medical imaging devices generate series of two-dimensional data "slices" that, when stacked, comprise a volumetric representation of the scanned subject. Computers can be used to re-interpret this data and produce perspective renderings of the volume data, giving the illusion of a three-dimensional representation of the subject[3]. The advantages these models provide over traditional two-dimensional views include easier recognition of spatial relationships and assistance in visualization. Secondary benefits may range from more precise surgical techniques to more diverse education methods[4].

Figure 1: A two-dimensional CT scan of an aneurism (left) and a three-dimensional version with aneurism highlighted (right) [image courtesy of www.volvis.org].

Despite the abundance of raw radiological data, there are relatively few Web-accessible databases that provide open access to such information; moreover, available data is often hard to use. Frankewitsch and Prokosch reviewed 48 image archive sites on the Web and reported on the methods of navigation used in these sites[5]. Only three sites they found index their datasets using a controlled vocabulary. In most instances the interface is cumbersome and limited. Thus, the availability and access to image data is limited. Existing sites, like the Bristol Biomedical Image Archive focus primarily on "flat" images. While the number of existing biomedical image archives being developed for teaching or research is large, proportionately few focus specifically on three-dimensional modalities[6]. Of those that do house volume images, many do not have a distinct organization or indexing structure; they are more akin to bulletin boards that post data.

Additionally, these sites generally house specific datasets, often focusing only on in-house research. Even comprehensive attempts at storing such images, like the University of Iowa College of Medicine's VIDA® site, contain only images from their hospitals. Yet, research organizations are taking up aggressive new work in medical image analysis, computer assisted diagnosis, computer assisted treatment planning and in the modeling and simulation of human anatomy and physiology. These new frontiers require copious amounts of data, organized and collected not only by demographics but also by diagnosis, prognosis, treatment and outcome. In addition, the emerging fields of medical visualization and simulation will also require supplemental data including microanatomy, physiology and practical data upon which new presentation techniques can be designed and tested.

In particular, researchers could benefit from a common source of datasets from which educational and research materials can be drawn. The availability of a common source would facilitate comparisons of diagnostic and analytic techniques. Teaching files created from common data sets can explore a single salient case at a variety of depths of interest and expertise. The National Library of Medicine (NLM) is in a unique position to undertake this project with its ongoing work on the Visible Human Project, the large volume of data available to it and its ability to function as a host of large amounts of data.

There are many tools being developed that can be brought to bear on this problem. In order to achieve a functional archive, rather than an unorganized repository of image data, this project identified the need to formulate a systematic schema of metadata associated with the image data itself. No current standard appropriately fit the scope of this project and thus a review of the current legal, design and technical issues surrounding this project was undertaken.

3. RESULTS

3.1 Legal Issues

Before the acquisition of any datasets, it was important to research the legal issues surrounding such a site. These considerations are shaped by the concerns of the various parties involved as data donors, data users and information organizers. Data donors will be interested in adequate copyright protection for the image sets they provide, in addition to possibly needing guidance on current privacy regulations concerning the subjects of their images. Similarly, data users

will need to be aware of copyright and fair use issues in using the information provided in the volumetric archive. As information organizers, the overseers of this project need to be sensitive to the varied, and perhaps even at times conflicting, interests of subjects, donors and users.

Accordingly, those issues addressed related to ownership of the data and patient protection, *i.e.*, copyright and licensing issues and obtaining permission to use the images and protecting patient privacy and confidentiality, respectively. Recent changes in the law governing copyright affect digital media. New privacy regulations planned for April 2003 add an additional layer of protection for patient records including image data with specific implications for databases and archives. Supporting ongoing collection of these digital datasets will require the creation of a new framework to ensure that the relevant legal and ethical issues are addressed.

Since NOVA is intended as a repository and distribution mechanism for volume datasets, among the most important issues are those of copyright and fair use of the data. In the common case of radiologists in the employment of hospitals, the copyright belongs to the institution[7]. Thus, a differentiation between data owner, producer and copyright holder exists. And, in order to acknowledge or regulate the fair use of such ownership, a set of guidelines must be created. Because no definitive fair use guidelines exist, however, their rather arbitrary nature demands the creation of licensing agreements. In following with NOVA's threefold mission scope, to include information useful to professional educators, researchers and clinicians, any agreement would suggest the fair and proper use of data for appropriate purposes. In order to facilitate both the ease of registration into the NOVA system and proper use afterwards, a web-based click-through agreement for both data set providers and data users is acceptable; copyright and privacy issues can be rolled into one license in order to simplify the submission and access processes[6]. In this way, both the veracity of the submitted data as well as the integrity of the users who are downloading it could be assured. The work of drafting the points to be covered and then having legal counsel review and create an actual license document remains outstanding.

Just as sensitivity to the image owners is important, so, too, is the privacy of the imaged subject. The broad nature of NOVA's collection, including all biomedical data, provides for human data. As such, a number of levels of protection must be observed, including the Common Rule and the more recent and specific de-identification guidelines of the Privacy Rule found in the Health Insurance Portability and Accountability Act's (HIPAA) regulations[6]. Implied in these guidelines are standards by which the subject is properly protected from invasion of privacy. Nonetheless, some information about a patient may be critical to understanding the image itself, particularly when used in a clinical or research setting. A balance was struck using legally allowed patient information and practically useful meta-information, *i.e.*, information about the data. This balance was then applied to create a unique set of data fields required to be completed along with data donation.

3.2 Design

The NOVA system was designed for a two-fold audience, data providers and data users. Although publicly accessible through the National Library of Medicine site, the range of anticipated users is narrow. Thus specific aspects and functions were integrated into the design with sophisticated users in mind. NOVA resides on a server at the Office of High Performance Computing and Communications and can be viewed at http://visual.nlm.nih.gov/evc/programs/nova/index.html.

3.2.1 Metadata elements

These data fields take the form of a set of descriptive tags that describe the image data. The information for these data fields is to be submitted concurrently with the image data itself. A primary goal of this project was to implement a feasible method by which to collect image data and its associated metadata; thus, the list of elements was created to maximize relevant information submitted while keeping the submission process as concise as possible. In this model, a donor provides information describing the archival, technical and medical features of the dataset, while concurrently uploading the data.

Current government regulations limit the type of data that may be used to identify a patient that can be included in non-patient care environments. The HIPAA[8] states that a patient must be completely de-identified, prohibiting references to medical record identification numbers, biometric identifiers like fingerprints and other distinguishing characteristics.

This project devised a schema of metadata that would properly comply with federal regulations while at the same time providing enough information about the dataset to be useful to those who might use or recreate the image.

Existing standards were consulted while developing the metadata scheme, such as DICOM and NCI's draft data elements for clinical trials[9, 10]. However, three-dimensional volume image storage is relatively new, so there is currently no single applicable standard. Metadata for each NOVA image dataset was sectioned into three parts, including archival (*e.g.*, the image copyright holder), technical (aspect ratio) and medical information (subject condition) tags. It is believed that this would provide the infrastructure on which to build a future searching tool as well as reasonable links between data. Future directions for the development of the metadata scheme include adding tags that will track the relationships between collections of images over time and modality of scan.

The level of detail in each of the three categories was based on perceived needs of the intended audience. For example, an educator looking for exemplary models of a particular type of subject may be interested in the demographic information, but less so on technical information related to it. Conversely, a researcher designing a program to identify lung cancer from CT scans will need to know not just the type of cancer but the number of nodes, their sizes in three dimensions, and where they are located in the image to evaluate the computer program's performance. For the purposes of developing the NOVA prototype system, then, we tailored our development of metadata to address the needs of researchers working in image registration and normalization. Thus we included basic demographic, treatment, diagnosis and outcome elements and more extensive technical ones.

3.2.2 Image format

Another factor influencing the inclusion of data was the file format itself. Just as there is no single standard governing metadata, there are numerous image file formats available, proprietary or not. Due to the large number of formats, it would be unrealistic to make data ubiquitously compatible. Therefore, a consistent internal standard was accepted.

Within NOVA, a user can first preview an image of the dataset and then download it. It was believed that users browsing through the archive would benefit from an initial image of the data before downloading it; these preview images, according to the Library of Congress, "allow users to judge whether they wish to take the time to retrieve a higher quality image" or continue searching[11]. These thumbnails are rendered versions of the data and are displayed either as JPEG or GIF files. These formats were chosen because of their prominence in Internet use. They also allow for reasonable download times and can provide an accurate representation of the data.

Internal standards were also applied toward the file format of the image data itself. Only raw data and TIFF files would be indexed in NOVA. These formats were chosen to provide lossless versions of the data, an important feature for medical images that require uncompromised detail. Raw data provides a format "free of the artifacts resulting from lossy compression"[6] while TIFF "is designed to promote the interchange of digital image data."[12] Thus, an "original copy" of the data would be available as well as a widely used and compatible second format.

3.2.3 Data submission

A prototype submission system was developed based on the metadata and image file format requirements. The Metadata and Image Registry (MIR) is the module through which data and metadata will be submitted and incorporated into NOVA. Designed to function in a linear manner, a user can only go forward or backward. This design was employed because "straight sequences are the most appropriate organization for training sites [where] the reader is expected to go through a fixed set of material and the only links are those that support the linear path."[13] Currently, however, MIR is a manual submission system that serves only for demonstrative purposes. Mimicking the future data entry process, this tutorial is the outline for the automated system.

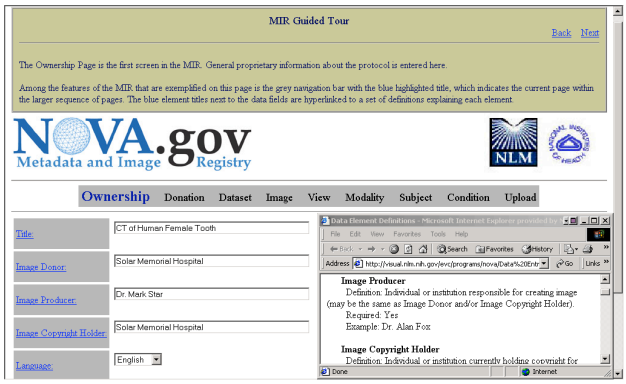

Figure 2: A screenshot of the Metadata Image Registry entry screen with Data Element Definitions present.

Anticipating a potential overlap of users, MIR is modeled on the Protocol Registration System (PRS) for *ClinicalTrials.gov*. It explicitly outlines the metadata required for submission while at the same time alerting the user to which information is required and which is optional. The metadata elements are sectioned into categories within MIR and limited to no more than six elements per page, thereby reducing the load on any given page and aiding the data enterer in submission[14]. Additional consideration was made for users who may range in familiarity with the data and its associated information. In anticipation of this need, each element title within MIR links to a set of comprehensive Data Element Definitions (DED). These DED open in a separate window and provide both a description of the element as well as a sample submission.

Presently, MIR serves only as a reference template. In its current form, it is a tutorial for potential users of the system, providing walk-through descriptions of each screen. Nonetheless, MIR does provide the framework for manual submission while at the same time laying the foundation for a future automated system.

3.2.4 Data pages

Once the data submission system was designed, end pages were compiled using test image datasets. After a search is conducted, a user arrives at a results page that lists all returned hits. Furthermore, these result pages allow users to scroll through various related datasets returned by the search and provide a chance to start a new search. This page includes thumbnail images of the data as well as a short description. After choosing a dataset, a link connects the user to the complete entry view.

Currently populated with data in the National Library of Medicine's collection, including the Visible Human data, these pages are the body of the NOVA system. They provide users with a two-dimensional preview image of the three-dimensional rendered volume as well as a short animated sequence that further displays the volume. Additionally, a set of all legally allowed metadata associated with the particular image is listed next to the image along with a link to download the dataset. These pages provide a chance to view the data before downloading it; the metadata also acts as a set of directions by which to load and render the volume once the data has been downloaded.

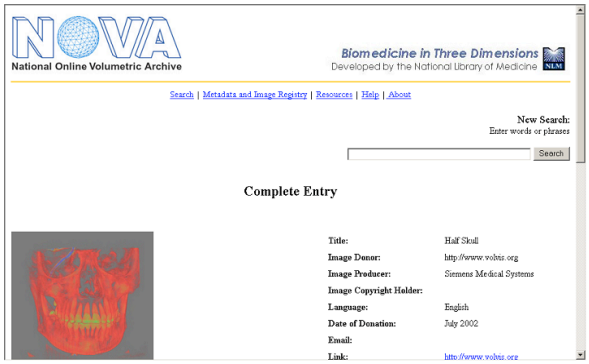

Figure 3: A complete entry view in the National Online Volumetric Archive displays a preview of the image and its metadata.

3.2.5 System Flow

After the metadata and data standards were chosen, a system flow was organized to link MIR submission with the end page data. In this model, the ultimate goal for both the data user and provider is the end pages. To this end, there are two pathways.

Once a data donor decides to submit a volume image to NOVA, an account is set up to verify the veracity of the data. This account provides the donor with an authorization and password and permission to use the MIR submission system. Once the data is loaded via MIR, it is dynamically generated into web pages that are populated with the data.

A data user proceeds via a similar mechanism. Entering through the public homepage, the user searches for an image set of interest. Once one has been located, a click-through agreement is displayed to ensure proper use of the data. After accepting these policies, the user gains access to the data pages.

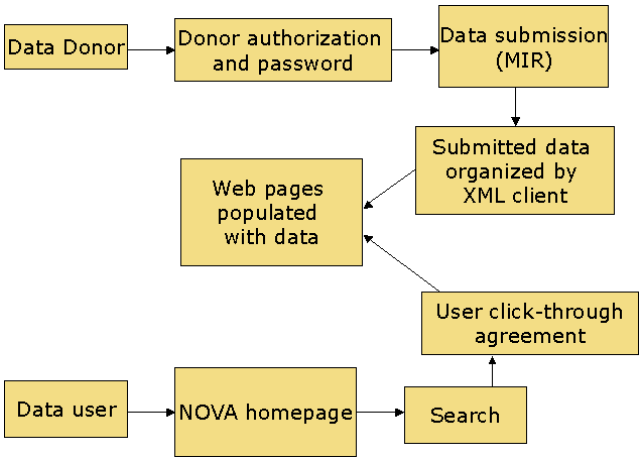

Figure 4: The system flow design for the National Online Volumetric Archive.

A main feature of this design is the site's homepage. Functioning as the public face of NOVA, the homepage was also developed to function as the hub for NOVA; it serves as the node from which to store and access the received data was needed. The homepage is the terminal to both the database and MIR.

The homepage features a simplified layout. Its appearance includes a themed logo that was adopted and applied to the homepage as well as being affixed to all other pages within the site. This creates a unified sense of domain. While graphics are minimally used throughout the site, along with a shaded navigation bar, they do provide sufficient contrast to aid in navigation[14].

The functional aspects of the site include its links to the various areas within the site, including creating a donor account, image viewing resources and links to information about the site. At the heart of the page, however, is its most important feature, the searching utility. Along with providing a variety of levels of searching, this tool is the passport to the rest of the data in the site. While this feature is not currently operational, an XML client was modeled that will eventually link the submission process to the search tool; this has the bimodal advantage of streamlining the acquisition and subsequent retrieval of data.

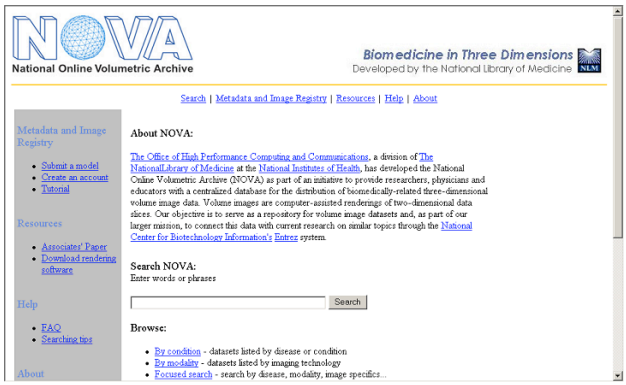

Figure 5: The National Online Volumetric Archive homepage with navigation bar and searching tools.

3.3 Indexing, Search and Retrieval

The searching system's main attribute is its ability to seamlessly link a user to the results page. This is made possible by the implementation of Image Content Groups (ICG). This is a concept that allows automated theoretical grouping of images based on metadata. An ICG is a well-formed XML document with numerous nodes and children that establish relationships between the metadata of image files. These relationships are automatically construed by their makeup and structure within the ICG.

Currently, there exists a large number of options for the storage and retrieval of metadata over the Internet. As the Internet primarily functions as a text-filled domain, however, the current mechanisms usually rely on an abundance of plain text or simple metadata entries. Even the best searching mechanisms like Google work through their use of textual relationships and are constrained by these requirements; similarly, most image databases use textual descriptions of the images like the file name to index them[15].

For example, when a search is conducted using the word "heart" in Google, it generally identifies all indexed documents that contain the word "heart" and sorts them according to how many times the word occurs. This type of search and retrieval system, while efficient for documents, does not work as well in a situation where it is necessary to interpret relevancy without the relying on recurrence of text, *e.g.*, in a three-dimensional image archive. In addition to the images themselves lacking a true description, using a system like this in NOVA would essentially eliminate the possibility of establishing any type of synonymy; as the medical lexicon includes so many closely related taxonomies, a searching system that does not return "heart attack" when "myocardial infarction" is entered is not operating efficiently. Thus, NOVA presents a scenario in which there is not a profusion of text to index. A need arises to build relations from something else to extract those crucial relationships and hierarchies with fewer words.

Thus, Image Content Groups provide a useful alternative. They allow automated decision making capabilities in connection with the relationship between the metadata and its related images. Used in conjunction with an already established structure like the NLM's Medical Subject Headings (MeSH), ICG can also assemble relationships based on synonymy in an "is-a" relationship. An ICG serves as a thin client between an interface and back-end data storage, consisting of nested groupings of relationships based on selected metadata entries. A given assemblage will consist of a <Field> parent node, which contains an attribute string indicating the corresponding field number that groups every child element. Inside the <Field> node there can be any number of <Group> nodes, each defined by a <Mesh> node, which gives a listing of all synonymous MeSH vocabulary terms. The <Documents> node contains references to all related documents fit with a <descript> providing its short description. One of the key aspects of these groups is that any number of images can belong to any number of groups, supplying invariable flexibility in relating different characteristics of images to others. This flexibility affords a tremendous increase in terms of the reliability of a search query because it relies less on textual representations and more on the implied relationships between images. Therefore, the relevancy of an image has more to do with its similarities to other relevant images as opposed to the proliferation of a given text value. This option prevails where relationships must be illustrated with text at a premium.

4. FUTURE WORK

NOVA represents a case-study in the development of a larger scale archive, and there are a number of features that remain under construction. The website has been created and more datasets are currently being acquired. With the addition of these datasets the XML client can be tested and employed by the site. The automated submission system and the online registration system for data suppliers and data users will be developed as the need for processing and presenting large numbers of new image datasets arises. These systems will serve to authenticate the donors' data and ensure that the data is only being used for appropriate reasons.

As the content of the site continues to grow, another avenue of exploration will to link image data with text in the NLM's bibliographic database, PubMed. Options include using the LinkOut feature and incorporating the archive into the existing Entrez system, providing visual links related to current research.

5. SUMMARY AND CONCLUSIONS

We discovered that related projects are usually smaller collections with limited scope; a small repository does not require an elaborate index and many informatics issues were not raised in previous image archiving attempts. We studied the problems of creating controlled vocabularies for indexing our data and identifying a limited set of common data elements for data retrieval. We also attempted to identify and include critical metadata necessary for visualizing volume datasets. We learned that the privacy and legal issues are difficult to navigate and keep changing over time, but they are manageable.

This project culminated in the launching of a website, NOVA, which includes downloadable datasets and a prototype data submission system, MIR. The metadata schema was created based on a comprehensive summary of technical issues and the current legal environment surrounding medical patient data release.

At this time, the site relies on manual submission and update of image data fields. It is anticipated, however, that with the expansion of the current site as well as the installation of additional components, a fully automated system will exist in the future. An XML client will eventually link the submission process to the search tool and help streamline the acquisition and subsequent retrieval of data. By then a fully automated system, it will be able to dynamically adapt to changes in the groupings and apply these relations seamlessly across platforms. At present, the NOVA system provides the foundation for and the impetus to continue production of a centralized and openly accessible repository of three-dimensional volume images data.

REFERENCES

1. A. Baxenanis and B.F.F. Ouellette, *Bioinformatics*, J. Wiley, New York, 1998.
2. *NIH data sharing policy*, National Institutes of Health Office of Extramural Research, March 2002. Available from http://grants1.nih.gov/grants/policy/data_sharing/.
3. A. Kaufman, D. Cohen, R. Yagel, "Volume graphics," *IEEE* Computer, 26(7), 51-64, July 1993.
4. M.J. Ackerman, "The Visible Human Project," *Proc. IEEE*. 86(3), 504-511, March 1998.
5. T. Frankewitsch and U. Prokosch. "Navigation in medical Internet image databases," *Med Inform Internet Med*, 26(1), 1-15, 2001.
6. T. Lee and C. Twose, *The NLM Public Volume Image Data Repository: Issues in Planning an Online Image Archive*, The National Library of Medicine, August 2002.
7. United States Copyright Office, "Copyright Basics," *Circular 1,* U.S. Government Printing Office, July 2002. Available from http://www.copyright.gov/circs/circ1.html.
8. Department of Health and Human Services, "Health Information Portability Act of 1996," U.S. Government Printing Office, 1996. Available from http://cms.hhs.gov/hipaa/hipaa1/default.asp.
9. National Electrical Manufacturers Association, *Digital Imaging and Communications in Medicine*, National Electrical Manufacturers Association, Rosslyn, VA, 2001. Available from http://medical.nema.org/dicom.html.
10. National Cancer Institute. "NCI Data Elements." January 13, 2003. Available at http://ciiserver5.nci.nih.gov:8080/pls/cde_public/cde_java.show.
11. C. Fleischhauer, "Pictorial Materials," *Digital Formats for Content Reproduction*, Library of Congress, July 1998. Available at http://memory.loc.gov/ammem/formats.html.
12. The University of Edinburgh Computer Science Department, "Tag Image File Format," Edinburgh, April 1987. Available at http://www.dcs.ed.ac.uk/home/mxr/gfx/2d/TIFF-4.txt.
13. S. Horton and P. Lynch, *Web Style Guide: Basic Design Principles for Creating Web Sites*, Yale, New Haven, 1999.

14. M.K. Evans and N. Finck, “An interview with Dr. Jakob Nielsen, usability expert,” *Digital Web Magazine,* January 5, 2003. Available at http://www.digital-web.com/interviews/interview_2002-11.shtml.
15. S. Brin and L. Page, “The Anatomy of a Large-Scale Hypertextual Web Search Engine,” World Wide Web conference, Brisbane, Australia, April 14-18 1998. Available from www7.scu.edu.au/programme/fullpapers/1921/com1921.htm.

DLeiman1@jhu.edu; phone 1 410 366 8582; fax 301 469 0586