The DLib Test Suite and Metrics Working Group: Harvesting the Experience from the Digital Library Initiative

Ronald L. Larsen
University of Maryland

DARPA's DLib Test Suite project[i] was an early attempt at organizing a rigorous and well-supported test bed to enable comparative evaluation of digital library technologies and capabilities. The test suite, part of the DLib Forum[ii], included a diverse and heterogeneous set of resources deliberately selected to foster research in interoperability. The DLib Forum also sponsored a Metrics Working Group (MWG)[iii] to develop quantitative performance measures. The test suite went largely underutilized by the research community, and the MWG, while making significant progress, found the stated objective daunting. Clearly, much remains to be done in both the conception of effective test beds and the instrumentation to assess progress.

The DELOS Workshop on Evaluation of Digital Libraries provides an opportunity to make further progress in this important area, engaging an international community and building on the collective experience accumulated from a larger and more diverse set of digital library projects. In this paper, the DLib Test Suite is briefly reviewed and the progress of the MWG is described.

The DLib Test Suite[iv]

The DLib test suite was conceived to address three needs: (1) lowering the barriers of entry for digital library researchers requiring access to large collections and information management services, (2) providing standard sets of data for quantitative and comparative research, and (3) supporting a distributed environment of heterogeneous resources organized to support interoperability experiments.

Six individual digital library projects participated in the test suite[v]:

- Carnegie Mellon University's Informedia Digital Video and Spoken Language Document Testbed (digitized and cataloged televised news)
- Cornell University's Networked Computer Science Technical Reference Library (computer science technical reports in a globally distributed set of repositories)
- UC Berkeley's Environmental Digital Library (images, databases, and scanned documents)
- UC Santa Barbara's Alexandria Digital Library (maps, images, and geo-located documents)
- The University of Illinois at Urbana Champaign's Desktop Link to Engineering Resources - DeLIver (online access to scholarly publishers' journals)
- The University of Tennessee – Knoxville's Netlib and the National High-performance Software Exchange (software, numerical databases, and accompanying documentation)

Together, these provided a very diverse set of information resources, services, and interfaces… an environment suitable for creative exploration of interoperability issues.

The model for interoperability experiments among the test suite participants was based on the view of digital libraries as *repositories* of *digital objects*. A *digital object* has a unique and persistent identifier, key metadata describing it, a data stream that can be invoked as a typed sequence of bytes, and a disseminator to map the data stream into a particular form for delivery. A *repository* instantiates digital objects and supports their use in a network environment. It also implements a level of abstraction over the underlying storage mechanisms and provides a secure environment for the management and use of digital objects. Fundamental to the operation of the repositories was a common *Repository Access Protocol (RAP)* that guarantees the integrity of digital objects and facilitates interoperability among repositories.

The DLib Metrics Working Group

The principal focus of the DLib MWG was on information discovery with a human in the loop, in which the information sought is distributed among a heterogeneous set of sources. The objective was to define a set of scientifically rigorous metrics and measures that would enable comparative evaluation of information discovery techniques and algorithms that yielded repeatable results over multiple experiments. As stated by Bill Arms[vi], "It should be possible for other researchers to repeat experiments, with different data and different implementations, and to replicate the basic results. The result should be evaluated against relevant, repeatable criteria, so that strengths and weaknesses of alternative approaches can be compared and improvements measured."

The MWG was chartered to consider evaluation issues in the system, user, and content domains. At the systems level, interest focused on interoperability, scalability, heterogeneity, reliability, and integration. At the user level, issues of relevance, specificity, timeliness, effort vs. effect, and usability dominated. In areas of content, measures of sufficiency, currency, and quality were sought.

Scenario-based evaluation was anticipated, and much of the work of the group ultimately revolved around definitions of canonical scenarios. A scenario was defined to include abstract classes, specific instances of those classes, and a common method of scoring. The use of simulation models as well as measurement of real systems was envisioned.

Three sub-groups were identified, although only two ultimately convened. The first sub-group, on metadata issues, focused specifically on metadata for interoperability or sharability, recognizing a spectrum of interoperability issues. Interoperability among systems using a common standard is clearly the easiest, but rarely fully achievable. More realistic is the expectation of a base standard with extensions to accommodate the specific characteristics of a particular collection or system. The most difficult is clearly interoperability among systems supported by fully divergent metadata sets. Metadata interoperability is required to support: (1) search and retrieval, (2) intellectual property rights management, (3) administration and preservation, and (4) evaluation and use. A system's ability to support interoperability in these areas is fundamentally dependent on the quality of the metadata, and the sub-group explicitly dealt with a range of metadata quality issues, including: (1) specificity, (2) completeness of fields, (3) syntactic correctness, (4) semantic correctness, and (5) consistency, as implemented through authority control.

The second sub-group addressed user-level issues and documented their progress in a series of reports, including (vii) and (viii) addressing scenarios for information discovery, dissemination, library administration, system operation, and other related activities. While the initial charge to the sub-group sought scenario-independent metrics, analogous to precision and recall for information retrieval, such metrics proved to be beyond reach. Scenario-based metrics, while less general, appear to be the best we can achieve at the current state of technology.

The third sub-group was intended to capture the interests of publishers more directly, but in the time the MWG had to conduct its work, the publishers' sub-group was unable to assemble. While the other two sub-groups made efforts to address issues of the publishing community, the development of effective metrics and test beds would benefit from their continued engagement.

DL Metrics

The range of potential metrics relating to digital libraries is immense. In the process of focusing their efforts, the MWG identified at least seven dimensions against which performance could be measured[ix]. These included: (1) system-wide vs. individual services, (2) user interaction vs. underlying system operation, (3) effort vs. effect (net benefit), (4) snapshot vs. session (temporal granularity), (5) capability vs. utility, (6) single user vs. scalability, and (7) collections and content vs. system capability and utility.

The MWG defined a framework for evaluation that addressed the two fundamental phases of an information-seeking scenario, *query* and *retrieval*. For each of these phases, four factors were considered: (1) timeliness, (2) sufficiency, (3) correctness, and (4) effort. The objective became one of identifying indicators for each of these factors that incorporate appropriate (and measurable) metrics. *Timeliness* clearly focuses on speed, and considers both objective measures (e.g., actual elapsed time to complete an operation) and subjective measures (e.g., the user's perception of how long it takes to complete an operation). *Sufficiency* measures the adequacy of the system's response to queries. Recall is the best-known objective measure, but it is typically applied to well-defined finite test collections. Scaling its use up to distributed heterogeneous digital libraries, even through a supported test suite, represents a significantly larger level of effort than was required for the evaluations conducted under the Text Retrieval Conferences (TREC)[x], for example, and yet these have been relatively costly affairs. Alternative means of building instrumented test collections may be required, or different measures entirely may be needed. Sufficiency also has its subjective element. Did the information seeker view the system's responses as adequately comprehensive for the intended purpose? As sufficiency is to recall, *correctness* is to precision. Correctness is intended to gauge the percentage of returned digital objects that actually are appropriate to the query. Subjectively, one asks the user if the returned objects are right, credible, useful, or reliable. Finally, *effort* addresses the amount of work required by the user to interact with the system, frame the appropriate query, and acquire the objects desired. Objective metrics could address search complexity, including the number of times the user must interact with the system or iterate the query to get it "correct." Subjective measures would consider the user's perception of the level of effort required. Is the system perceived as "easy" to use, for example, or does the user leave in despair, finding the system's operations to obscure to comprehend?

Query-phase Metrics

Each of the four factors (timeliness, sufficiency, correctness, and effort) has a number of components particular to an operation's phase. The MWG found the query phase to pose a rich set of questions with opportunities for metrics. Considering timeliness, for example, leads to measuring the time required to prepare an adequate query, the time for the system to respond to the query, the perceived responsiveness of the interface mediating between the user and the system, the currency of responses (are they current, or up-to-date, as judged by an informed observer?), and the novelty of the responses (would an informed observer recognize them as new or particularly relevant?).

For sufficiency, a measure identified as *availability* is the proportion of sources that the digital library has direct or indirect access to that an informed user would judge as relevant to a particular query. *Interface guidance* addresses the degree to which the system offers useful guidance or options for alternative query formulation. *Coverage* refers to the breadth of system resources that contribute to building the set of returned references. *Actual* and *perceived recall*, as discussed earlier, measure the comprehensiveness of the set of returned items, as measured against a standard, and as perceived by the user.

For correctness, in addition to the traditional *precision* measure (both perceived and actual), the MWG included *interface power*, by which was meant the ability of the user interface to suggest more powerful and correct search terms, strategies, or tactics. A measure of this could be the proportion of suggestions that are actually chosen by the user and that substantially contribute to the resulting set of appropriate responses. Another metric considered was *redundancy*, which measures the proportion of responses that duplicate other material in the same set. As with precision and recall, redundancy can be measured objectively and subjectively (did the user notice actual redundancy or perceive redundancy that was not present?).

Considering effort, potential measures include *interface usability*, *query complexity*, and *response complexity*. Interface usability assesses both objective and subjective measures of a user's ability to efficiently and effectively construct and submit an accurate query. Objectively, one can count the number of queries constructed in the process of finding the sought material or measure the time taken to complete the search. Subjectively, one can ask the user to rate the relative difficulty of using a particular interface. Query complexity is intended to assess the difficulty of formulating the appropriate query for a particularly abstract or complex problem specification, including, for example, the number of search terms required, or the number of iterations required to formulate the successful query. Response complexity attempts to measure the difficulty the user has in interpreting the returned query response, either by subjective evaluation, or by measuring the time required for the user to take the next step.

Retrieval-phase Metrics

Retrieval is taken to mean the delivery of disseminations of requested digital objects identified as a result of performing a query. The same four overall measures are suggested. For timeliness, the *dissemination time* (the time between the user requesting a dissemination and its presentation to the user) would be measured. For sufficiency, one could consider a metric such as *presentation appropriateness*, where attempts are made to disseminate retrieved objects in a form tailored to a particular audience. This could be as simple as recognizing that a .pdf document will be more useful than a .txt one, or as complex as translating a document into a different natural language. Another sufficiency metric for retrieval is simply *retrievability*. How many of the references returned refer to actual retrievable items?

The correctness metric is *retrieval correctness*, and is defined as the probability that a retrieved dissemination is, in actuality, the correct one. Effort is measured as *selection effort*, or the difficulty the user encounters in selecting or extracting desired disseminations of digital objects from the set of references returned from a query.

Summary

The DLib Metrics Working Group summarized its evaluation metrics for distributed digital libraries as shown in the following table. Much work remains to be done to realize these in viable digital library test suites and to develop valid and understandable comparisons useful to the digital library community.

Evaluation dimensions	Query	Retrieval
Timeliness	Query preparation time Query response time Interface responsiveness Currency Novelty	Dissemination time
Sufficiency	Availability Interface guidance Coverage Actual recall Perceived recall	Presentation appropriateness Retrievability
Correctness	Response correctness Interface power Actual precision Perceived precision Redundancy	Retrieval correctness
Effort	Interface usability Query complexity Response complexity	Selection effort

Table 1 Evaluation Metrics for Distributed Digital Libraries

[i] See http://www.darpa.mil/ito/psum1999/G834-0.html for the DARPA Project Summary

[ii] See http://www.dlib.org for information on the DLib Forum

[iii] See http://www.dlib.org/metrics/public/index.html for the record of the MWG's work

[iv] See http://www.dlib.org/test-suite/ for more information about the DLib Test Suite

[v] The Corporation for National Research Initiatives (http://www.cnri.reston.va.us/) managed the test suite and supported researchers using it.

[vi] Arms, William Y., "Replication of Results and the Need for Test Suites," (http://www.dlib.org/metrics/public/PositionPapers/arms.html), January 1998.

[vii] Leiner, Barry M., Leah Lievrouw, Tassos Nakassis, and Mike Sullivan, "Seeker Scenarios for Distributed Digital Libraries," (http://www.dlib.org/metrics/private/papers/Seeker_Scenarios-rev4.htm), October 1999

[viii] DLib Working Group on Digital Library Metrics,"Digital Library Challenge Problems and Metrics" (http://dlib.org/metrics/private/papers/Challenges.html), September 1998.

[ix] Leiner, Barry M., "Types of Digital Library Metrics," (http://dlib.org/metrics/private/papers/types.html), January 1998.

[x] See http://trec.nist.gov/ for complete information regarding the Text REtrieval Conference (TREC)