Image Retrieval Benchmark Database Service:
A Needs Assessment and Preliminary Development Plan

A Report Prepared for the Council on Library and Information Resources and the Coalition for Networked Information

Jennifer Trant
Archives & Museum Informatics
www.archimuse.com

Original draft: October 1, 2003; Updated January 2004

Table of Contents

Acknowledgments

This study was initiated as a joint project between the Coalition for Networked Information (CNI) and the Council on Library and Information Resources (CLIR) and was funded by the Atlantic Philanthropies. It has benefited from the thoughtful groundwork of Clifford Lynch (CNI) and Anne Kenney (then CLIR, now Cornell University Library).

The problems of assessing image retrieval were explored by the participants in several planning meetings held prior to the commissioning of this report. One such session, entitled "Planning Meeting for Test Database for Digital Visual Resources" and convened by Clifford Lynch and Anne Kenney in May 2001, was particularly helpful in shaping my initial thinking. Participating in this session were Joseph Bush, director, Solutions Architecture, Interwoven; Don D'Amato, MITRE Corporation; Corinne Jörgensen, School of Informatics, Florida State University; Donna Harmon, Text REtrieval Conference (TREC)/ National Institute of Standards and Technology (NIST); Peter Hirtle, Cornell University Library, Cornell University; Matthew Kirschenbaum, Department of English, University of Kentucky; Max Marmor, Arts Library, Yale University (now with ARTstor); Worthy Martin, Department of Computer Science, University of Virginia; Beth Sandor, University Libraries, University of Illinois; Don Waters, The Andrew W. Mellon Foundation; and John Weiss, Digital Library Production Service, University of Michigan.

I would also like to thank all those with whom I've discussed this question. I am particularly grateful to Margaret Graham, Corinne Jorgensen, Anne Kenney, and James Wang, who commented on drafts of the manuscript.

My personal thanks to David Bearman for his insight into our discussions of this and many other problems in the digitization of cultural heritage information.

1. EXECUTIVE SUMMARY

The rapid increase in the quantity of visual materials in digital libraries—supported by significant advances in digital imaging technologies—has not been supported by a corresponding advance in image retrieval technologies and techniques. Digital librarians sense that much could be done to improve access to visual collections and hope, perhaps vainly, that users' needs to identify relevant digital visual resources might be met more satisfactorily through search strategies based on visual characteristics rather than on textual metadata associated with the image, which are expensive to produce. However, digital librarians currently have no tools for evaluating either content-based or metadata-based image retrieval systems. Consequently, they have difficulty assessing existing systems of image access, evaluating proposed changes in these systems, or comparing metadata-based and content-based image retrieval.

Some have proposed benchmarking as a solution to this problem. An image retrieval benchmark database could provide a controlled context within which various approaches could be tested. Equally important, it might provide a focus for image retrieval research and help bridge the significant divide between researchers exploring these two search paradigms: metadata-based vs. content-based image retrieval. If so, such a database could spur advances in research, as comparative results make it possible to evaluate the effectiveness of particular strategies and thereby add value to studies supported by many funding agencies.

Creating an image retrieval benchmarking service would be a significant undertaking. A benchmarking database is more than a collection of images. Benchmarking requires a set of queries to be put to that test collection. Each image in the test collection must be assessed to determine whether it is relevant to that query. Assessing the performance of systems requires a set of evaluation metrics that make it possible to compare one system with another and to rank results. Developing a test collection requires an investment in data collection,

documentation, enhancement, and distribution. Most significantly, maintaining an image reference benchmarking service requires that a community of researchers make a long-term commitment to its use. Without a community vested in the development of the database—and publishing research based on it—the collection remains a chimerical solution to advancing the state of research and improving the retrieval of visual materials in the digital library.

2. PROBLEM STATEMENT

2.1 Images in Digital Libraries

Digital libraries—managed collections of digital information assembled and curated by institutions (such as libraries, archives, and museums) or individuals and made available for use—are complex, hybrid environments in which many kinds of materials are brought together for the first time. Their promise of integrated access to information is not being fully realized because of the different ways in which apparently similar materials are currently described. As digital librarians—those professionals responsible for the creation, management, maintenance, and provision of access to digital libraries—struggle to improve services to their traditional users, they also feel pressure to increase the use of their collections in new communities whose needs are less well known and understood. The desire to make collections accessible to nontraditional users makes clear the need for measures of success in information retrieval and evaluation of the user experience.

Images appear to offer great potential for interdisciplinary use. Researchers concur that retrieval of images is going to be increasingly important for a range of commercial, governmental, and academic purposes. They also concur that large aggregations of images, measured in hundreds of millions of images and picabits of data, will soon be a standard searching target. Space science, medicine, trademarks, and patents are far along in the implementation of large-scale image databases. Scholarly databases of art and cultural images are still in their relative infancy but are growing fast (a comprehensive cultural heritage database would comprise many hundreds of millions of images). New applications are emerging in

home-based image services addressing the hundreds of millions of personal images created with digital cameras.

Collections of visual materials bring the information retrieval problem into sharp focus. Resources originally created for use by a single department, such as art history, become a resource for users from many departments when digital image collections are made available on the campus network. Ironically, the potential for broad use of images is often frustrated by the same highly developed disciplinary descriptive and indexing systems that make them especially useable within a specialist community. One could imagine the same images being of interest in the creative arts, art history, the humanities (including history, languages, and philosophy), the social sciences, and anthropology. But for the historian, the artist-focused organizational systems of art history do little to surface the subject matter of images. For the cultural theorist, taxonomies in a biological database documenting an exploration do little to identify images depicting early cross-cultural contact. Scientific imagery is also growing greatly in volume, as satellite, meteorological, biological, and geographical images are gathered that document current and past conditions in detail; however, their electronically captured metadata do not identify the things depicted, making it difficult for the historical geographer to find a road's site using its name. These image collections are themselves of interest to computer scientists and those in information retrieval as well as to earth and space scientists.

Visual information has developed a significant role in our culture, and, it appears, in our research methods (Rhyne 1995, 1996). As more and more information resources are made available, their use and reuse become more difficult to predict and to assess. Recent surveys of image retrieval make the point that the users of such systems are drawn from many disciplines. Those cited by Venters and Cooper (2000) include "art galleries and museum management; architectural and engineering design; interior design; remote sensing and earth resource management; geographic information systems; scientific database management; weather forecasting; retailing; fabric and fashion design; trademark and copyright

database management; law enforcement and criminal investigation; picture archiving and communications systems." Known users of the Art Museum Image Consortium (AMICO) Library range from art-history researchers, to information and computer scientists, to language teachers and other individuals at the graduate, undergraduate, and K–12 levels, to lifelong learners. E-commerce audiences for image access to product catalogs are growing rapidly; they are just one of many areas of heavy image use outside the research and educational communities.[1]

In framing a study of image use at Pennsylvania State University, Henry Pisciotta asks, "Are all picture collections now interdisciplinary?" (2001). If we accept this rhetorical statement, how does it complicate our challenge to provide access to collections for all users?

2.2 The Issue for Digital Libraries

As the content of digital libraries increasingly varies in form and grows in size, librarians more and more frequently wish that they could reliably move beyond text-based retrieval. Those developing digital library services (expressed in the meetings leading up to this report) fervently hope that user needs for identifying relevant digital visual resources might be met through search strategies based on visual characteristics rather than solely on those represented in textual metadata associated with the image. This hope is in part born from a frustration with current access methods and in part a reflection of the presumed cost of creating image metadata for retrieval, even if metadata-based image retrieval did work. Even among specialist librarians, there is a sense that much could be done to improve access to visual collections, both in the use of existing description and indexing schemes and in the application of new technologies (Eakins and Graham 2000). Perhaps the solution is to match the retrieval methods with the materials and to use more visually oriented retrieval tools to provide access to visual collections.

This sense of dissatisfaction with current retrieval methods co-exists with a —perhaps unfounded—sense that content-based image retrieval (CBIR or CBR) systems could enhance the effectiveness of resource delivery in the digital library.

Unfounded, because there appears to be a lack of understanding of what CBIR systems do, how they function, and how well they work within the digital library community. Digital librarians have no readily available assessments of the various strategies for content-based image retrieval; the few comparative studies that exist (such as the recent Joint Information Systems Committee/Joint Technologies Application (JISC/JTAP) report (Venters and Cooper 2000) are neither expressed in librarians' language nor focused on the humanistic researcher or library administrator. This lack of knowledge exists along with a technology transfer gap. The call for applied research (expressed by Beth Sandore and others during meetings leading up to this report) echoes the sense that ongoing work is not being related to the needs and requirements of the digital library community and is not being integrated into their service-delivery environments. Pure research that identifies more-effective search algorithms is not being transferred into tools that could be deployed in digital libraries. Users, and the digital librarians who serve them, do not see the benefits of these technological developments.

The failure of technology transfer into the digital library application realm does not result from a paucity of image retrieval research. A review of the literature shows that content-based image retrieval is a vital area of computer science. The challenge is to operationalize services based on these technologies and to integrate them into digital libraries. Before we reach that point, we must develop methods to compare and contrast various strategies and to assess where progress has been made and where investment is required in order to create robust technical services. Such methods of direct comparison could improve the caliber of image retrieval research because the relative value of differing strategies could be directly known and the results of different experiments compared.

3. IMAGE RETRIEVAL TO DATE

3.1 Overview

Image retrieval is a large and active area of computer and information science, described as "breathtaking" in its pace in a recent survey (Smeulders et al. 2000). Many large groups maintain extensive teams and support multiple avenues of

research that is well supported by national funding agencies and foundations (see References: Funding Support). Governments and industry are investing substantial amounts, and significant portions of their information research budgets, to the issue (though the field is expanding to include moving image retrieval as well). Several conferences each year are devoted exclusively to these issues and there are many conferences with major components for image retrieval (see References: Conferences). Hardly a major university worldwide is without a research group working on the problem (see the References: Research Groups for a sampling).

Image retrieval research has taken two distinct, and discrete, paths. The first is focused on metadata-based retrieval, where images are found on the basis of associated textual descriptions and indexing. The second is based on feature-driven CBR or CBIR, where computational methods are used to identify and abstract the visual elements of an image. In metadata-based retrieval, a searcher's chosen text strings are matched to those used to describe the image (with or without lexical aids such as thesauri or word stemming). In CBIR, a query image (selected or drawn) is compared against the image database, and images similar to it are retrieved. This ability to use the inherent features of an image to retrieve it is attractive to digital librarians, as metadata-based image retrieval brings with it many problems of disciplinary perspective, intercataloger consistency, and incompatible metadata schemas.

A summary review of the literature shows an exceptionally active community of researchers in CBIR. Smeulders et al. recently reviewed the research focus of more than 200 papers judged important to the field (2003); Rui et al. have summarized research in more than 100 papers (1997). Numerous conferences on the subject are listed in the references. Some specialized subsets of this research area, such as latent semantic indexing (Brinkley 2001, Zoran 1997) or progressive feature searching (Castelli et al. 1998), are quite large and have developed their own conferences, publications, and research and evaluation methods. CBIR research has been well funded and is most often concentrated in large, ongoing research teams within computer science.

Metadata-based image retrieval is a less coherent field (Chu 2001) that conducts research focused on image retrieval for particular disciplines (Shatford 1986, 1999; Roberts 2001) or formats (Hunter 1999, 2002) or on theoretical improvements in the way images are indexed (Jaimes and Chang 2000, Greenberg 2001). Metadata-oriented image retrieval research has less funding, is conducted by individuals or small transient teams, reports its results in a wider range of journals, and is concentrated in information science schools.

Metadata-based image retrieval research appears to have little if any impact on indexers or image metadata developers. For example, at the CNI/Online Computer Library Center (OCLC)'s Image Metadata Workshop Third Dublin Core Workshop (Weibel 1997), no reference was made to retrieval research in the two days of deliberation over data elements minimally required for image retrieval. Instead, an element set with a focus on information retrieval was created by practitioners, on the basis of their experience describing images. This irony is not unique to image retrieval (Bates 1999).

Cawkell (1992) identified the fundamental flaw in image retrieval research: little or no crossover between researchers using "visual" vs. metadata-based methods of image retrieval. This is borne out by a citation analysis conducted by Persson (no date) that was based on authors cited in Rasmussen (1997) and reaffirmed in recent reviews of the literature (Chu 2001) and the research agenda (Jorgensen 2001). Chu's citation study reinforces the gap, noting that the journals with the highest citation rates in CBIR were outside the normal disciplinary discourse of the metadata-based image description community. She also points out the significantly greater volume of literature published by CBIR community, by inadvertently dropping out all metadata-based researchers when choosing the most frequently cited for further analysis. (Margaret Graham of the School of Informatics, Northumbria University reported that this gap is one of the motivators for the creation of the Challenge of Image Retrieval/Challenge of Image and Video Retrieval conference series Personal communication, 2003.)

Methods using both visual characteristics and textual descriptions in retrieval have recently emerged (Enser 2002, Goodrum et al. 2000, Perez-Lopez et al. 2000). Lewis et al. (2002) summarize the issues as they relate to cultural heritage objects and posit the development of a visual [multimedia] thesaurus that will assist in identifying concepts in images, and thus bridge the "semantic gap." Barnard et al., in Clustering Art (2001), Barnard and Forsyth (2001), and Li and Wang (2003), among many others, including a group of papers presented at Internet Imaging 2003, have been exploring the relationships between visual characteristics and keywords. Image searching tools on the Web such as MetaSEEK (http://ana.ctr.columbia.edu/metaseek) and SIMLPIcity (http://wang.ist.psu.edu/IMAGE) use image and text in combination. However, evaluation is lacking (Chen and Rasmussen 1999). The methods and measures for comparative evaluation of the two approaches, or of hybrid approaches, are poorly articulated and untested (Sormunen et al., Wang et al. 2003, and Bernard and Shirahatti 2003 represent early attempts).

3.2 Evaluating Image Retrieval

Research is beginning into methods of measuring, evaluating, and benchmarking image retrieval systems. In particular, a collaborative effort to develop a CBIR benchmarking environment is under way involving the University of Geneva, the Viper Group (http://www.viber.nige.ch/benchmarking/) and the Benchathlon group (http://www.benchathlon.net), with benchmarking events taking place at imaging conferences such as Internet Imaging.

Although some researchers have created data sets against which to test their own methods and some have made these data available to others, there is no widely used data set and no generally accepted set of benchmarks against which to evaluate new methods. Sormunen et al. (1999) explored the use of a task-oriented evaluation framework and a test collection to evaluate CIBR. Müller et al. (2001) described a process for evaluating image browsers that attempts to define the

"contribution of low-level, feature based systems to retrieval success" and posits the existence of a set of well-described images as a means of evaluating CBIR systems (though they do not define "well described"). Assessing the relative effectiveness of any image retrieval methods can be costly and frustrating (Venters and Cooper 2000). It is difficult to assess the relative utility of metadata-based systems and to compare them to CBIR.

There are few significant studies of users' needs for or experiences with image retrieval systems. The Consortium for the Computer Interchange of Museum Information (CIMI) (1995) summarized work about access points in the cultural heritage community. Jorgensen (1999) looked at the relationships between naive user query presentation language and some image classification systems. Rodden (1999) explored the utility of incorporating CBIR "intelligence" into interfaces. Markkula and Sormunen (2000) looked at the specific use of a digital newspaper photo archive, building on the work of Eaken, Enser, and others. Pisciotta et al. (2001) reports on an ambitious study now under way at Pennsylvania State University. But as Rasmussen (2002) points out, no balance has been achieved between system-centered and user-centered evaluation of information retrieval. It should therefore come as no surprise that there is neither a consensus on the most promising approaches to image retrieval nor an agreement on how proposed approaches, systems, and tools should be evaluated for effectiveness (Eakins and Graham 1999).

The digital library community, which is a major organized consumer and creator of image databases and has historically had a substantial interest in retrieval effectiveness as part of its service mission, is concerned that the various approaches to image retrieval have not been assessed against common standards, that too little is known about success factors, user needs, and retrieval methods. On the basis of the perceived effectiveness of the NIST/TREC program (http://trec.nist.gov/overview.html, the Council on Library and Information Resources (CLIR) and the Coalition for Networked Information (CNI) sponsored a series of meetings in 2000/2001 to explore whether a shared testbed for image

retrieval research could address these problems. These meetings probed a specific question:

Could an image retrieval benchmarking database focus the image retrieval community and further its results?

4. TOWARD AN IMAGE RETRIEVAL BENCHMARKING DATABASE AND RELATED SERVICES

4.1 Why Benchmarking?

While exploring the efficacy of an Image Retrieval Benchmarking Database the participants in the planning meetings (see Acknowledgements and References: Project Documents) articulated a number of questions that reflected the desire of the digital library community to understand aspects of retrieval and possibly integrate CBIR methods into their services. These included:

- *Can we compare CBIR and metadata-based image retrieval?*
- *How do we compare different methods of CBIR?*
- *How do we compare different methods of metadata-based retrieval?*
- *Howe do we evaluate new methods?*
- *What are the optimum levels of metadata for description of images (e.g., most cost-effective)?*
- *What are the costs of creating effective metadata?*
- *What is the most effective balance between description and retrieval metadata?*
- *What needs to be described in images?*

Each of these questions can be answered only by benchmarking.

Benchmarking involves the comparison of the results of two or more different methods of performing a known task with a known result (the benchmark) in order to establish relative effectiveness. Benchmarking is a critical component of

establishing best practices, key performance indicators, or performance metrics, all of which are other terms often applied to benchmarking or its results.

We can compare two methods of CBIR by using them to answer the same questions by searching the same data set.

We can compare different methods of metadata-based retrieval by seeing how well images described using different methods are found when the same questions are asked of them and they exist in the same data set.

We can assess optimum levels of metadata for the description of images and establish things such as cost effectiveness in image description *if* we can compare the ease with which images are retrieved in controlled circumstances against the cost of creating metadata.

We can test the balance between descriptive and retrieval metadata by assessing how well images with different levels of metadata are retrieved when standard queries are run against a common data set.

We can begin to identify what needs to be described in images if we review the results of query effectiveness tests and compare them to the standard kinds of metadata assigned to images. Provided that the queries reflect real user need, we can see what kinds of metadata are most and least likely to be used (if the queries reflect real user need).

We can compare QBIR and metadata-based image retrieval and, possibly, assess the degree to which these two approaches are complementary, if these two methods are used to ask the same questions of the same data set and the results are compared.

Key to the success of all these studies is the existence of a benchmarking environment where research can be done in controlled circumstances.

4.2 Who Does Benchmarking?

What kind of community of interest is required to sustain an image reference database?

Benchmarking is well established throughout the economy, in areas from automotive manufacturing to knowledge management to higher education. All benchmarking initiatives are committed to sharing information and to improving business processes or performance. Through shared measures, assessments can be conducted that provide comparable results in different contexts. The emphasis in benchmarking is on reliability and comparability.

Benchmarking can be of benefit whenever it is necessary to compare the results of different systems accomplishing the same task.

Benchmarking is valuable to process managers because it permits them to compare the outcomes of their processes with those of a standard measure. For example, it might form a key component of a quality management program under ISO 9000/9001. Benchmarking is valuable to the consumer because it enables the comparison of different processes, by using them to produce a common product. Where the product is a service, it enables cross-supplier comparisons.

Benchmarking succeeds when it has a recognized place in a community of producers or researchers, for it is their work that is benchmarked, and in many cases, they also participate actively in the benchmarking tests.

Benchmarking is usually conducted by those creating products and used by those consuming products. Producers often organize benchmarks to ensure fairness in the measures and their application. A good example of this is the Embedded Microprocessor Benchmark Consortium:

> EEMBC, the Embedded Microprocessor Benchmark Consortium, was formed in 1997 to develop meaningful performance benchmarks for processors and

> compilers in embedded applications. Through the combined efforts of its members—more than 45 of the world's leading semiconductor, intellectual property, compiler, and RTOS companies—EEMBC® benchmarks have become an industry standard for evaluating the capabilities of embedded processors and compilers according to objective, clearly-defined, application-based criteria. . . . EEMBC benchmarks reflect real-world applications and the demands that processors encounter in these environments. [. . .] Mission: EEMBC will work collaboratively to develop a suite of performance benchmarks that will target key applications of embedded systems. These benchmarks will help provide customers an objective means of evaluating processors and controllers (http://www.eembc.org/About.asp).

Benchmarking is rarely carried out by consumers alone, with the possible exception of product testing (Consumers Union product tests come to mind; see http://www.consumerreports.org/).

When the results of research processes are being benchmarked, the active participation of the researchers is critical.

4.3 How Is Benchmarking Done?

Regardless of the process being measured, the steps in benchmarking are similar.

1. Create an accepted test environment that the community of researchers sees as valid for the purpose of establishing a benchmark. It must be reflective of real-world circumstances, yet controlled enough to be scientifically credible.
2. Define a benchmark for that environment that establishes the attributes of complete success for each task to be studied. Benchmarks must reflect real-world goals, with attributes that can be readily identified when a task is completed.

3. Conduct a number of studies. These may be done by one or more groups, in one or more research projects, using different methods of achieving the task.
4. Compare the results of each to the measures established in the benchmark and evaluate their effectiveness.

Ideally, results are studied further to establish what factors contribute to success or failure (and to what degree) and the significance of each factor to the user community in terms of cost, impact, or other factors. Such analyses can aid in making decisions about deploying particular of image retrieval technologies.

4.4 When Is Benchmarking Valuable?

Other questions discussed at the planning meetings centered on when it was appropriate to introduce benchmarking. Participants asked:

- *Is it too early for benchmarking?*
- *What benchmarking resources exist?*

Benchmarking may be introduced as soon as there are multiple methods available and it is important to know which ones work best. For example, Edward Jones describes the establishment of a benchmark for fingerprint-identification systems (Jones 1997). He argues that the extent to which civilian agencies of government, including local government, already rely, or will soon rely on biometric identification systems and the numbers of new firms offering technologies in this area mean that sound public procurement practices require standard and reliable benchmarking of systems in their real-use environments.

Comparability of results is important both for assessing the overall value of a method and for exploring how to improve the way a method works.

A sufficient requirement for benchmarking would exist if the community funding image research, conducting image research, or consuming image research demands it. In an area where there are already a large numbers of studies being funded that

are producing results that are not comparable, there is a significant argument for benchmarking, i.e., comparable results across research groups offer added value. The advancement of the field is dependent upon the evaluation of the efficacy of a method and the assessment of results. Benchmarks can provide a clear point of comparison.

Image retrieval research is ripe for the introduction of benchmarking:

- A community of image users cannot make assessments. Users are being offered image retrieval systems that claim effective performance in finding results in the new, large, digital image resources now available. The developers and deliverers of digital libraries are unable to evaluate the various products offered to them.
- Many groups are producing research results that are not comparable. Reported enhancements, improvements, and advances have not been evaluated.
- A significant amount of funding is being devoted to image retrieval research. It is difficult for funders to determine value of this work. It is equally hard for potential implementers to determine what methods could improve practical retrieval software systems.
- The image retrieval community itself is calling for a standard test set as a means to compare research and assess progress (Rui et al. 1997, Jorgensen and Srihari 1999, Eakins and Graham 2000, Goodrum 2000, Smeulders et al. 2000, Venters and Cooper 2000, Muller et al. 2001c, Smith 2001, Jorgensen 2002). Indeed, researchers have published the results of experiments that show that different data sets influence the evaluation of the same system. (Muller et al. 2002).
- There is a need to improve access to visual information and extend digital library service in this area (Pisciotta 2001, Graham 2001, Burford et al. 2003). It is hoped that improved access will increase the use of the digital visual collections themselves.

If image retrieval really worked, there would likely to be a group of ready users. Benchmarking can help us understand whether the technology works.

4.5 What Could Be Benchmarked, and How?

What is the difference between being a "research resource" and a benchmarking database?

What are the costs of creating and sustaining an image reference database?

Developing an image benchmark database is a significant task, made more complicated by the many aspects of image retrieval that might be tested in such an environment.

Processes that could be benchmarked relevant to image data include the following:

- capture
- compression
- color management
- lighting
- database and storage
- image displays and user interfaces
- printing and replicating
- retrieval
- indexing methodologies, tools, and vocabularies
- transmission

Each of the aspects identified as important to image quality (Williams 2000, CLIR/RLG/DLF 2002) could be isolated and tested against a known data set.

In many cases, there are fundamentally different technical options for the same process, e.g., initial capture on microfilm or by digital scanning; retrieval by associated text (metadata) or by image content. We could evaluate these options by constructing measures applicable to either method or to a hybrid of the two. For

example, we could measure time to prepare each image source for capture associated with a "successful" scan (defined as the measure we create) in scanning vs. microfilming and also when both processes were being conducted simultaneously to generate two copies. Similarly, we could measure effectiveness in retrieving a pre-identified subset with specified characteristics using metadata-based or content-based retrieval or a combination of the two.

The same visual data set could be given to different groups to test the effectiveness of different metadata formats and indexing schemes for supporting retrieval. For example, Tam and Leung (2001) and Jörgensen (1999 and 2001) have each proposed a new method of organizing and generating visual descriptors. Tam and Leung use a structured natural language description to support enhanced retrieval of visual materials; Jörgensen proposes a conceptual method for classifying visual descriptors. Retrieval using such a new method of description could be compared against retrieval using more traditional methods (e.g., MARC/AACR2).

Similarly, metadata-based retrieval schemes could be compared with schemes that use primarily the computational analysis of visual characteristics, or retrieval strategies that use a combination of the two schemes.

4.6 How Are Benchmarks Established?

How do you create a benchmark for retrieval systems that operate in different ways?

A benchmarking database is more than just a collection of data. To enable assessment, such a system must include the benchmarks, or statements of what would constitute success for different tasks as applied to specific subsets of the data. The defining characteristic of a benchmarking database is these baseline measures, or ground truth, that, if constructed with care, reflect what is important to a user community. The benchmarks should be created in consultation with experts within the prescribed user domain that represent the intended users of the result, or across domains for interdisciplinary uses.

Thus, a set of measures of "local government response to citizens requests for permits" would establish the kinds of permits requested by citizens and the typical questions posed by those seeking permits. It would also establish baselines for the number of referrals the first contact would make, the amount of time to answer the question, the accuracy of the answers, and whether the citizen received the necessary form for filing the permit as a result of the contact or had to go somewhere to get them, among other things. A set of measures for an image-retrieval benchmarking database would establish the kinds of images requested by the user community, the uses to which they are put, and the methods used to identify them. Baselines could be identified for the number of images used, the amount of time it takes to identify relevant images, the number of irrelevant images retrieved, and the number of times nothing was retrieved (though it might exist in the benchmarking database).

Image *capture* benchmarks (whether scanning or microfilming) must be established on the basis of the intended use of the digital image. If the process is intended to create an archival copy, archivists and conservation professionals would create the benchmarks. If a digital image was designed to be used by researchers or the general public, members of these audiences would be brought to the table. What is key is that users are involved in constructing the benchmark in order to ensure that characteristics important to them are measured.

Similarly, an image *retrieval* benchmark could be constructed by having a group of experts with content domain knowledge (e.g., art historians, anthropologists, astronomers) characterize a subset of the database in a way that established which items should be retrieved in response to particular domain- relevant queries.

An image *quality* benchmark could be established by having experts representing one or more communities of image users characterize a subset of the database according to its appropriateness to a particular use.

All these benchmark databases could reside in the same repository and be served by common management and technical facilities, but the ground truth for each benchmark would have to be established by an appropriate community of users, the "relevance testers."

The purpose of the benchmarked data set is to establish a known and validated result. A variety of retrieval or image-capture methods could be compared even though each was different, because what is compared in the end is the result of one method to achieve the benchmark.

Ideally, all stakeholders should agree that benchmarking is needed, because a significant investment is required to compile the data and develop the queries and relevance assessments. Researchers must undertake an additional effort to use new benchmarks in future research. Consumers will also have to invest in the process, assisting in the construction of validated benchmarks on the basis of their domain expertise and helping formulate meaningful domain problems that can be bench tested.

4.7 Questions a Benchmarking Database Cannot Answer

Some of the questions posed at the planning meetings could be answered by research that is conducted using image sets drawn from a benchmarking database but are not answerable by the database itself. For example, participants asked:

- *What are the best strategies for quality digitization?*
- *How do we assess image quality?*
- *What subsets of metadata are of interest to which communities of interest?*
- *Does facial recognition (or another specific use) require the same or different qualities of images than other uses (for example, art historical research)?*

In each of these cases, it is necessary to have several studies of different methods, as well as comparison of these studies to a common benchmark, in order to answer the implicit or explicit questions about relative value or effectiveness. It is not the task of a study of the feasibility of constructing a benchmarking database to answer

these questions, but it is encouraging that there are many such questions of importance to the community of image users, since they are one of the three identified stakeholder communities.

4.8 Other Issues

Other questions raised at the planning meeting showed a need to understand the domain of image databases more fully:

- *What research fields use images?*
- *What studies are there of users of image databases?*
- *What types of queries do researchers put to an image database?*

While such questions wouldn't be answered by an image retrieval benchmark database, they could be addressed in the planning for an organization that might host such a resource. The current-awareness function could become one of the elements of the organization's mission.

Questions were raised that relate to the practicality of finding existing data sets, with domain- appropriate metadata and existing user communities, relevant queries, and potential for benchmarking. The group also asked questions about strategies for developing a research resource, including the following:

- *What existing data sets are there?*
- *How could they be made available?*
- *Are there rights problems in using them?*

The answer is that there are many existing databases available (See References). An assessment of which databases, or parts of databases, to draw upon needs to be made within a larger context of organizational, research and community development objectives.

4.9 An Environment for Research

An image retrieval benchmark database would not provide answers to many of the questions posed during the planning meetings. It would, however, provide a predictable place in which to ask them and to evaluate the results, thereby

becoming a precondition for quality research in this area. It is possible that over time the same test image database could be used in different tests, for different purposes.

It is important to focus attention initially on image retrieval in order to limit the scope of the problem to a manageable size. Other aspects important in the creation of image database services, such as perceived image quality and cost-effectiveness of capture or description methods, could be topics of future research.

An examination of the issues identified makes it clear that rather than a static, network-accessible database, the community is searching for a service that enables and supports the evaluation and comparison of various image retrieval techniques.

5. Planning an Image Retrieval Benchmark Service

5.1 Goals for a Research Benchmarking Service

By focusing the discussion on developing a benchmarking system to support image retrieval research initially, the scope of the problem can be narrowed and the results can be more achievable. A successful image-retrieval benchmarking database would

- improve understanding of image retrieval research in the digital library community, through the careful articulation of research questions and benchmark tests
- enable researchers in the field of image retrieval to measure the effectiveness of their retrieval methods against a common standard
- provide researchers, digital librarians, and potential users of image retrieval systems with a means to assess the effectiveness of different methods and with a basis for making decisions about their deployment
- enhance the value of research on retrieval methods through comparative measures
- improve the effectiveness of image retrieval through comparison, evaluation, competition, and cooperation

- raise the overall quality of work in the field through the creation of a focused center of excellence
- encourage cross-method collaboration through the development of an interdisciplinary community of research

Known assessments of effectiveness would, it is hoped, stimulate the application research results. Results could influence practice in the creation of metadata for image collections, resulting in more effective investment. The design of production-level image retrieval systems would also benefit, because tools known to be effective could be incorporated.

5.2 Audiences/Users of the Benchmarking Service

It is easy to confuse the audience or user community of an image retrieval benchmarking service with the audience interested in the results of studies of retrieval effectiveness. In reality, these are two distinct communities. The users who will both interact with the benchmarking environment and contribute to it (both data and tools) will be image retrieval researchers. Unless the community of image retrieval researchers—both those working in metadata-based retrieval and CBIR—finds benchmarking useful and accepts it as a significant part of their research process, the service will not succeed. If the service becomes a source of and a repository for images used in research studies and is referenced a standard benchmarking facility, over time there will be evaluations of image retrieval effectiveness that the image retrieval system consumer community can use. The digital library community will have to wait some time for usable information retrieval research results.

It is highly unlikely that the same data set will serve subject specialists who are end-users of image databases. It will not have been selected to support end-user interests; indeed, it will have been selected for its heterogeneity of users and content. Furthermore, it is unlikely that the permissions granted for use will extend beyond image retrieval research studies.

It is possible, however, that a database created to support image retrieval benchmarking could provide data sets to researchers in other aspects of imaging systems, such as compression, watermarking, printing, and color calibration. This is a secondary benefit. The selection criteria for inclusion in the image retrieval benchmark database are, at least initially, dictated by the image retrieval benchmarking requirements, and they must remain so. Using the data sets for other purposes will generally require the creation of distinct benchmarks, or ground truth, for these other evaluations and may require the gathering of different metadata. A retrieval focus is essential for initial success.

5.3 Components of an Image Retrieval Benchmark System

In his article "Quantitative Assessment of Image Retrieval Effectiveness," John R. Smith of the IBM T.J. Watson Research Center outlines the requirements for an image retrieval testbed, reminding us that the issue is not simply one of collecting a number of images and making them available for reuse (Smith 2001). Other evaluation literature, such as that which describes TREC, the "model" text retrieval initiative, concurs on the following components, which owe much to the Cranfield methods of information retrieval research (Rasmussen 2002).

5.3.1 Collections of Test Images

5.3.1.1 Which Test Images?

The set of test images is what is queried. There is general agreement that such a test should be large (though no agreement on what "large" means. Jorgensen (2002) cites more than 100,000 images; Smeulders et al. (2000, 1373) say "a state of the art paper in content-based retrieval reports experiments on thousands of images." Existing studies use anywhere from a single image (JPEG's Lela) to "over 200,000" (Huang and Zabih 1999). There is no real measure for this number. The selection of current data sets often seems to be the result of convenience, rather than of a systematic attempt to represent a variety of visual genres in a balanced manner.

There is also no agreement, and no easily discernible answer, to the question of the ”content” of such a digital collection. The close relationship between the collection and the kinds of image retrieval research questions it will support needs further exploration with image retrieval researchers and digital librarians. Different end-user disciplines will have different opinions about relevance.

Some general characteristics are clear. An image test database must be free of intellectual property (IP) restrictions that would limit the distribution of research results or the replication of experiments. While ”freely accessible over the Web” might be an ideal, it is unclear whether rich, diverse collection of images could be assembled with this as the IP framework. Image owners may be more receptive to scenarios that limit the redistribution of their images in ways that still respect the research goals of the image retrieval community. Required is the development of a set of agreements (i.e., terms and conditions of deposit and use) that balance the needs of image owners with those of retrieval researchers.

A test set should satisfy the kinds of queries likely to be asked of within a particular domain or domains. Random collections of images “crawled” from the Web will not satisfy this requirement, nor will collections of stock photographs, because the questions asked must be nontrivial from the perspective of the end-user discipline. It is possible that a domain could be as broadly defined as the humanities, in order to test assumptions about cross-disciplinary use. Whatever the domain, a balance must be established between the test image set and the research questions that will be supported.

Even though these images will need to be described in a number of ways in order to provide for ground truth in different domains, images should be diverse and represent a full range of visual information types, from all kinds of works of art, to contemporary news and historical photographs, to natural and physical science images, to schematic and graphical representations. In addition, the test set will need to contain a set of known textures and colors as targets for very specific CBIR queries.

These two requirements, homogeneity and diversity, may seem be in conflict, but in fact they simply dictate the creation of a fairly large data set assembled from identifiable, smaller sets with some subject- or content-based coherence.

5.3.1.2 Existing Image Databases

The creation of an image research database from new materials would involve a level of time and investment that would probably not be sustainable. In addition, real-world data are valuable in and of themselves as a basis for research. Therefore, the test data set should be composed of existing images, drawn from extant image databases, and created for real communities of users.

Images and metadata from many existing databases could be made available to the image retrieval research community for use as part of a benchmarking data set. Today, image databases support research in virtually every academic discipline and in many commercial activities. Finding image databases that will permit research use, that contain domain-relevant metadata that have been independently validated by their use in a primary use community, and that are technically specified in a comparable way, should not be impossible. The challenges will be to determine what size data sets to draw from each such source and to gather appropriate tertiary data about the data sets (e.g., the costs of indexing and metadata creation) and profiles of existing users that enable the support of image retrieval research.

Data sets used in specific image retrieval research projects (often over many years) may also be available from the researchers participating in a benchmarking database initiative. These will need to be evaluated, particularly with regard to user needs and, if from CBIR studies, with respect to metadata requirements, before their incorporation into an image-retrieval benchmarking database. Care must be taken to set the requirements threshold to avoid excluding too much a priori (too high a threshold for metadata excludes most resources).

The preliminary list of existing image databases found in the References provides a sense of the range of image databases available. It is in no way intended to be comprehensive or necessarily to report the largest or most important databases in any discipline. Instead, it is designed to give a concrete flavor to the variety of uses such databases serve, and hence the range of what would be perceived as relevant visual content by their primary use communities.

The selection of image sources must proceed hand-in-hand with a refinement of the domains that will supply the research questions that will be benchmarked in retrieval tests.

5.3.1.3 The Problem of Image Description

Many different metadata standards are used to describe visual materials. These range from the highly structured MARC format and its accompanying cataloguing rules (AACR2), through less prescriptive, though still structured domain-specific standards such as the Categories for the Description of Works of Art or the VRA Core, to free-form keywords assigned by indexers or researchers testing CBIR systems, to the captions, nearby text, and narration that are providing context to CBIR researchers using the Web as their laboratory.

Any group of images incorporated into the test data set should include some metadata about the content of the visual images. These metadata should be structured according to one of the approaches in widespread community use (whether "standard" or not). Since it is undesirable that all new metadata be created for this project, those constructing the database must be prepared to accept metadata in its native format and map this to a format adopted for the test data set. While this requires some attention at the first instance, the experience of both RLG's Cultural Materials Initiative and AMICO in creating The AMICO Library™ have proved this feasible.

Developing a test set of data from any preexisting image resource will require additional analysis and the assignment of additional metadata in order to provide

the ground truth necessary to determine relevance rankings for many queries. For example, color is one of the key retrieval characteristics in CBIR, but it is seldom described explicitly in metadata from other sources. If color is to be one of the test query elements, then this attribute needs to be assigned to a significant number of images in the test data set.

The metadata for the images in the test data set will have to satisfy the three levels of description described by Smith (2001 drawing on Smith and Benitez 2000). These levels are

- semantics: what is shown in the image (literally and figuratively)
- structure: the relationship of the various visual elements (i.e. image composition)
- features: individual visual characteristics that make up the images (e.g., color, texture, and shape)

James and Chaing (2000) offer a possible way of approaching the integration problem through a series of conceptual levels expressed as a pyramid. Greenberg 2001 approaches mapping of metadata systems through a functional lens. Burford et al. (2003) propose a multilevel taxonomy of image content. All strategies may prove helpful in defining a superset of image metadata elements that accommodates all existing descriptive schema. XML/RDF seems a promising approach to mapping ontologies.

A distinction will need to be maintained between whatever descriptions "come with" images incorporated into a test data set from a preexisting source and descriptions that are supplemented in order to describe an image for new domains or retrieval methods. It is hoped that manual indexing may be enhanced in part with automated indexing tools (see Barnard et al. 2001, Barnard et al. 2002, and others). Additional or alternate image metadata might also be supplied by the research groups that use the test data set: If a research goal is to test the effectiveness of a particular descriptive or indexing schema, then a group might choose to describe a subset of images in a novel manner and prove this enhances

retrieval in subsequent tests. Following the experiment, these new metadata could be made available for reuse within the data set, though they would also need to be sourced, along with all other metadata assigned to the images.

5.3.1.4 Requirements for the Creation of a Set of Test Images

The six requirements for creating a set of test images are as follows:

- identification of research domains to be covered and user communities for which ground truth will be established
- shortlist of data sources
- terms and conditions for acquisition and use
- statement of metadata requirements (minimal metadata needed with image set)
- an extensible "meta-metadata" schema to incorporate all possible data sent with images
- systems to compile and store images and metadata

5.3.2 Benchmark Queries

Benchmark queries are the standard questions asked of each system to be evaluated and are answered using its particular method(s). Defining test queries for an image retrieval benchmarking requires particular attention to the different manners in which questions are posed of image data sets. Some are purely textual, and metadata based, some are visual, defined using a target image, or through the provision of a sketch.

If performance is to be evaluated successfully, queries must also reflect the domain for which an image set was developed. While some generic questions about image features might be asked of any image, different semantic questions will be asked of art databases than of astronomy databases.

Clear articulation of test queries is critical to the success of the tests themselves. This is a significant problem when the query is visually expressed, and the retrieval method is one of determining visual similarity.

5.3.2.1 What Should be Queried?

Queries made of an image benchmark database should represent the full range of user queries of such materials. These have been characterized by Eakins and Graham (1999) as

- primitive features (color, texture, shape)
- derived or logical features (identity of objects depicted)
- abstract attributes (named events, meaning of images)

This should be compared with the levels of subject identified by Panofsky (1962), reflected in the Categories for Description of Works of Art (CDWA*)* and commonly used in the discussion of subject indexing of visual materials in the arts (Lane, Shatford Lane, Roberts 2001):

- pre-iconographic: the basic, named elements of an image (man and woman with apple)
- iconographic: the identification of an event, or personage (Adam and Eve)
- iconological: the meaning of an image or its symbolic significance (the Fall from Grace)

These levels are often described as what an image is *of* (pre-iconographic) and what it is *about* (iconographic and iconological). Eakins and Graham have defined two levels of *of-ness* and one level of *about-ness.* Panofsky, alternatively, offers one level of "of" and two of "about."

Examples of queries in these areas might be expressed as

- find all images of men
- find all images of men with beards
- find all images of Abraham Lincoln
- find all images of male authority in the nineteenth century

When images are used as more than illustrative material, the questions themselves become more complex.

A researcher studying the uptake of new technologies might ask, "What was the proportion of horse-drawn to self-propelled vehicles on the major streets of American cities in 1900, 1905, 1910, 1915, and 1920?" Or, a sociologist might ask, "What proportion of advertisements in women's magazines relates to cosmetics in the second half of the nineteenth century?"

Some queries will be expressed not in words but in images, either drawn or chosen (the retrieval challenge being to find an image "like" this one). These kinds of queries include the following as "CBR challenges" identified by the Benchathlon network:

- find this image
- find an image that this image is a part of
- find an image that looks like this sketch
- find the compressed version of this image
- find other versions of this image
- find other images of this subject (person, place, thing, etc.)

The Benchathlon group also define a series of system performance measures, for example, the length of time it takes to add an image to the database, that are unrelated to image retrieval but significant in image database management.

While much literature on image indexing focuses on it, Subject Matter is only one of over 30 information areas defined by the Art Information Task Force in the Categories for the Description of Works of Art that could be used to cluster images and retrieve sets from image databases. Other areas emphasize the physical nature of the image, the context of its creation, its history or use over time, and the reactions to it. These are defined as information elements in the AMICO data specification. Used in combination, they cluster images according to the needs of a particular user community.

Queries should be directed toward different kinds of image characteristics. However, because one of the significant research questions is determining how combining text and image features can enhance retrieval, all retrieval methods should be evaluated for use in answering all queries.

Combining retrieval methods might support complex questions such as

- What proportion of Picasso's paintings in his Blue Period was predominantly blue?
- What role did black play in the paintings of the Impressionists?
- When did paisley appear as a pattern in British textiles? How quickly did its use spread?

Such questions as these illustrate the relationship between the queries and the test data set. If there are no images of Impressionist paintings in the data set, putting the question is futile.

5.3.2.2 How Many Queries Should be Used in Each Test?

The number of queries in a test run contributes to the validity of the experiment. When defining the number of queries used in each test, TREC uses 25 as a minimum, with 50 as a norm; they accept that 25 queries are too few (Voorhees 2002). These seem to be practical judgments and are not based in theory.

5.3.2.3 Requirements to Create Benchmark Queries

- communities of experts (image-using researchers) to establish typical queries that could be answered by the visual information represented in the test set
- knowledge of user-studies literature (to be sure that queries are grounded in broad-based experience)
- knowledge of test image set (to be sure answers to queries are available)
- expression of queries in plain language, along with a characterization of the kinds of documents that would answer them.
- a system to compile queries and the descriptions of target resources

5.3.3 Relevance Assessments

Each image in the test set should be ranked vis-à-vis each benchmark query, so that the results of each retrieval system tested can be assessed. Systems that retrieve more of the relevant images, and fewer of the irrelevant images, are said to perform better. The assessment of relevance should be based on the domain-specific statement of the kind of document that would answer the query.

When results are reported, it must be possible to determine, via these relevance scores, how well a particular system performed by identifying the following:

- how many relevant images were retrieved
- how many irrelevant images were retrieved
- how many relevant images were not retrieved

If the systems in question assign relevance rankings in the process of retrieval, a comparison of those rankings, such as "least relevant," could also be a useful measure.

Establishing relevance theoretically requires that all images be ranked vis-à-vis a particular query, so that their relevance can be assessed if they are retrieved. Ranking is a substantial investment and may need to address a significant number of works. Metadata to support the generation and recording of relevance assessments are complex and must address many aspects of the visual characteristics of an image. The investment to create this is difficult to quantify. Reporting results of a pilot study, Bauer (1997) estimated that it would take three to five hours to make relevance judgments for 500 images.

5.3.3.1 Reflecting Relevance Judgments in Metadata

Ground truth—the data about the database that will be used to determine whether images meeting specified descriptions were in fact retrieved by queries designed to retrieve them—is based in specialist descriptions of the images, including their expert relevance judgments (Is it a cathedral? Is it a Dutch master drawing?). It may be possible to derive assessments of relevance for one type of query from the

results of another; for example, subject terms could identify the images that should also be found using visual characteristics. These "pre-queries" could enhance effectiveness in assigning relevance judgments. Barnard and Shirahatti (2003) and others have modeled systems for recording relevance judgments in a simple, straightforward manner.

Because there is a nearly endless number of ways to characterize images in the diverse communities that use them, trying to build a new database and establish ground truth for each image on the basis of observations by many specialists in numerous disciplines is likely to be impractical. Aggregating data from known sources, each with internally coherent metadata that can be characterized themselves, is likely the most effective initial approach to constructing the benchmark data set. Subsets of the larger image database could be used in particular tests, even if the entire data set has not been consistently characterized. Because distinct chunks of the larger whole will have an internal coherence, comparison of effectiveness of certain search strategies across them could be a tool for establishing effectiveness in generic environments.

5.3.3.2 Identifying Visually Based Relevance

When the query is expressed visually (i.e., as an image itself) and the challenge is to find images 'like' it, determining relative likeness in a large data set may be impossible. Howe (2000) explores this issue and proposes a methodology that uses a set of query images that have been altered in known ways as query input in order to provide comparable known results for query evaluation. Sormunen et al. (1999) and Barnard and Shirahatti (2003) have also proposed methods for assessing relevance.

To reduce the burden of relevance assessment, TREC has developed a practice of *pooling*, whereby the universe of retrieved documents is ranked and documents not retrieved by any participant in the trial are deemed to be irrelevant. While a significant method for reducing the effort required to analyze the results of a test

query run, this method is criticized by Smith (2001) for falsely judging all items not retrieved as irrelevant.

A similar strategy, of analyzing the retrieved sets of all test participants rather than analyzing the data set as a whole, might be the only viable approach for a large image-retrieval benchmarking database.

5.3.3.3 Requirements to Establish Relevance Assessments

- metadata analysis to identify which slices within the larger data sets could support particular queries
- methodology for making assessments
- human and computational evaluation of images in relation to test queries
- system to support assignment of relevance judgments and to record them vis-à-vis particular queries and works

5.3.4 Quantitative Evaluation Metrics

Smith (2001) adds a set of metrics for evaluation to the components of a benchmarking database more commonly cited. The assessment of relative relevance is as important in the analysis of results as in the initial relevance judgment. Scheduling time for evaluation is critical if the results are to influence research developments.

In traditional retrieval research, results of any systems queries are evaluated against the relevance assessments, in terms of precision and of recall. Further analysis might break out performance in these areas by any one of a number of factors related to areas such as system design, metadata design, image content, or cost.

5.3.4.1 Requirements for Quantitative Evaluation Metrics

- information retrieval expertise to define methodology
- statistical expertise to evaluate submitted results of test queries as executed by different systems.

- digital librarian input to identify areas of interest to evaluate

5.3.5 Community of Researchers

To succeed, an image reference database benchmarking service must develop and maintain a community of researchers who are vested in its development and success. The benchmark database must engage researchers at all stages in their research process: defining research questions, identifying test data sets, creating and testing image retrieval systems, determining evaluation metrics and methods, and reporting results.

The community of image retrieval researchers must participate in the trials, help set the queries to be benchmarked, assist in evaluating the results, and contribute to the development of the test image set.

For a few years, as the community commits to and builds a benchmark service, much of its research will be directed towards establishing what the service will be and how it will operate.

5.3.5.1 Requirements for Developing a Community of Researchers

- consensus on the form and nature of benchmark system that has been developed through workshops and meetings
- conference at which research results are reported and discussed
- dissemination of results through the Web and publications
- mechanisms (meetings, conferences, publications, Web site) to develop awareness and facilitate community contributions as test set is planned, developed, deployed, and used

5.4 Success Factors in the Creation of an Image Retrieval Benchmarking Service

In moving toward the development of a community with shared interests and goals in furthering the state of image retrieval research, many factors come into play, as discussed below.

5.4.1 Sponsorship

The legitimacy of the database and the backing of the international community are crucial to its success. Sponsorship from a major organization or foundation dedicated to research in this area (such as The Andrew W. Mellon Foundation) or from governmental sources with long-term interests (such as the European Union Information Society Directorate General, or the National Science Foundation [NSF]in the United States) is essential, not just to pay the costs of creating and sustaining the resource but also to signal its importance. If funding agencies are not convinced that a major investment in comparable evaluation is justified at this stage in the history of image retrieval research, this expensive overhead to the research process will not be sustainable and the undertaking should be abandoned.

5.4.2 Community Buy-In

Substantial proportions of funding during the planning stages must be expended to involve, and maintain involvement of, the major research players. Over time, it will be the commitment of these groups to the benchmarking resource that will make it viable and make it necessary for other researchers to reference it.

Ultimately, if the benchmarking service is successful, researchers will need to use the resource because editors of peer-reviewed journals and members of the research community will expect it.

5.4.3 Governance

Involvement of the image retrieval research community in the governance of the benchmarking service, as members of technical advisory committees governing boards as contributors of test sets and as active and regular users of the benchmarks, is critical to the success of the effort. Involvement of the digital library community and, if possible, of visually based research communities will help bridge the gaps between areas of interest.

5.4.4 Creating Incentives to Use

Injecting the image database into the research process is the most challenging aspect of its creation and launch. The initial impetus to use the benchmarking

service could come from funding agencies that require it as a condition of funding. The impetus could also come from journal editors who require it as a condition of publication – a form of peer review. However, over time, the impetus will need to come from within the community of researchers.

As a case in point, the National Science Foundation funded, in its Small Grants for Exploratory Research (SGER) "An Online Repository of Large Data Sets for Data Mining Research and Experimentation" in 1998 for about $100,000, to create "benchmark testbed to enable researchers in data mining (including computer scientists, statisticians, engineers, and mathematicians) to scale existing and future data analysis algorithms to very large data sets" (see http://www.fastlane.nsf.gov/servlet/showaward?award=9813584). However, it appears this project did not take hold in the community it was to serve. Up-front consideration of usage incentives could help avoid a repeat of this scenario for an image retrieval benchmarking service.

For example, if four of the top six research groups committed to using the test set, a significant center of gravity would have developed to change the community dynamic. We can propose that a measure of success be that within three years, 15 percent of the newly published research in the field would be benchmarked against this database. We could hypothesize that at this stage of use, discourse would turn to the comparative causes of differences in effectiveness. At that time, whatever external requirements for benchmarking had been put in place (e.g., special funding, publication requirements) could be relaxed since the community itself would regulate further use of the benchmark data set. Mechanisms that supported this, including supporting the sharing of research methods and results, would encourage studies aimed at replicating results and enhance comparability and reliability of benchmarking

One factor that should be explored in stimulating initial use is how major research groups could receive credit for contributing to the databases and the tools they might develop (and write about in the literature) that could be used by future

researchers. Structuring "membership" in an image-retrieval benchmarking consortium in a way that gives founding members a sense of purpose could be important. By contrast, the Standard Performance Evaluation Corporation (SPEC) *pays* contributors of applications if their software / task clusters are used as benchmarks in a distributed set of hardware benchmarking tools (http://www.specbench.org/osg/cpu/CPU2004/search_program.html).

Ideally, the community would adopt and use an image retrieval benchmarking service without these kinds of requirements. However, if a funding agency makes a commitment to start-up, such tie-ins will accelerate the benefits of the program.

5.4.5 Technical Success Factors

5.4.5.1 Data Integration/Ingestion

Developing methods for integrating relatively large data sets (on the order of 10,000 images at a time) from various disciplines and projects at a time, establishing their characteristics, and integrating their metadata will provide the benchmarking service with a mechanism to continue to grow as research methods and applications change. It will also provide it with one of its significant documentation challenges. Methods will need to be developed for image transfer, verification (linking), and quality assurance throughout the process.

5.4.5.2 Data Documentation

Since one aspect of measuring effectiveness of retrieval methods is to compare approaches to creation and indexing, it will be necessary to have data from source file contributors about the manner in which the images and metadata they contribute were created.

The need to merge descriptive schemas to gather metadata documenting the content of images from varying sources has been discussed. As well as metadata about the images in the test data set and their content, the benchmarking service will require technical metadata about the digital image files themselves (ideally using elements drawn from the National Information Standards Organization

[NISO] Technical Metadata Standard for Still Digital images [NISO Z39.87-2002]). Technical metadata will support comparisons of capture and storage methods. It is unlikely, however, that full technical metadata will be available for all data sets. It may be possible to retrospectively create such data for some; for others, they may by unattainable.

Administrative metadata about organizational provenance and the terms under which data are contributed to the test data set will need to be recorded to document the creation of the test image data set. The Dublin Core's administrative core or the NISO/Digital Library Federation (DLF) electronic resource metadata elements might prove useful (seehttp://www.library.cornell.edu/cts/elicensestudy/home.html).

Applied image retrieval research will also require information about the costs of the creation of image data and metadata, so as to support the comparative study of technical methods and indexing methodologies. It is only with such pre-established baseline data that we can consider studies that compare the costs of CBIR with the costs of metadata-based image retrieval, for example. Before data sets are acquired, cost measures for their creation and for the creation of cost-related metadata will need to be agreed on if possible.

5.4.5.3 Research and Development

A number of tactical decisions or development issues must be confronted in designing the benchmark system. These include how a single database can contain subsets of images that are representative of the images needed in specific domains and how to enable the sourcing and management of metadata, particularly if researchers add to the metadata as a result of tests. The primary the mission of the service, however, will be the compilation and distribution of image data. This area is not likely to raise any fundamental research questions.

The image retrieval benchmarking service could also consider the development of tools associated with the database. These tools could support internal functions

such as indexing and the assignment of relevance assessments or the identification and distribution of subsets of data. Initially, however, all manipulation of the data required to perform specific research would be the responsibility of the researcher.

Over time, the service might also provide tools to the researcher, for example, to permit a researcher to degrade the data in ways that would enable comparisons between methods, e.g., to create test sets with different resolutions (for CBIR) or levels of metadata (for metadata-based retrieval).

5.5 Ancillary Costs to the Research Community

The creation of this sort of community resource requires a funded project and the engagement of leaders in the community. In the short term, attention will be directed to the development of tools and methods for benchmarking, rather than to the primary activity of designing image information retrieval systems. This must be recognized up front. This "distraction quotient" may limit the participation of some.

A resource of this kind cannot be created once and left for future use. It requires ongoing care and feeding —growing with the addition of new data sets and requiring that its characterization (ground truth) be extended as new types of questions and new approaches to retrieval are explored. The long-term community cost of maintaining an image retrieval research database is significant.

6. SCENARIOS FOR DEVELOPING THE IMAGE RETRIEVAL BENCHMARK DATABASE

The costs and levels of effort involved in developing an image retrieval benchmark database will depend on the kinds of studies the different benchmarks it is designed to support.

A number of scenarios can be envisioned for developing an image retrieval benchmark database service. Each of the models discussed here needs to be

evaluated in terms of the goals and objectives of the community and its particular costs, benefits, and risks of each strategy.

There are several approaches creating this service. They include the following:

- introduce a new track within TREC
- expand the Benchathlon activity
- sponsor a new image retrieval benchmarking service, inviting participation by interested parties
- create a consortium of image retrieval groups in industry and academia, and make one of its program elements the maintenance of an image retrieval benchmarking service

6.1 TREC Model

The most commonly cited model for a widely available, well-documented test set is the TREC, hosted by the NIST (http://trec.nist.gov/overview.html). Since 1991, TREC has coordinated annual benchmarking tests "to encourage research in information retrieval from large text collections" and sponsored an annual conference and publication to disseminate results. TREC is widely cited as *the* example of a research data set around which a community can focus its work.

In a paper directed to the music information retrieval community, Ellen Voorhees summarized the goals of TREC:

- to encourage research in text retrieval based on large text collections;
- to increase communication among industry, academia, and government by creating an open forum for the exchange of research ideas;
- to speed the transfer of technology from research labs into commercial products by demonstrating the impact that improvements in retrieval methodologies can have on real-world problems; and
- to increase the availability of appropriate evaluation techniques for use by industry and academia, including development of new evaluation techniques more applicable to current systems (Voorhees 2002).

The image retrieval community shares these goals; however, the communications challenges appear greater in image retrieval because of the distinct natures of the text- and image-based retrieval communities.

6.1.1 New TREC Video Tracks: The TREC Digital Video Test Collection and The Open Video Project at the University of North Carolina

In 2001, TREC added a video-information component to its sets of reference data. Developed in consultation with video information retrieval researchers (Schmidt and Over 1999), the set has now been used in TREC tests. Slaughter et al. (2000) describe the framework for another test collection of video designed to provide for the video research community many of the same benefits as have been projected for an image retrieval benchmark database, i.e., to have an available set of reusable video; to organize a collaborative video retrieval research community that shares methods and results; and to create in a collaboratively developed test collection a resource that is of higher quality than any individually developed set.

Finally, for researchers interested in the methodology of retrieval evaluation, publicly available documentation of the test set is essential. This collection, now hosted at the University of North Carolina, (http://www.open-video.org/) was used in the TREC 2001 and 2002 video retrieval tests.

That a new community focused on a novel area of information retrieval research could come together under the TREC umbrella is encouraging for the image retrieval initiative, for it shows that some aspects of the TREC method are extensible to other types of information retrieval challenge.

6.1.2 Emerging TREC Communities

6.1.2.1 Genomics

William Herch of Ohio State University is developing a TREC track focusing on retrieval within the genomics literature (http://medir.ohsu.edu/~genomics/). This is a much more traditional TREC-based exercise that focuses on the definition of questions that reflect a particular disciplinary discourse and the development of a

test data set against which these questions can be put. Its popularity has led to the development of a procedure for proposing new TREC tracks.

6.1.3 An Image Retrieval TREC Track?

A request could be made to NIST to expand TREC to include a rack devoted to image retrieval. Previous expansions in areas such as sound associated with video files are an indication that this should be possible.

However, given TREC's interests, it is likely that a TREC image track could become focused on metadata-based searching methods. It is not entirely clear whether the TREC context, which is text based, could accommodate the needs of the CBIR community, and hence the needs of the metadata-based image retrieval community, for a common benchmarking environment. It is also unclear whether a track within TREC would provide the community focus needed to "gel" the image retrieval field.

However, the TREC foundation is the strongest available, and its status would nearly ensure success of the venture to some extent. Expertise in benchmarking is readily available within the TREC community and should be drawn upon as an image retrieval benchmarking service is developed.

6.2 Genesis from within the Computer Science Research Community: Benchathlon Expansion

Benchathlon is a loosely organized, nonprofit group based in Geneva (http://www.benchathlon.net). Its members seem to be more interested observers than active participants. The organization's principal activities seem to be conducted by the group at the University of Geneva.

The Benchathlon group is trying to develop a CBIR benchmarking activity in conjunction with the Internet Imaging Conference, and it has hosted events at the conference for several years. The weakness of this model is that the community of researchers lacks the funding required to support an ongoing effort of this scale. In addition, expanding this activity raises the same problems as building on a TREC

approach does: There is a vested community with methods based in only one of the two approaches to image retrieval. The people involved in this effort should definitely be brought into any further activity because they have thought a great deal about the problem and have already identified numerous potential participants.

6.3 New Service Created by Funding Bodies

Foundations and governments currently funding research in image retrieval or concerned about the applicability of results of these studies could lead the creation of a new function within an existing organization. Or, a freestanding, not-for-profit organization could be formed to manage the benchmarking service. The latter would have the advantage of being a neutral place where content-based and metadata-based researchers could meet with digital librarians. It would also enable linkages between funding and use of the test data set in benchmarking results.

The disadvantages of an organization dedicated to this purpose include its need to establish itself legitimately and to build itself "from scratch." Housing the new initiative within an existing organization would provide needed support during incubation and added credibility. Such an organizational host should not, however, be closely affiliated with one or the other communities of image retrieval researchers. This might jeopardize the "bridging" goal.

6.3.1 Music Retrieval

An example of how foundation support can coalesce a community can be found in the area of music retrieval. J. Stephen Downie is leading an initiative to establish a music information/music digital library evaluation framework (music-ir.org). Using community consultation methods, Downie has built a consensus around music retrieval problems and possible methods to move that particular information retrieval community forward. White papers have been solicited from stakeholders, and several discussions and workshops held at music information retrieval conferences (JCDL 02 and ISMIR 02, SIG-IR 03, documented in Downie 2003c). Having garnered significant support from the NSF and The Andrew W. Mellon Foundation, those involved are moving toward the creation of an International

Music Information Retrieval System Evaluation Laboratory (IMIRSEL) at the University of Illinois at Urbana-Champaign (Downie 2003b).

The goals of this group are similar to those of the group discussing an image retrieval research database. They include "a formalized set of evaluation methods," "a set of test music databases of substantial size, and varied content" so that researchers may "properly compare and contrast techniques under a variety of scenarios" (Downie 2002a). The group has criticized some of the TREC methodologies as being focused too narrowly on system performance. It seems to be moving toward more open-ended, user-centered evaluation methods. How these will be supported by a test data set remains to be determined. Preliminary architectures for a test system are outlined in Downie (2003b).

6.4 An Industry Consortium

Benchmarking is often undertaken by a professional association, an industry trade group, or government regulators. However, an industry-led consortium model does not seem to fit with the nature of the image retrieval community. Since implementation has not proceeded very far, there is an immature image retrieval industry that is unlikely to be able to support a benchmarking service. Multiple professional associations represent the different directions of research, and the community of consumers does not have the ability to require benchmarking. However, industry interest is strong, and its involvement should be solicited from image retrieval and asset management systems vendors.

7. STAGES IN DEVELOPING AN IMAGE RETRIEVAL BENCHMARK DATABASE

7.1 Phased Approach

It is proposed that development of an image retrieval benchmarking service be undertaken in four phases. The successful completion of each phase is a precondition for proceeding to the next. This will ensure that a solid foundation is

build for the initiative and that it does not get too far ahead of the community it is designed to serve.

The four phases are

1. establish a case, identify sponsors, and recruit research participants
2. establish organization
3. launch service
4. operationalize service

Rough cost estimates, based on an organizational model of an independent benchmarking function—the genesis of an independent entity and established within an existing organization—are provided for each phase. These are offered purely to establish the level of support required and to determine whether there is adequate commitment on the part of funders, interest on the part of stakeholders, and willingness on the part of end beneficiaries to forgo other research that might receive these funds if the benchmarking resource were not constructed.

7.2 Phase 1: Establish a Case, Identify Sponsors, and Recruit Research Participants

Goals: Test receptiveness of community to concept; identify concerns; identify partners and participants; gain buy- in; validate need and approach

The first step is to establish whether the potential sponsors and stakeholders believe that a benchmarking service, which could cost more than $1.5 million to establish before its first cycle of results is reported, has sufficient potential value to warrant detailed planning.

If so, the next step is to organize a steering committee. Functions of this committee would be to spell out a clear and convincing definition of the project and to obtain commitment to it from the principals of major research groups in the image retrieval field worldwide.

Timeframe: One year

Funding Required: About $250,000

Fixed costs: Consultant facilitation, reporting, travel

Variable costs: Workshop participant travel, research, community perspective reports (could be contributed white papers at little or no cost).

7.2.1 Form Steering Committee

A group of high-profile representatives of the content-based and metadata-based image retrieval communities should be recruited to serve on the project steering committee. These individuals should be joined by representatives of the digital library community and the evaluation/benchmarking community. This group would initially provide direction for the consultant, solicit input into the community consultation, and participate in shaping the nature of the image retrieval benchmarking service.

7.2.2 Issue Request for Comment

A formal request for comment (RFC) on the feasibility of an image retrieval benchmarking service should be issued. (The text of this report could serve as the RFC.) Input from all interested researchers, organizations, and institutions should be solicited in the form of written comments on the study and white papers.

The RFC should present the concept of a benchmark database for image retrieval research within the communities of content- and metadata-based researchers, as well as in digital libraries community for a response. It would test the waters for interest in pursuing the idea and could identify partners in further stages of the effort.

7.2.3 Hold Workshops

If responses to the RFC are sufficient and interest is shown in pursuing the articulation of the nature of an image retrieval benchmarking service, a series of workshops should be held to examine responses to the RFC and probe the needs of specific communities.

The workshops should address the CBIR community, the metadata-based image retrieval community, the digital library community, and the community creating digital image resources. Venues could be solicited at conferences such as Internet Imaging, the Joint Conference on Digital Libraries (JDL), ACM's SIG-IR [spell], the European Conference on Digital Libraries (ECDL), the International Conference on Image Processing (ICP), the European Conference on Computer Vision, the Computer Vision and Pattern Recognition (CVPR), or Challenge of Image and Video Retrieval (CIVR).

Alternatively, funding could be sought for several invitational workshops that brought together a cross-section of participants from each of these areas in separately convened venues. There are pros and cons to each approach.

The project consultant would facilitate each workshop. Participants would discuss the RFC and the responses received, address the issues raised there, bring to light critical success factors for the image retrieval benchmarking service, share research to date, and document tasks for the future.

7.2.4 Draft Implementation Plan

Drawing on the responses to the RFC and the discussions of the workshops—and only if a consensus exists or can be developed—a full implementation plan and proposal requesting funding would be developed. The proposal would be predicated on the model favored in the workshops and RFC responses. It would include a budget for the first two to four years.

7.3 Phase 2: Establish Organization

Goal: Create a prototype collection and skeleton service

Timeframe: 18 to 24 months

Funding Required: About $250,000 per year

Fixed costs: Advisory committee meetings; project manager, administrative assistant, metadata expertise, and evaluation expertise; programming,

computer, and network services; data storage; technical support; travel; reporting/communications/Web site

7.3.1 Establish Governance

7.3.1.1 Create Advisory Board

At this phase, the steering committee would be reconstituted as an advisory board for the emergent organization. The board would be responsible for determining the direction of the organization, identifying an appropriate focus and location for its activities, determining the nature of initial tests of image retrieval benchmarking, establishing terms of participation, and maintaining community liaison.

7.3.1.2 Hire Project Director

The board would identify as project director an individual with knowledge of image retrieval (both metadata based and content based) and with experience in developing organizations and collaborations that could manage the development of the benchmarking service. This person would work with the board to identify an appropriate hosting relationship with an organization willing to incubate the service.

7.3.2 Issue Request for Proposals for Host

Responsibly for data set development must be placed either within the funding organization or within a university with strong programs in information science and retrieval evaluation, but outside of a group conducting image retrieval research per se. A request for proposals (RFP) for hosting such a service might not only produce the lowest long-term costs but also expose differential cost issues that might otherwise go overlooked. Groups with expertise in measurement and evaluation and in construction of research data sets would be better suited to create and maintain the resource than would groups interested primarily in image retrieval, who will be the primary users.

Responses to the RFP will be submitted to the project director and assessed by the board.

7.3.3 Issue Call for Participation

The board and director will issue a formal invitation to all image researchers and image database creators to participate in the development of the image retrieval benchmarking service. If a consortium model prevailed in the planning, the terms of consortium membership would be articulated here.

7.3.4 Issue Call for Data Sets

A call for data sets to form the nucleus of the image-retrieval benchmarking database would be issued. The board and the project director would need to determine the size of the data sets needed and to establish a method of selection. The framework for database assessment should be shared with the community.

Terms and conditions for the deposit of image data sets, as well as for research use, need to be established at this time, to ensure that no data are collected without a clear understanding of any associated intellectual property rights.

The board and the project director would evaluate the responses to the call for data sets.

7.3.5 Prototype Integration of Initial Data Sets

Once the responses have been evaluated and key data sets identified, the benchmarking service should prototype the acquisition and integration of a small number of data sets. Steps include the following:

- acquire means to ingest and store data
- identify administrative, technical, and descriptive metadata required
- review available metadata from each image set and identify enhancements needed
- map the metadata from each data set to a minimum common denominator, such as the Dublin Core, but keep original, richer, domain-specific metadata (in database- or domain-specific extensions)

- ensure that all images are captured according to a basic specification (at an established minimum, or higher, resolution and bit density and in a known color space)
- ensure presence of required technical metadata

7.3.6 Establish Test Queries

A preliminary list of questions to be posed to the prototype data set should be developed. These are likely to be very basic questions, without much domain relevance; complex questions are unlikely to be answerable given current state of technology. "Starting simple" will provide a way to speed the availability of results.

7.3.7 Establish Test Ground-Truth Assessments

Ground-truth assessments will need to be made for each of the queries established. Methods of assigning relevance should be tested and compared, and ways to use metadata to assist in making the relevance judgments prototyped. Means of identifying image subsets should be explored. A methodology for establishing ground truth for the data set, in whole or part, should be set.

7.3.8 Release Test Data Sets without Ground Truth

One possible strategy to maintain involvement of community, spur development, and test TREC's idea of pooling as a strategy to establish relevance is to release the data set for use prior to the establishment of ground truth. Relevance assessment could proceed in parallel with research and could take results of research as input.

7.4 Analyze and Report Prototype Results

The results of this prototype phase should be reported and assessed, and a full plan developed prior to the launch of an operational service.

7.5 Phase 3: Launch Service

Goal: Conduct first rigorous benchmarking tests and report results

Timeframe: 18 to 24 months
Funding Required: About $400,000 per year

Fixed Costs: Director, project manager, programmer, metadata expert, indexer(s); computer and network services; data storage facility; evaluation expertise; data storage; technical support; travel; advisory committee meetings; reporting/communications/Web site; offices
Variable: Content acquisition (if included in the plan)

7.5.1 Construct Production Systems

On the basis of the experience with the prototype systems and first test run, a robust repository/server to house sets of image data to be used in research should be constructed.

7.5.2 Obtain Data (Image And Metadata Sets)

Additional data sets should be targeted from the group identified in the previous phase and be integrated into the benchmarking database.

7.5.3 Establish Queries

Three or four high-profile image using domains should be identified, and queries relating to those domains developed in conjunction with the ingestion of data sets. Construction of queries should be seen as an opportunity to solicit end-user input into the direction for image retrieval research.

A query management server would facilitate the ongoing development of queries, assignment of relevance, and assessment of results.

7.5.4 Establish Relevance Judgments

Relevance judgments need to be made for the queries identified. The query management server would provide facilities for recording of query results from the variety of retrieval methods being tested (for comparison over time with results from other methods). The server would also enable the characterization of the data set, or segments of it, which is essential in order to establish the ground truth.

7.5.5 Launch Test

The first full-scale test would then be launched, with time schedules set for the reporting and assessment of results.

7.5.6 Convene Image Retrieval Conference

An international conference should be planned to report results, with papers available as proceedings or a journal issue. This would reinforce the relevance of the image retrieval benchmarking service, raise its profile among the community, and build the need for researchers to use the benchmarking service in future research. A conference provides a way for researchers to meet, to be encouraged to provide data sets, and to update benchmarks to reflect new findings.

7.6 Phase 4: Operationalize Service

Goal: Carry out ongoing image retrieval research supported by the community

By phase 4, the image retrieval benchmarking services is fully operational—registering users and their methods, "lending" image data sets for research, receiving additional image and metadata sets, "filing" search results from participating projects, developing/receiving analytical tools and making them available for use by community, managing research methods/benchmark application peer review process, and reporting/archiving results. An annual Image Retrieval Conference provides a focus for research activity and reporting.

The service plans for ongoing enhancement of services and increases the sophistication of tests over time. Staff and researchers begin exploration of use of assembled data sets beyond the initial scope of image retrieval. The use of the same data set for different tests in a single supportive environment offers a possibility for economies and synergies. A repository for research methods and results is created. Image retrieval research centers around the use of the benchmarking service.

The effects of the image retrieval benchmarking service are studied over time. Has it produced benefits to the field? Have costs decreased? Has retrieval improved? Has the availability of benchmarking data enabled decision making about image retrieval technology deployment in the digital library service environment (for example, has it been possible to make a determination about what is "good enough" to support some kinds of retrieval, on the basis of costs)?

Timeframe: Ongoing

Funding Required: About $500,000 a year

Fixed costs: Director, project manager, programmer, metadata expert, indexer(s); computer and network services; data storage facility; evaluation expertise; data storage; technical support; travel; advisory committee meetings; reporting/communications/Web site; offices

Variable costs: Conference support, research, travel, support for content acquisition (if included in plan)

8. CONCLUSION

Integrating images into the digital library poses a challenge to librarians committed to developing a quality service in an interdisciplinary environment. As resources developed in a single department are used beyond that boundary, existing disciplinary perspectives of retrieval needs and discipline-specific descriptive schema impede use and reuse. Digital librarians feel a strong need to improve access to the visual information that forms part of the digital library; however, metadata-based retrieval systems have not been entirely successful, either because concepts such as "Romantic" are not equivalent across disciplines or because access points deemed key in one area are inconsequential in another. Many librarians hope that retrieval methods that use the content of the image, rather than only the metadata associated with it, might further the use and usability of image collections.

Development of image retrieval research has been hampered by a strong division between content-based and metadata-based image retrieval. CBIR is a large, active area of computer science. Metadata-based image retrieval, in contrast, forms a small, loosely organized area of information science. There is little cross-citation between these communities and little relationship between the theoretical research work in image retrieval and working image retrieval systems deployed in digital library service environments.

Digital librarians currently have no tools with which to evaluate or assess available image retrieval methods or systems. There is no clear path for the deployment of image retrieval tools from research to service delivery. The field faces a significant challenge in developing services based on emerging image retrieval technologies. An image retrieval benchmarking service might provide a means to focus research, assess its results, and further its applicability in the digital library.

Benchmarking involves the comparison of results of two or more methods of performing the same task against the known result of that task. Differing methods are assessed by analyzing how closely they reach the desired result. An image-retrieval benchmarking database would provide a controlled environment within which to explore questions critical to the development of the field. A successful image-retrieval benchmarking database could offer significant benefits, enabling the comparison of retrieval strategies: content based, metadata based, and hybrid (involving both at different points in the process). It could improve the understanding of image retrieval in the digital library community through the careful articulation of research questions and benchmark tests. It could enable researchers in the field of image retrieval to measure the effectiveness of their methods against a common standard. It could provide researchers and potential users of image retrieval systems with a means of assessing the effectiveness of different methods. It could enhance the value of research through comparative methods. Over time, the focus provided by an image retrieval benchmarking service could improve the quality of image retrieval, through comparison, evaluation, competition, and cooperation. It could encourage cross-method

cooperation through the development of an interdisciplinary community of research. Over time, it could reduce costs by improving methods.

Such benefits do not come without a price. The creation of an image retrieval benchmarking service would require the cooperation of a broad community of researchers who were willing to devote time and energy to the development of evaluation methodologies. Image retrieval scientists would be needed to use the test set in their work (research, evaluation, and reporting). Metadata retrieval researchers would be needed to test schemas on a know data set and conduct research, evaluation, and reporting. Disciplinary representatives drawn from areas where visual information forms a key research resource would be needed to express their needs in terms of a series of queries and to establish the relevance of an assembled set of images to those questions. Digital librarians, image curators, and information scientists would be needed to define assessment measures and evaluation frameworks and to evaluate the results of the retrieval systems. Funders that are willing to support the transformation of a discipline over the longer term would be needed.

An image retrieval benchmarking service could further the development of image retrieval science and the delivery of digital library services. However, creating an image retrieval benchmarking service would be a significant undertaking. A benchmarking database is more than a collection of images. Benchmarking requires the development of a set of queries that can be put to that test collection. Each image in the test collection must be assessed to determine whether it is relevant to that query. Assessing performance of systems requires a set of evaluation metrics that can be used to compare one system to another and to rank results. Developing a test collection requires an investment in data collection, documentation, enhancement, and distribution. But most significantly, maintaining an image reference benchmarking service requires the long-term commitment to its use on the part of a community of researchers. Without a community vested in the development of the database, and publishing research based upon it, the collection

remains a chimerical solution to advancing the state of research and improving the retrieval of visual materials in the digital library.

Is the time ripe for the development of an image retrieval benchmarking service? This is a question for the many constituencies required to make the service a success.

[1] While one can imagine a large number of commercial applications for image retrieval technology, this report focuses on the digital library domain. Any move forward with an image retrieval benchmarking service would likely look to the e-commerce community for input and support.