CoLIS 3 Workshop on Metrics

Thursday 27 May 25, 1999
Chair, Edward A. Fox

Resources available from: D-Lib Working Group on Digital Library Metrics

Outline:

- Summary of D-Lib Working Group Discussions and Documents
- Adding an LIS Perspective to the D-Lib Activities
- New Measures?
- New Scenarios?
- Projects?
- Experiments?
- Summary Report

D-Lib Working Group on
Digital Library Metrics

D-Lib Working Group on Digital Library Metrics

This Working Group is aimed at developing a consensus on an appropriate set of metrics to evaluate and compare the effectiveness of digital libraries and component technologies in a distributed environment. Initial emphasis will be on (a) information discovery with a human in the loop, and (b) retrieval in a heterogeneous world.

Working Group Charter
Other Working Group Documents
Working Group Private Area

This is an open working group, and anyone interested in the subject and in contributing to the work of the group is encouraged to join. For further information or to join the group, contact Barry Leiner <BLeiner@cnri.reston.va.us>.

The Working Group recently sponsored a Workshop on Digital Library Metrics, organized by Bill Pottenger and Bob McGrath, and held 27 June 1998, just after the DL'98 conference, .

The D-Lib Program is based at the Corporation for National Research Initiatives and is sponsored by the Defense Advanced Research Projects Agency (DARPA) on behalf of the Digital Libraries Initiative under Grant No. N66001-98-1-8908.

prepared by Barry Leiner
last modified 8/25/98
bl/bw

Working Groups

D-Lib Working Group on Digital Library Metrics

CHARTER

Purpose

This Working Group is to develop a consensus on an appropriate set of metrics to evaluate and compare the effectiveness of digital libraries and component technologies in a distributed environment. Initial emphasis will be on (a) information discovery with a human in the loop, and (b) retrieval in a heterogeneous world.

Background

Much of digital library research is experimental or exploratory. Research projects lead to demonstrations, pilot systems, and eventually to deployment in production systems. Currently, there are few ways to evaluate the effectiveness of research, or to measure progress towards long-term goals. A notable exception is information retrieval, which has been greatly enhanced by the existence of the well-established measures of precision and recall. These metrics, in conjunction with standard corpora that can be used for testing and evaluation, have helped further the state of the art by allowing researchers to do comparisons and evaluations on a fair comparison basis.

While these measures have been very useful in evaluating and comparing "single site" search and retrieval mechanisms, the richness of the digital libary environment demands a much richer set of metrics. Metrics are required to deal with issues such as the distributed nature of the digital library, the importance of user interfaces to the system, and the need for systems approaches to deal with heterogeneity amongst the various components of the digital library.

Working Group Objectives

The objective of this working group is to develop usable and useful metrics for the more complex world of distributed heterogeneous digital libraries. A parallel initiative plans to establish a test suite of library collections that can be used for collaborative research and as a set of corpora for evaluation. Efforts to accumulate evaluation software for standardization and reuse will also be encouraged.

The working group will initially focus on metrics for information discovery and retrieval. Information discovery in digital libraries is more complex than the classical problem of information retrieval. One difference is that there is a set of

seeking tasks with varying criteria for success. In practice, most information discovery is an iterative process that includes searching, browsing, filtering, and other complex interactions between human understanding and computer processing. Current metrics measure the performance of discrete steps in this process, but not the overall success.

Likewise, retrieval in the distributed digital library environment requires increased attention. As library objects become more complex, it becomes increasingly common for a user to discover the existence of an object, but not be able to access it effectively. Problems range from incorrect references (broken links), access restrictions, mismatches between the MIME types that are supported, system incompatibilities, and so on.

Because of the importance of the human in the loop, we expect to draw metrics from a broad set of relevant fields, including but not limited to those as diverse as psychology, engineering, and human communications. Our ultimate objective is to be able to measure and document the impact of particular system concepts or features, in specific settings, for specific user communities with specific purposes. Therefore, ideal metrics will be meaningful to all stakeholders, reproducible, and inexpensive.

The D-Lib Working Group on Digital Library Metrics is co-chaired by William Arms and Barry Leiner

Last modified 5/7/99 bw

Working Groups

D-Lib Working Group on Digital Library Metrics: Other Documents

- The Scope of the Digital Library. Draft Prepared by Barry M. Leiner for the DLib Working Group on Digital Library Metrics, January 16, 1998. Revised October 15, 1998.

  At the kickoff meeting of the DLib Working Group on Digital Library Metrics (WG), held January 7-8, 1998, at Stanford University, some discussion was held as to what did we mean by the term "digital library". We concluded that it would be valuable for our own deliberations to document a common understanding of the term, but agreed that such an understanding could only be for the purposes of our deliberations, i.e., we could not and would not aim for a general consensus. This document is intended for that purpose, and this draft is intended to start the discussions.

- Position Papers from Kickoff Meeting

  In preparation for the kickoff meeting of the Working Group, held 7-8 January 1998, members were encouraged to submit position papers. These are some of those received:

  Replication of Results and the Need for Test Suites
  (Position Paper by William Y. Arms)

  Technological and Social Change and Implications for Digital Library Metapolicy
  (Position paper by Bob Este)

  Position Statement - Approaching D-Lib Metrics
  (Position paper by Ed Fox)

  White Paper
  (Paul B. Kantor)

  Digital Library Metrics: Uncertainty and Failure
  (Initial Draft by Carl Lagoze)

  Evaluation of Concept Space and Category Map Semantic Indexes
  (Position paper by William M. Pottenger, Bruce R. Schatz, Duncan H. Lawrie, Robert E. McGrath

---

The Scope of the Digital Library

Draft Prepared by Barry M. Leiner
for the DLib Working Group on Digital Library Metrics
January 16, 1998
Revised October 15, 1998

Preface

At the kickoff meeting of the DLib Working Group on Digital Library Metrics (WG), held January 7-8, 1998, at Stanford University, some discussion was held as to what did we mean by the term "digital library". We concluded that it would be valuable for our own deliberations to document a common understanding of the term, but agreed that such an understanding could only be for the purposes of our deliberations, i.e. we could not and would not aim for a general consensus. This document is intended for that purpose, and this draft is intended to start the discussions.

Introduction and Background

The term "Digital Library" has a variety of potential meanings, ranging from a digitized collection of material that one might find in a traditional library through to the collection of all digital information along with the services that make that information useful to all possible users. As the WG discussed possible scenarios and challenge problems to drive our discussion of metrics, we found the need to come to at least a loose agreement on the scope of the digital library. This document is intended to serve that purpose.

Much of the question about the scope of the term is how broad a view should be taken of the digital library. Does it encompass all of information management or is a more tightly constrained view appropriate? In this document, and for the purposes of the deliberations of the WG, we choose to take a very broad view. This is driven by the recognition that to do otherwise would require setting boundaries that are fairly artificial.

The structure of this document is as follows. In the first section, a brief definition of the term "digital library" is given, as a set of characteristics. The remainder of the document elaborates each of those characteristics.

Definition

At the kickoff meeting of the WG (held January 7-8, 1998 at Stanford University), the following definition was proposed:

The Digital Library is:

- The collection of services

- And the collection of information objects
- That support users in dealing with information objects
- And the organization and presentation of those objects
- Available directly or indirectly
- Via electronic/digital means.

The collection of services

A digital library is much more than just the collection of material in its repositories. It provides a variety of services to all of its users (both humans and machines, and producers, managers, and consumers of information). Thus we start our definition with the notion of the collection of services that the digital library represents. There are a large and varied set of such services, including services to support management of collections, services to provide replicated and reliable storage, services to aid in query formulation and execution, services to assist in name resolution and location, etc.

The collection of information objects

The basis for a digital library, however, must be the information objects that provide the content. A basic characteristic of the digital library is that the information objects are found in collections with associated management and support functions. The types of information objects vary from traditional "documents" through to live objects (e.g. sensor readings) or dynamic query results.

Supporting users deal with information objects

The goal of the digital library is to assist users by satisfying their needs and requirements for management, access, storage, and manipulation of the variety of information stored in the collection of material that represents the "holdings" of the library. Users may be humans or they may be automated processes acting on behalf of or in support of human needs. Users also vary and include those who are "end" users (those not involved in the management and operation of the library but rather are the customers), library operators, and information "producers" who want their material available through the library.

The organization and presentation of those objects

The key to effective collections management is to implement simple structural organizations and be able to present those organizations in a way that library users find useful and can understand easily. In traditional libraries, books are primarily stored by subject, title, author, and date, and accessed by following signs to the appropriate floor, room, bookcase, shelf, and spine-labeled book. The size and relative celebration of each portion of the collection gives patrons information about the collection and can reveal the library's collection management objectives as well.

A library is created to serve a community of users. Users who participate in the digital library should be

aware of its design and be able collectively to refine that design to better serve their own information needs. Therefore, the ongoing human usability of a digital library depends on the clear and unobtrusive exposure of the library's design, its near-term goals, and its overall objectives.

Furthermore, digital libraries should continue the ongoing tradition of coupling utility with aesthetics in the organization and presentation of materials.

Available directly or indirectly

These information objects may be digital objects or they may be in other media (e.g. paper) but represented in the library via digital means (e.g. metadata). They may be available directly over the network (e.g., using a query service of the library to find and then retrieve electronically the information object) or indirectly (e.g., the result of the query may give instructions on how to obtain the object, but that is done outside the scope of the library itself.)

Electronic/digital availability

Although the objects may not even be electronic, and although the objects themselves may not be available directly over the network, the objects must be represented electronically in some manner through, e.g., metadata or catalogs. Otherwise, we would not consider the objects to be part of the *digital* library.

White Paper

Paul B. Kantor

Rutgers

This note addresses some complexities which cannot be avoided, and suggests a way to approach them with some logical coherence. It is not intended to oppose the "poor but simple" thrust, but to help us focus attention on what aspects of the evaluation problem should be respected on the road to honest poverty.

Scenarios.

For an evaluator, scenarios play the same role that the "market basket" plays for the economist. Our basket of scenarios needs to contain some examples of each of the "economic necessities" of users of digital libraries. This raises two problems: (1) identifying the necessities, abstract classes (2) deciding which specific instances should be used to represent each of the important classes. Whatever method of scoring performance is eventually selected, it must be applicable to elements from each class independently. The problem of producing a single overall measure can then be "evaded directly" by encouraging each user of the evaluation process to select a rule of combination (for example, a weighted sum of the scores on several types of tasks) with a form (the specific weights) which reflects the importance of the corresponding benefits and costs to that particular user of the evaluation. The MWG may of course suggest some typical baskets and weights representing say, an elementary school teacher, an industrial spy, or a disaster relief planner. I am working with Linda Hill of UCSB to develop a classification scheme for GIS tasks, which already has some 6 distinct dimensions.

It is important, however, that the suite of tasks retain its essential "vectorial" character, so that the basic elements of score are available to those who would reweight them in a different way. This ensures that the effort of measurement, which will be substantial, is not lost in some simple condensation of all that is known.

It is important to seek (if there are any) the *ceteris paribus* monotonic contributors to the overall "goodness" of a system. In the language of Orr, these are parts which, when done more well, can only make the overall system do "more good". A natural candidate would be improvements in the speed of algorithms which accomplish the same calculation on the same processor. Perhaps, from the human side, improvements in monitor resolution, at no added cost, would fall into the same category. On the other hand, squeezing more information onto the screen might or might not bean improvement from the user’s point of view.

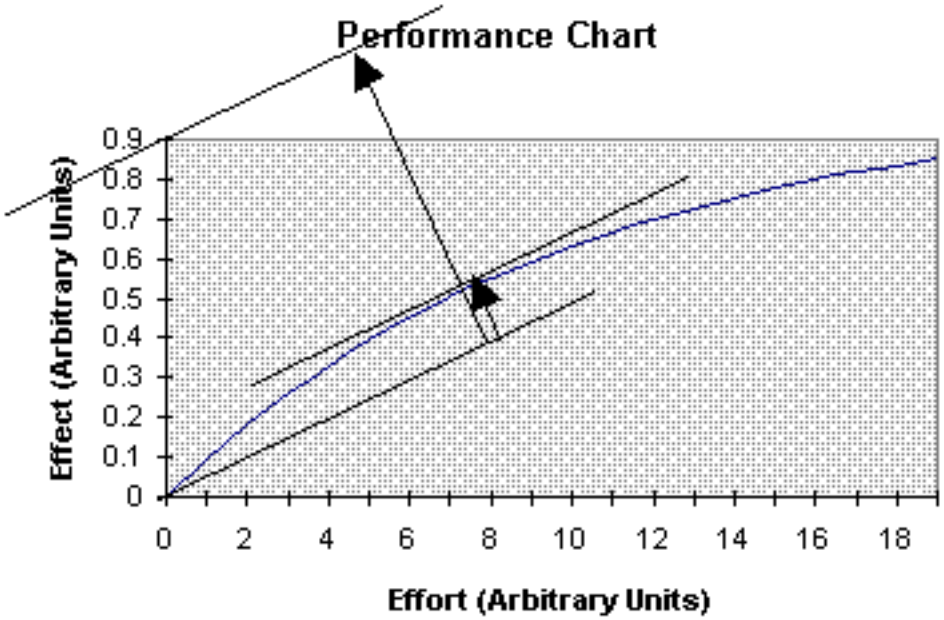
Measurement and observation.

Much current discussion of digital libraries, in their working environments, emphasizes the importance of "observation". Observation is quite open-ended, and can lead wherever the skills and insights of the investigator are able to take it. [Belkin and collaborators at Rutgers]. It is the essence of qualitative method, when done well. Observation is an essential step in measurement. In fact it occurs twice: once in deciding what to measure, and once in making the observations that provide that measure.

A central issue in developing measures for digital libraries is that "performance" viewed even as an abstraction, is not a

"quality" but a "relation". For example, the acceleration of a sports car is given in terms of the number of seconds it takes to reach 60 mph. but this is simply one point on the curve of speed versus time (or distance). The reality is more complex. Operating manuals for airplanes give several key speeds for climbing, corresponding to maximum rise versus distance, maximum rise versus time, and maximum rise versus fuel expended. And an airplane (at least a Cessna 172) is a good deal simpler than a digital library.

At a minimum, I suggest that we approach each dimension by looking for a pair of variables, whose relation over the entire range of likely use, characterizes the performance of a library, or of a component of that library. Generically, we can call these "Effect versus Effort" curves. If the system is "efficient" in a certain natural sense, these curves will be concave, as shown in the example below. That is, the greatest benefit comes early, and further benefit is available for those who wish to persist into regions of diminishing returns.



The Human in the Loop

For measuring how a system interacts with the human in the loop, this concept specializes to selecting a particular measurable definition of the effect, and of the effort. If the users are "sufficiently homogeneous" then the time they must expend is probably a good surrogate for effort. For effect we hope to go beyond a naive counting of the number of retrieved items that might be judged, by some experts, to suit the stated purpose or problem. Ideal would be surrogates for an end-to-end approach, in which there is some task to be performed, and the measure of effect corresponds to seeing how much the digital library advances that task, or how well the task is performed.

It is also possible to devise inverse schemes (as was used in the TREC5 Confusion Track) in which the task is to find a known (in that setting, presumably "vaguely remembered" document. The measure of effort (for a basket of tasks) is the total effort required of the user. The measure of effect is the total number of target documents found. [See Voorhess and Kantor, TREC5 proceedings]. With a body of representative users, it is possible to convert the end-to-end approach into a group of known item tests. That is, we can develop estimates of how well the task is done when item I is found, for each of

several useful items I, and then track the effort required to unearth each of them. We must note that items will, in most real cases, be not single documents, but sets of documents which together support a better solution to the task problem.

When it becomes necessary to compare several different systems, each of which has an "operating curve" as shown in the figure, there are many issues to be decided. Among them are: micro averaging versus macro-averaging; counting of cases (which has a tractable statistical basis) versus summative averaging (which doesn't), as well as the issues about weighting mentioned above. These are troublesome, but they are manageable in simple spreadsheet type layouts, which permit various users of the evaluation to select among schemes for comparison. An interesting wrinkle is to consider the relation among the three lines shown in the graph. The lowest represents the performance of a "mindless" system. The highest represents an ideal system. The middle one represents the performance of this system. The ratio of the length of the short arrow to the long one is a clean statement of how far the system carries us from mindlessness to excellence. In the spirit of Data Envelopment Analysis, one might change the slope of the lines to make each system look as good as possible. Some systems will thus emerge as "non-dominated" or "Pareto Optimal" while others simply are not best in any situation.

Distributed Systems

The curve of effect versus effort can be defined, and perhaps more clearly, for addressing the effectiveness of distributed systems. In this case there is a natural "gold standard" which is the fully integrated system. So the "effect" axis can be measured as "fraction of the integrated system performance which is achieved". The effort axis is still complicated, as it must combine the cost, processor requirements, bandwidth, management, stability, etc. etc. of the distributed scheme. An example of this kind of measure can be applied to the distributed indexing schemes discussed by Dolan et al at UCSB.

Position Statement Approaching D-Lib Metrics

Ed Fox

---

1. Scope of Problem:

As argued by several, our group faces an enormous challenge in dealing with a wide-open opportunity. Thus, in these early days of DL work, when definitions, prototypes, and commercial systems are just being developed, we hope to help to develop methodologies, and launch a series of scientific studies according to those methodologies, that will help advance the state of the art.

2. Broad Approach:

If we take the union of the definitions advanced for "digital library" we find that we are confronting some of the most complex and advanced information systems proposed. This suggests that we conclude:

"D-Lib Metrics WG's broad agenda should be to encourage all studies that help the DL field, providing guidance in any aspect of a digital library. Such studies may deal with individual aspects of a digital library (like searching, or the interface) and draw on methodologies developed in particular research communities (e.g., IR, HCI) and reported in existing conferences (e.g., TREC, CHI). However, it is encouraged that studies targetted to help Digital Libraries consider at least two such specialty areas, whenever possible, to help with the integration that is at the heart of the DL field."

3. New Problems

More at the core of interest of the D-Lib Metrics WG should be issues that have emerged because of DL systems and not before. Some of these have been mentioned in other position statements: interoperability, scalability, and heterogeneity. Others are concepts that arose in other fields but take on new dimensions in the DL area, such as quality, timeliness, usability, integration, reliability, relevance, and specificity.

As we look at properties and metrics, capabilities and tasks, services and scenarios, situations and environments, benchmarks and hypotheses, failure modes and comparisons, we must revisit "metrics" in our title. What can we measure? How accurately? With what generality (statistical, inferential)?

Since the field is wide open, let us try the most obvious studies first:

* Compare existing digital libraries used in real-life contexts, and apply a mix of metrics drawn from areas like IR, HCI, Hypertext, distributed processing, etc., with the aim of better serving users.

* Use performance measurement approaches, including modeling and simulation, to analyze internal and inter-system communication (especially through published protocols) of distributed digital libraries, with the aim of improving operations.

* Deeply think about, and work toward concensus, on understanding the essence of digital libraries, so the most important metrics and studies can be proposed. These may extend the current set of Challenge Problems and Scenarios. One starting point is the architectural work at CNRI. Clearly, there also is relevance to work on documents, digital objects, collections, repositories, terms and conditions, shopping models, and federated systems.

4. Related Work

In an advanced graduate class on digital libraries (http://ei.cs.vt.edu/~cs6604) in Computer Science at Virginia Tech, Fall 1997, a number of students worked on projects relevant to our WG activities. One group developed a visual simulation of NCSTRL, to allow testing of the effects of varying topologies, number of backup servers, number of users, amount of network traffic, server speed, and other parameters (like timeout interval).

Another group did a classic HCI study comparing 4 digital libraries. They had 4 comparable tasks for each system, and had 48 users work with all 4 tasks on all 4 systems (balancing on order). We are still mining the results on timing, failures, pre and post-surveys, synchronized videotape records of both the user and the screen, and critical incident comments of observers.

A third effort focussed on developing a theoretical framework, based on sets, to describe DLs.

In addition to these efforts, a number of the students are working with the W3C effort for HTTP-NG to understand through log analysis how the WWW operates, and to consider user behavior and proxy/server performance.

D-Lib Working Group on
Digital Library Metrics

D-Lib Working Group on Digital Library Metrics: Working Area

(Private).

This page contains pointers to various working documents of the Working Group.

Working Group Members

The Working Group is open to anyone interested in working on the issues as described in the charter. Current membership is shown here.

Last Meeting

Meeting/Teleconference 29 October 1998

- Minutes(Draft), Working Group on Digital Library Metrics, Open Meeting, October 29, 1998, Kahuku, HI
- Slides from the meeting

Prior Meetings

Data Gathering

To aid in the gathering of data for candidate metrics and the challenge problems (see working paper below), a web-based approach has been developed. See http://www.dlib.org/metrics/private/forms/challenge_problems.html for the forms being used to gather and index the challenge problems and http://www.dlib.org/metrics//private/forms/metrics_in_use.html for those being used to gather and index submitted metrics relevant to digital libraries.

Working Papers

It is intended that once the papers are sufficiently matured, they will be moved to the open working documents area.

- Leiner, Barry. Two Simple Scenarios for Distributed Digital Libraries [Draft Prepared by Barry M. Leiner for the D-Lib Working Group on Digital Library Metrics, May 2, 1999].

  In this paper, we identify two simple scenarios and develop some

specific but simple metrics for those scenarios. These scenarios are each intended to capture a common aspect of the distributed digital library. Furthermore, they are framed explicitly to permit development of standardized testing in the context of existing testbeds in a test suite, e.g., the D-Lib Test Suite.

- Leiner, Barry. Abstracted Metrics for Distributed Digital Libraries [Draft Prepared by Barry M. Leiner for the DLib Working Group on Digtial Library Metrics, October 15, 1998].

  The next step in the process of developing a consensus set of simple metrics is to abstract candidate metrics from those specifically proposed for the challenge problems. These candidate metrics are then to be examined by the group and community as a basis for the consensus metrics. This document lists abstracted metrics based on the Challenge Problems gathered to date. For each metric, the challenge problems that cited that metric (or a metric that got aggregated to the shown one) is listed. The metrics are sorted into two major categories: those that try to measure the effort required by a user in doing a task, and those that measure the efficacy of the results.

- Leiner, Barry. Types of Digital Library Metrics [Draft Prepared by Barry M. Leiner for the DLib Working Group on Digital Library Metrics, January 26, 1998].

  At the kickoff meeting of the DLib Working Group on Digital Library Metrics (WG), held January 7-8, 1998, at Stanford University, some discussion was held as to the various kinds of metrics that might be applicable to the digital library. In preparation for our discussion on the specific metrics, we felt it useful to lay out the various dimensions of such metrics. This document is intended for that purpose, and this draft is intended to start the discussions.

- Leiner, Barry. Digital Library Challenge Problems and Metrics [Draft Prepared by Barry M. Leiner for the DLib Working Group on Digital Library Metrics, September 30, 1998].

  At the kickoff meeting of the DLib Working Group on Digital Library Metrics (WG), held January 7-8, 1998, at Stanford University, as well as in prior email discussion, the WG concluded that an critical initial step to developing metrics for the digital library is to identify a set of challenge problems/scenarios. For each of these challenge problems, we would then identify appropriate metrics. This could then be the basis for future discussion to identify commonality in the metrics and select a suitably small set of metrics that adequately cover the domain of interest. This document is intended to start that process. The intent is that this document be a major work product of the WG.

- Nakassis, Tassos, Metrics for Digital Libraries, May 7, 1999.

  > "What follows is an attempt to recast the metrics issue in such a way that a good part of our definitions of metrics will be easily translatable in measurable quantities and measurable procedures. While there are no implicit or explicit innovation claims in the following pages, I believe that they provide an explicit framework that will make some of our work easier."

- Wesley, Rebecca and Lagoze, Carl, Traditional Library Measures for Digital Libraries, August 26, 1998.

  > Libraries have a long tradition of measuring all aspects of library service. Anyone attempting to create a comprehensive bibliography of metrics in libraries will have a difficult time indeed. Library literature is full of studies both formal and informal (many Master's theses) and many articles and books summarize the research.
  >
  > In order to get one's mind around metrics that might be used to measure digital libraries, the scope must be narrowed. This article will indicate the traditional library metrics that can be useful for digital libraries.

Email

- Send email to dlib-metrics@cnri.reston.va.us.
- Hyperarchive of mail to D-Lib Metrics Working Group.

prepared by Barry Leiner
last modified 5/7/99 bl/bw

Metrics for Digital Libraries

Tassos Nakassis
National Institute of Standards and Technology
May 7, 1999

1. **Introduction**

   What follows is an attempt to recast the metrics issue in such a way that a good part of our definitions of metrics will be easily translatable in measurable quantities and measurable procedures. While there are no implicit or explicit innovation claims in the following pages, I believe that they provide an explicit framework that will make some of our work easier.

2. **What do metrics measure?**

   Metrics provide clues as to future behavior. I.e., if we knew that the observed precision for a number of representative queries varied between 0.60 and 0.70, we would expect similar performance in the future.

   The logic of this statement can be summarized as follows:

   - The behavior of a library is a random vector B
   - The distribution of B depends on parameters x[1], x[2], ..., x[n] to be estimated
   - Metrics are estimates of such parameters. obtained through well defined procedures.

3. **How are parameters measured?**

   No big magic, parameters are measured through observations. Some measurable things (e.g., phenomena related to distant galaxies, earthquakes, and such) require that we observe the phenomenon if and when nature produces it. But in many instances (and digital libraries are such an instance) we can measure through experimentation.

4. **Metrics and Digital Libraries**

   To come back to digital libraries, I would summarize the situation as follows:

   - Metrics relate to measurable phenomena
   - Such phenomena can be measured directly or indirectly by experiments
   - Standard statistical techniques exist and can be used to summarize and classify the measurements (e.g., do we have one or multiple populations when the observed recall measurements cluster around two values such as 0.4 and 0.8?)

5. **Experiments for digital libraries.**

   The Metrics group has defined a number of desirable behaviors and metrics without explicitly stating how we would measure them. For some (e.g., recall) there is an implicit measurement procedure that unfortunately relies on comparing an observation against an idealized behavior which, most likely, we do not know. For others we have a desired behavior but no associated metrics. E.g., we would like a query to retrieve related documents (there is an example in which a researcher asks for economic information on a country A and is directed toward documents treating related regional issues). For such metrics it is important to recast the definition and to

derive an experiment Examples:

- Recall:
    1. Definition: The probability p that a record R matching query Q will be retrieved;
    2. Experiment: Start with an (R,Q) pair; insert R; submit Q; observe if R matches Q
    3. Measurement: Count the percentage of experiments in which the inserted record is succesfully retrieved
- Retrieval of related non-matching information
    1. Definition: The probability p that a record R semantically related to query Q is referred to in the answer to the query.
    2. Experiment: Start with an appropriate (R,Q) pair; insert R; submit Q; observe if R is referred to in the answer.
    3. Measurement: Count the percentage of experiments in which the desired behavior was observed.

6. **Related issues**

   We have up to now conveniently bypassed several thorny issues the most important of which is that of the querying language. It is not all that obvious how to formulate queries and it is even less obvious how to formulate queries that a library will accept.

   Example: An image indexing scheme might index images through texture; A query for "images containing mountain lakes" could be translated as "Images containing rocks, wooded areas, and water". The library may answer accurately and completely the query submitted but the answer will miss mountain lakes that do not have trees in their proximity and will probably retrieve a few islands, maybe, even fjords. The library will have done a perfect job in terms of the queries it accepts, possibly a less stellar one from the point of view of the user who wishes to see images with mountain lakes. Even worse: Assume that the local library is federated with other libraries that support more sophisticated schemes. If the mapping (mountain lakes)-->(forest, water, rocks) took place outside the library, the original query will never be seen by the other libraries and the quality of the answer will be degraded. In short:

   - We need to consider behavior as observed at the user level, not at the library level, if we perceive that the mapping of queries results in information loss
   - We need to explore the relation between "Queries expressed in a natural language", "queries in a general, formal language", and queries that are acceptable at the library interface.

7. **A program of work:**
   - We need to create some toy libraries that duplicate the look and feel of real libraries but whose purpose is to be experimented upon;
   - We need to populate these libraries with synthetic multimedia data that are doubly indexed:
       - Index 1: The indexing scheme the library uses
       - Index 2: A description of the process through which the specimen was created.
   - The behavior of the database can be assessed by mapping the queries on index 1 (to mimic library behavior) and on index 2 to accurately retrieve what we query for.
   - We need to link these libraries in ways that reflect their federated properties.

- Once these tools are in place we can develop experiments to be conducted and reference materials to be used

8. **Implementation**

   It would be desirable to have access to schemes that mimic reality as accurately as possible. My group at NIST cannot do this by itself. Therefore, unless better alternatives materialize, we plan to implement the following:

   - A doubly indexed reference database based on whatever querying and indexing techniques are available;
   - A set of experiments along the lines of what was delineated above;
   - A semi-automated scheme for carrying out the experiments referred to above.

Two Simple Scenarios for Distributed Digital Libraries

Draft prepared by
Dr. Barry M. Leiner
for the
D-Lib Working Group on Digital Library Metrics
2 May 1999

Introduction

The goal of the DLib Working Group on Digital Library Metrics (WG) is

to develop a consensus on an appropriate set of metrics to evaluate and compare the effectiveness of digital libraries and component technologies in a distributed environment.

A first step in this process has been to compile a set of scenarios, or challenge problems, intended to capture the richness and diversity of the digital library environment. For each of these challenge problems, appropriate metrics were identified. This process of scenario compilation continues, and the current set of scenario descriptions may be seen either through a document *Digital Library Challenge Problems and Metrics* <http://www.dlib.org/metrics/private/papers/Challenges.html> or through a set of web forms <http://www.dlib.org/metrics/private/forms/challenge/c_index.html>

In this paper, we identify two simple scenarios and develop some specific but simple metrics for those scenarios. These scenarios are each intended to capture a common aspect of the distributed digital library. Furthermore, they are framed explicitly to permit development of standardized testing in the context of existing testbeds in a test suite, e.g., the D-Lib Test Suite (http://www.dlib.org/test-suite/index.html)

Two Simple Scenarios

Two simple scenarios have been suggested as starting points. The first is simply the ability to retrieve for a user a specified dissemination of an object. The scenario is there are a number of repositories, with perhaps different access protocols/mechanisms etc. A user wants to receive a dissemination of one of the objects in one of the repositories. The system under test would be responsible for dealing with the different repositories and associated protocols, etc., and retrieving the desired dissemination.

The second scenario is a simple query across multiple repositories. A set of desired objects within the various repositories is defined with some description. The scenario is that the user wants to identify the objects in the combined testbeds that are within the desired set.

Simple Metrics for Simple Scenarios

For each of these scenarios, we could define a sizable number of relevant metrics. To start with, though, we define a few simple, easy to understand metrics that can be applied easily across different instantiations of the scenarios.

Metrics for Retrieval Scenario

The following are candidate simple metrics that could be used for the retrieval scenario:

- Proportion of Supported Disseminations. A priori, the set of disseminations are identified for a particular request. The request for the disseminations is made, and the number of successful retrievals measured.
- Set of Supported Formats. As the user attempts to retrieve disseminations in a variety of formats, we track which formats get retrieved successfully.
- Set of Supported Protocols. Similarly, as the user attempts to retrieve disseminations from various repositories, we track which repository access protocols are used to successfully or unsuccessfully retrieve disseminations.
- Delay in Retrieval. The statistics of the time involved in retrieving a dissemination are tracked. This could involve mean, standard deviation, minimum/maximum, or other statistical measures of the time. The time measured could be from a user perspective (time from when the dissemination is first requested until it is actually obtained and presented to the user) or system perspective (time from when the request is issued to the remote system until when it is returned.)
- User Effort. User interface metrics could be applied to measure how difficult it is for the user to prepare the request and then deal with the result.

Metrics for Simple Query Scenario

The following are candidate simple metrics that could be used for the simple query over multiple repositories scenario:

- Precision. The proportion of identified items that lie within the pre-defined set is measured.
- Recall. The proportion of the pre-defined set that lies within the identified item set is measured.
- Time Delay. The statistics of the time required to obtain the query result is measured.
- User Effort. User interface metrics are applied to measure the effort required to prepare the query, any special knowledge required to manage the query, and effort required to assimilate the query results.

Applying the Metrics

To validate the metrics and demonstrate their utility, we identify a specific subset of the D-Lib Test Suite (http://www.dlib.org/test-suite/index.html). DeLIver (http://www.dlib.org/test-suite/uiuc.html) and NCSTRL (http://www.ncstrl.org) deal with overlapping domains, and similar types of material in the sense they both contain heavily text-based reports on computer science research. DeLIver has them in the form of papers in technical journals, and NCSTRL has more informal documents. It would be desirable if user-issued queries (e.g., I need papers on brain implanted computers that support rapid eye movement measurements), provide results about work in both forms (tech journals and gray material). We would like to post a query and get an integrated response from both.

However, the formats of the material are different (SGML with a little PDF for DeLIver, PDF and other text-based forms for NCSTRL), the query formats are most definitely different, etc.

This then is an ideal example of the kind of test being discussed. For example, a test could be set up so that the system under test would be a user interface supporting query across multiple servers.

The combination of DeLIver and NCSTRL are an interesting combination of libraries in the context of federated libraries. Each themselves is a federation of sorts. NCSTRL is explicitly a federation with defined protocols for linking them. DeLIver provides an integrated way of accessing a variety of repositories of journals. Trying to work across the two (and perhaps others as we move forward) will help us understand the issues in such loose federations.

Abstracted Metrics for Distributed Digital Libraries

Draft prepared by
Dr. Barry M. Leiner
for the
D-Lib Working Group on Digital Library Metrics
15 October 1998

Introduction

The goal of the DLib Working Group on Digital Library Metrics (WG) is

to develop a consensus on an appropriate set of metrics to evaluate and compare the effectiveness of digital libraries and component technologies in a distributed environment.

A first step in this process has been to compile a set of scenarios, or challenge problems, intended to capture the richness and diversity of the digital library environment. For each of these challenge problems, appropriate metrics were identified. This process of scenario compilation continues, and the current set of scenario descriptions may be seen either through a document *Digital Library Challenge Problems and Metrics* <http://www.dlib.org/metrics/private/papers/Challenges.html> or through a set of web forms <http://www.dlib.org/metrics/private/forms/challenge/c_index.html>

The next step in the process is to abstract candidate metrics from those specifically proposed for the challenge problems. These **candidate** metrics are then to be examined by the group and community as a basis for the consensus metrics.

The following are abstracted metrics based on the Challenge Problems gathered to date. For each metric, the challenge problems that cited that metric (or a metric that got aggregated to the shown one) is listed. The metrics are sorted into two major categories: those that try to measure the effort required by a user in doing a task, and those that measure the efficacy of the results.

It should be noted that, while these metrics are relatively simple to understand, many of them are only meaningful in the context of a scenario. This implies that the way these metrics are likely to be used is as a standard basis for defining specific metrics for specific scenarios. These scenarios, then, with their associated metrics would be the basis for comparing different approaches or quantifying the evolutionary progress of a system or approach.

Effort

Ease of Use

- Doctor wants to know if there is any new information in an online update on item Z.
- User wants to build and maintain an SDI profile on topic X

Required level of expertise

- I wish to publish to a specific audience
- Disseminate software to a known audience

Understandability and Utility of Results

- User wants to build and maintain an SDI profile on topic X
- High School Student who is interested in topic X wants to know if there are contrary opinions.
- Query across media boundaries.

How much annoyance was created

- Who else is interested in what I just wrote?

Frequency of usage

- Who else is interested in what I just wrote?
- I need email address of person X at institution Z.
- Measuring the usage of downloaded software

Effect

Accuracy of information

- I need email address of person X at institution Z.
- Dependence on results from unknown sources.
- Who is reading my articles?
- How credible is this information?
- How reliable is this software?

Precision

- Query across media boundaries.
- Dependence on unknown results.
- I wish to publish to a specific audience
- Disseminate software to a known audience
- Artifact assembly

Recall

- Query across media boundaries.
- Dependence on unknown results.
- I wish to publish to a specific audience
- Disseminate software to a known audience
- Artifact assembly

Completeness and availability of information

- Who is reading my articles?
- How credible is this information?
- Library Director wants to know how many electronic articles publisher X made available last year for Y dollars paid.

System response time and timeliness

- Dependence on results from unknown sources.
- Dependence on unknown results.
- I wish to publish to a specific audience
- Disseminate software to a known audience
- Distributed software dependencies
- Provisioning of computing resources rather than downloading software.

Types of Digital Library Metrics

Draft Prepared by Barry M. Leiner
for the DLib Working Group on Digital Library Metrics
January 26, 1998

Preface

At the kickoff meeting of the DLib Working Group on Digital Library Metrics (WG), held January 7-8, 1998, at Stanford University, some discussion was held as to the various kinds of metrics that might be applicable to the digital library. In preparation for our discussion on the specific metrics, we felt it useful to lay out the various dimensions of such metrics. This document is intended for that purpose, and this draft is intended to start the discussions.

Introduction and Background

The Digital Library encompasses a broad variety of types of content, users, and usage paradigms. Furthermore, the parties involved in the digital library have various concerns, ranging from user interaction through to efficiency and effectiveness of the operation of the library. The range of possible metrics for the Digital Library is equally broad.

At the kickoff meeting of the WG (held January 7-8, 1998 at Stanford University), we identified a number of dimensions to the metrics for the Digital Library. This document attempts to lay out these dimensions and therefore the scope of the metrics we will deal with.

System-wide vs. individual services

The Digital Library is viewed as a collection of services providing a capability for users to deal with a broad range of information. Some metrics will deal with system-wide capabilities while others will try to measure the performance of individual services.

User interaction vs. underlying system

The goal of the digital library is to support the users. Like in any system, though, it is often useful to measure underlying system aspects as well as direct user interaction/support effects.

Effort vs. effect

Measurements may be made of the effort required of a user to perform a task (where effort is interpreted broadly to include items such as patience in waiting for a response). Measurements can also be made of the effect of those efforts coupled with the system performance, measuring the resulting performance in

user terms. Often the most useful measurements of a system couple the two, generating a curve relating effort and effect.

Snapshot vs. session (temporal granularity, adaptability)

Some measurements are of the system at a given time with little or no "memory" in the metric. For example, one can measure the precision and recall of a particular query, or one can measure the processing delay for an action. Often, though, the system response depends on prior context and what is of interest is the system's (including the user) ability to adapt and learn as a task proceeds.

Capability vs. utility

The digital library represents a certain set of capabilities - the aggregate set of its services. These capabilities can be measured and metrics associated with them. Ultimately, though, the interest is in the utility that the system brings to the users - how useful is the digital library in assisting users accomplish their tasks? Metrics can be developed for both aspects.

Single "user" vs. scalability

From a singer "user" perspective, what is important is the performance that user sees. Many metrics are appropriate to that performance. From the perspective of the operator of the system, it is important that the individual performance be maintained over a large set of users at reasonable cost. Thus, metrics addressing the scalability of the system are critical.

Collection/content vs. System/capability/utility

The digital library provides a set of services to help users deal with content. Thus, metrics are appropriate for the system and its associated services and also the parameters of the content of the collection (e.g., how complete is it in the domain of interest to the user?).

Conclusion

As we have seen, there are a large number of aspects and perspectives associated with metrics for a digital library. The challenge for our Working Group will be to identify a small enough and simple enough set of metrics that will be useful for the community to assess the fruits of its labor.

Back to Working Group Private Page

Digital Library Challenge Problems and Metrics

DLib Working Group on Digital Library Metrics
September 30, 1998

Preface

At the kickoff meeting of the DLib Working Group on Digital Library Metrics (WG), held January 7-8, 1998, at Stanford University, as well as in prior email discussion, the WG concluded that an critical initial step to developing metrics for the digital library is to identify a set of challenge problems/scenarios. For each of these challenge problems, we would then identify appropriate metrics. This could then be the basis for future discussion to identify commonality in the metrics and select a suitably small set of metrics that adequately cover the domain of interest. This document is intended to start that process. The intent is that this document be a major work product of the WG.

Introduction and Background

The Digital Library encompasses a broad variety of types of content, users, and usage paradigms. Furthermore, the parties involved in the digital library have various concerns, ranging from user interaction through to efficiency and effectiveness of the operation of the library. The range of possible metrics for the Digital Library is equally broad.

The goal of the DLib Working Group on Digital Library Metrics (WG) is

> *to develop a consensus on an appropriate set of metrics to evaluate and compare the effectiveness of digital libraries and component technologies in a distributed environment.*

To accomplish this goal, the WG is compiling a set of scenarios, or challenge problems, intended to capture the richness and diversity of the digital library environment. For each of these challenge problems, appropriate metrics will be identified. The intent is that this compilation be used as the basis for abstracting a small set of simple, easily understood metrics that can be used across the diversity of the digital library.

This document is meant to address the first part of the process - the compilation of a set of challenge problems and associated metrics.

We have organized the challenge problems into several categories as follows:

> Discovery - the user is attempting to find and retrieve information with specified characteristics
>
> Dissemination - the user is attempting to make material available through the digital library

Other Usage - the user is attempting to use the digital library to accomplish some task requiring access and/or manipulation of information in the library

Library Administration - scenarios typifying the interactions of the library administrator/operator/librarian with the digital library

System Operation - scenarios typifying the interactions of system developers and support staff with the digital library

Other - scenarios not covered by the above or difficult to categorize

These challenge problems have been input by various members of the Working Group, and are available at http://www.dlib.org/metrics/private/forms/challenge_problems.html. Working Group members and others are encouraged to submit challenge problems through the form available on the Web, and to comment on this document or the challenge problems through email to the Working Group mailing list, DLib-metrics@cnri.reston.va.us

Discovery Challenge Problems

In these challenge problems, the user is attempting to find and retrieve information with specified characteristics.

Doctor wants to know if there is any new information in an online update on item Z.

Problem Description:

Let us assume that the online update is to a medical reference book. In this case the user does not want to be notified every time an update comes out on the book, she only wants to see the information updated since she last logged on. Furthermore, she only wants updates on item Z, not updates on all of the book information.

Issues:

- It is clear that the above update service would have to track the user by some identification and remember what the user had seen in the past. This raises privacy issues.
- This also raises problems of granularity. If the medical reference book were on anatomy and item Z were treatment for cancer of the esophagus, would the update on cancer be too much for the doctor to wade through.

Metrics:

- Functionality
- ease of use
- granularity of information

User wants to build and maintain an SDI profile on topic X

Problem Description:

SDI is Selective Dissemination of Information, also called updates. SDI's have been around since the 1970s particularly as part of commercial services. The basic notion is that each person creates a profile which includes particular databases and a search strategy. This profile is loaded onto the server and when a new database update arrives (sometimes monthly, could be weekly or even daily) the user profiles are run against the new information. If there are search results they are sent electronically to the user.

Most users have no idea that these services are available nor do they know how to develop and/or maintain their profiles. A user needs to not only create a profile but also needs to know how to change it or delete it.

Issues:

- How do users find out what SDI services are available to them (including cost)?
- How do users create and test a profile? Profiles need to be tested to make sure they are narrow or broad enough (before the user is sent 10,000 records).
- How do users find out how to change or delete the profile? What do users do if they have lost their id and/or password?

Metrics:

- Individual service: Measure ease of use of profile system.
- Measure utility of results from SDI profile and ease changing profile to obtain higher quality results.

High School Student who is interested in topic X wants to know if there are contrary opinions.

Problem Description:

On most search engines today users query with keywords which pull up all kinds of information on the topic. There is no easy way to limit the search results to those that are pro and/or con to a specific issue (topic X). Or if there is a way to limit results to pro/con users do not know about this feature.

Issues:

- How do users find out what functionality is available on a system?
- Most search engines have features users don't know about or complexity they don't care to use. Many services use bibliographic data with special fields that would allow users to limit their results to be more useful but users are unaware of the functionality or feel it is too difficult to learn.
- Consistent semantics from one system to the next.

Metrics:

- functionality: ease of use
- functionality: user understanding

Who else is interested in what I just wrote?

Problem Description:

The potential of the web is to bring authors and readers together spontaneously without invading privacy. Services already do this (sort of) via links and home pages, and authors can include their email address and/or url in their documents. New ways of improving on this rather informal method of bringing authors and readers together WILL emerge.

Issues:

- We do not know how well the current system works.
- We do not understand the privacy issues.
- We do not know how to test the extent of invasion of privacy.

Metrics:

- Measure how many times authors encourage users to contact them by inserting email, phone, url into their documents.
- For authors who encourage contact, how much annoyance was created.
- Measure if over time author has quit putting contact information in the documents.

I need email address of person X at institution Z.

Problem Description:

Most academic institutions have online directories that identify phone and/or email addresses of faculty, staff, and students. These directories are usually found on the www using http://www.institution.edu. Other institutions like corporations have been slower to make this information available to the general public. There is no standard way to find email addresses of people, be they children, CEOs, Congressmen, or Plumbers.

Issues:

- Email has been most effective in enhancing communication around the world. The downside of email is two-fold: 1. Too much too fast; 2. Spamming and advertising.
- In today's hectic world often the best way to communicate with someone is via email, if only the email address was known. Most people give up if they cannot locate the email address within five minutes.

Metrics:

- Individual Service: How many times did users of the service need an email contact address which was conveniently provided?
- Individual Service: For existing email directories, are they accurate and up-to-date?

Query across media boundaries.

Problem Description:

The problem here is an information query that crosses different systems, media, etc. For example, a user trying to understand the intellectual property issues associated with the case between GM and VW would have to trace the property from GM to VW. Therefore, there is a need to correlate the information across different languages and publication media. A similar problem might have to deal with different types of media such as voice, text, and images.

Issues:

- Consistent semantics across multiple media and languages.
- Integration and presentation across multiple media and languages.

Metrics:

- Precision and recall of query across multiple languages (based on expert assessment of relevance of documents.
- Understandability by user of relevance of document to query.

Dependence on results from unknown sources.

Problem Description

The problem here is an information query that requires a result but the user may not know from whom. An example here is the problem of analyzing weather trends. You might need the temperature in Troy, NY, but not know what sources there might be for that temperature. Furthermore, you probably wouldn't care what the source was as long as you had confidence in its reliability and accuracy.

Issues

- Characterizing information suitably to retrieve without knowing the source
- Assessing suitability of sources and balancing system response/performance

Metrics

- Accuracy of result
- System response time

Dependence on unknown results.

Problem Description:

The problem here is an information query that in turn depends on another set of correlated data, but the dependence is "data dependent". For example, an analyst may be trying to understand trends in Far East financial markets. Another analyst may have turned up a correlation between Tokyo and Hong Kong market indicators. If you knew that such a correlation existed, you might query for it. But the system needs to identify and retrieve such indirectly related data.

Issues:

- What kinds of analytical tools need to be built into the infrastructure to support analysis-based queries?
- Semantic interoperability

Metrics:

- Precision and recall of such complex queries
- System response.

Dissemination Challenge Problems

In these challenge problems, the user is attempting to make material available through the digital library.

I wish to publish to a specific audience

Problem Description:

Given an open (to any search by anybody) digital library and an author publishing a digital object (e.g., book, image of painting, program), how does one support the concept of notifying a specified audience?

Issues:

- How does one characterize the target audience?
- How does digital library know the characteristics of searcher - registration, deduction at search time, questions at search time?
- Do we actively notify audience members (e.g., e-mail) or passively (e.g., whenever member searches notice of new digital object is given)?

Metrics:

- functionality: target audience characterization, notification
- system: required level of expertise (both publish and search); timeliness of notifications
- user(publisher): % of potential users(readers) identified; % of potential users reached; % of

potential users' replies(that goes to user friendliness of notification)

Who is reading my articles?

Problem Description:

Most authors want to know about their readers. They want to know at least how many individuals downloaded their document. Many would like to know more about these individuals. Information services must weigh the author's desire to know with the readers desire for privacy.

For decades authors using citation indexes have been able to find out how many authors have cited their works. Many authors assume if their work is cited it was also read. These citation indexes have given authors a pretty good idea of the specific individuals reading their works.

Issues:

- In the electronic environment it will be much easier to give specific readership information to authors, easier to know, easier to collate.
- Should authors be able to get data on which web pages have links to their documents? Is this equivalent to citation information?
- Should authors be given data on which individual machines or users have downloaded their documents?

Metrics:

- What specific information is collected from a service?
- What specific information is available to the public?

Disseminate software to a known audience

Problem description:

When releasing a new software package, how does one notify all users who might find it useful? (related problem: How can users be notified of updates to software they have already downloaded and are currently using?)

Issues:

- How does one notify not just other numerical analysts, but also application scientists in the broader user community who might not recognize the athematical terms in which the software is described or know that the **a**lgorithms are pertinent to their applications?
- Do we actively notify users or wait until they search the repository?

Metrics:

- functionality: target audience characterization, notification

- system: required level of expertise (both publish and search); timeliness of notifications
- user(publisher): % of potential users(readers) identified; % of potential users reached; % of potential users' replies(that goes to user friendliness of notification)

Other Usage Challenge Problems

In these challenge problems, the user is attempting to use the digital library to accomplish some task requiring access and/or manipulation of information in the library.

How credible is this information?

Problem Description:

Most people agree that provenance is a critical part of information. The value of the information is to a certain extent based on knowledge and guarantees of the source, and trust in the source. So there are three parts to this question:

1. What is the source?;
2. Do I trust the source?
3. How accurate is the information?

Issues:

- It is often difficult to know and guarantee the information source of electronic information. In the print world the author and publisher were prominently displayed on the cover and title page. Although book and journal pirates exist, the discerning print reader could usually tell that the material was a fake. Not so in the electronic environment.
- There is a whole spectrum of information accuracy; from fact to opinion. Each reader (or listener) takes in information and shades it with his own beliefs about the credibility of the information. For this reason most authors highlight their education and experience which lends credibility to their words.

Metrics:

- Measure the availability of the author’s name and institution/company; also any additional information about the author.
- Measure the system’s ability to guarantee machine or institution source.

How reliable is this software?

Problem description:

(related to "How credible is this information?") The user wants to know how reliable is a piece of software that they are downloading from a software library.

Issues:

- Is the software what it purports to be, or has it been modified accidentally or maliciously (e.g., introduction of a virus or Trojan horse)?
- Has the software been evaluated and tested? If so, how can testing and evaluation results be provided to the users in a useful form?

Metrics:

- percentage of software with verifible digital signatures, strength of encryption scheme upon which signatures are based
- software quality metrics (e.g., level achieved in level-based scheme)

Artifact assembly

Problem description:

This problem is quite different from typical queries. The problem is the assembly of related information across the system. It is easiest described as a game. There are a large number of artifacts distributed randomly across the Internet. A subset of these artifacts have a defined relationship. The challenge problem is to find and retrieve the subset and assemble them into a desired result.

Issues:

- Information characterization to allow definition of the relationship

Metrics:

- Precision and recall of retrieved artifacts

Distributed software dependencies

Problem description:

Software packages often require that other software packages also be installed and linked to. It is often difficult for the user to locate the correct version of these related packages. (This problem is related to the artifact assembly problem)

Issues:

- How are relationships between resources, as well as their locations, cataloged and made available so that the correct required related resources can be located? Can the process of locating and downloading related resources be automated?
- How can availability of necessary related resources be assured?
- Under what conditions can substitutions be allowed if the specified resources are unavailable?

Metrics:

- Time to correctly assemble and install complete composite package

Library Administration Challenge Problems

These scenarios typify interactions of a library administrator/operator/librarian with the digital library.

Library Director wants to know how many electronic articles publisher X made available last year for Y dollars paid.

Problem Description:

There are many questions Library Directors, Curators, Bibliographers, Funders, and Auditors might ask in the electronic environment that were not possible to get in the print environment. The above question is just one of many regarding quantity of information obtained for costs expended.

Issues:

- How do we create an open statistical environment for electronic information?
- Many publishers of electronic information do not want purchasers to know the details of availability and cost for fear that the buyer will find the cost too high.
- Once we succeed in getting publisher’s to recognize the need for this data, how do we get publishers to agree on a standard format for consumers to use.

Metrics:

- for every shopping model (subscription, site license, per session rates, etc.) publishers will need to provide information to the consumer on how much "data" (articles, books, data sets, etc) were available and at what price.
- additionally publishers may want to provide this information to the general public in order to advertise their product.

Measuring the usage of downloaded software

Problem Description:

What software that has been downloaded from the repository is actually being used and for what purposes?

Issues:

- With http access, caches, etc., recipients of downloaded software are frequently not identified.
- There may be many more users at a site than just the person who downloaded the software.
- Surveys have low rate of return.

- Does requiring registration so as to be able to contact users discourage downloads?

Metrics:

- download counts
- usage statistics (difficult to obtain)

System Operation Challenge Problems

These scenarios typify interactions of system developers and support staff with the digital library.

Provisioning of computing resources rather than downloading software.

Problem description:

In the case of a resource (e.g., a computer program) that is large and/or difficult to download and install and/or that will be used only once or a few times, can we provide a service that enables the user to use the resource where it is as an alternative to downloading it to his own machine? (This problem is pertinent to other types of digital libraries -- e.g., data archives that want to support processing of data at the archive site)

Issues:

- Who pays for the computational cycles and how are they accounted for (preferably not by having the user get an account on the repository server, since this would take much longer than downloading and installing the resource)
- What are the security implications of allowing the user to access computational resources on the repository site, including possibly uploading and running his or her own program or function on the server (e.g., a user-defined function to be executed by a repository optimization routine)

Metrics:

- total turnaround time from locating a resource to being finished using it

Other Challenge Problems

These scenarios are not covered by the above or are difficult to categorize.

Unclassified Challenge Problems

(editor's note: The following are challenge problems identified during the kickoff meeting held at Stanford on 7-8 January 1998, but have not yet been elaborated according to the structure used in the document: problem statement, description, issues, and metrics. We identify where each group of

problems came from so that the original source can update and elaborate. It is preferred that this elaboration be done through the web form at http://www.dlib.org/metrics/private/forms/challenge_problems.html)

(Group I Challenge Problems)

I'd like to identify the characteristics of "too long queries"

- Queries that have too long a response time for users
- Numbers of words, numbers of clicks, media; places
- Class of entities to deal with
- Elicit, process, reports

I'd like to know all the tricks to access The Library

- Review self knowledge, learn user profile
- Elicit, processing, dispense

I'd like to blackmail the premier of Netonia

- Elicit, process, dispense

Which parts of my federated library cause the (worst/most) problems?

- Wrong content
- Down
- Slow/timeouts
- Metadata but no full document
- Numerous spelling errors
- Non-standard multimedia formats
- Low quality of service on streams
- Multiple identical copies (not "de-duped")
- Poor metadata - short records, no authoritycontrol

Who has plagiarized one of my publications?

- All of it
- A part or section
- A table or figure
- A reference (with a deliberate typo)

Which works are most cited?

- Pairs of works?
- Works in a particular topical area?
- Authors of methodological works in Nobel-prize areas?
- Paired with rebuttals/disclaimers?

In what fraction of cases has the system (permitted/denied) access in (clear?) violation of the stated terms and conditions?

(Group II Challenge Problems)

The "Control Zone" problem.

A librarian wants to include items in their collection but not, as a result, include all items transitively linked from those items.

The "Refind" problem.

A person has found something in a library a long time ago (e.g., 15 years) and wants to be able to find it again. This is a "fuzzy known item" problem.

The "loose ends" problem.

A person, for example a software engineer, is working on a large project, for example a mult-part software project. This person needs to keep track of the open tasks in the project and then, when resuming the project, find the open tasks and what needs to be done with them.

Selection "cost/benefit".

In the traditional library world collection developers have numerous shortcuts such as automatically buying whole series from certain publishers. The challenge is to develop shortcuts in the digital world where objects may be available from a wider variety of sources.

Legal constraints

I want to add items to my collection and want to know if there are any legal encumbrances or if any one will be offended by any of the items I want to add.

(Group IV Challenge Problems)

Change in Document State

I want to create a change in state of a collection of documents involving the modification, annotation, and creation of documents. Further, this change in state must pass a separate validation before it is made public.

Idea distribution/remuneration

I have a great idea that I want to distribute to an interested class of consumers (targetted, specific) and receive fair remuneration.

I want to publish a document and track all derived works.

I want to publish a document that ensures the best possible presentation for a given user:

- Associated software
- Track alternative representations
- Restrict acess
- Index, abstracts

Publication Workflow

I want to create and issue an edition of a journal subject to complex workflow issues (minimize time to market)

Context Representation

I want to represent the implicit context that is associated with a document when the value of cataloging a document is small.

Edited by Barry Leiner
Last modified 9/30/98

Traditional Library Measures for Digital Libraries

Libraries have a long tradition of measuring all aspects of library service. Anyone attempting to create a comprehensive bibliography of metrics in libraries will have a difficult time indeed. Library literature is full of studies both formal and informal (many Master's theses) and many articles and books summarize the research.

In order to get one's mind around metrics that might be used to measure digital libraries, the scope must be narrowed. This article will indicate the traditional library metrics that can be useful for digital libraries.

Here are the categories traditional libraries measure:

1. measurement of staff performance
2. measurement of integrated library systems
3. measurement of a collection (both quality and quantity)
4. measurement of use of a collection
5. measurement of user satisfaction
6. measurement of service (both quality and quantity)

Staff performance may be indirectly useful for measuring digital libraries but the assessment elements will be entirely different. Perhaps when digital libraries are more fully established, measurement of staff performance may be relevant.

Measurement of integrated library systems is very specific to traditional library functions like circulation and serial check in. For more detail on specific functions of integrated library systems click here.

The other four categories of traditional library measurement are relevant to digital libraries.

Measurement of a collection (both quality and quantity).

A book by Baker and Lancaster, "The Measurement and Evaluation of Library Services" gives six basic evaluation approaches on the collection itself: (Baker: p41)

1. Outside evaluators can conduct a subjective collection review and give their impression of collection adequacy.
2. Holdings can be checked against lists of best books or standard bibliographies in a subject area
3. Holdings can be checked against lists of sources cited by researchers in a discipline
4. The absolute size of a collection and its growth rate can be compared against quantitative standards issued by the profession or other formulas that state the optimal size of a collection to meet patron needs
5. The size of various subject collections in learning institutions can be determined and compared against the emphasis placed on these materials in the curriculum
6. Collection depth, or comprehensiveness, can be estimated.

Add to this list, citation analysis, which can be done at large research libraries to analyze one subject area at a time.

Quantitative measures of traditional libraries usually start with the number of volumes per student/faculty/staff.

Measurement of use of a collection

In the past few decades computers have aided libraries in evaluating the use of a collection. Using automated circulation systems use data can be collected on material that leaves the library and also material that is used within the library. Many libraries will ask patrons to leave the library books and journals on tables or return shelves so that the barcodes can be scanned to gather what is called in-house use.

Library use statistics are heavily used to select material for

- weeding
- remote storage
- cancellation

Library use statistics also indicate if funding for staff and materials is adequate.

Use statistics are gathered on the basis of a physical piece and usually cummulated to figures like number of items circulated vs number of times circulated.

Measurement of User Satisfaction

Traditional Library

User satisfaction can be very difficult to evaluate. The more specific the user's need the easier it is to evaluate. Known item searches are the easiest. The user is looking for a specific book, article, videotape, or whatever, either the library has the item or it does not.

Patron satisfaction issues to consider are the quality and quantity of service and the cost.

Baker and Lancaster give three levels and types of evaluation:

1. Effectiveness must be measured in terms of how well a service satisfies the demands placed on it by its users.
2. Cost-effectiveness is concerned with its internal operating efficiency"
3. Cost-benefit evaluation is concerned with whether the value (worth) of the service is more or less than the cost of providing it.

User surveys are often used to obtain data on user satisfaction. Thousands of user surveys have been done. "The vast literature of 'user studies' defies effective summarization." (Baker, p.369).

For more information on user surveys see the Hernon or Baker books listed below.

Digital Library

Many user studies have been done to determine usability of search services.

One example is, QUIS, Questionnaire for User Interface Satisfaction was developed at the University of Maryland to measure human-computer interaction.

Measurement of service (both quality and quantity)

Services in traditional libraries generally include reference, document delivery (interlibrary loan), online search services, technical service, and collection building.

Two examples of services are document delivery and reference (Cronin):

Document Delivery: Possible measurements

- number of requests from library users
- number of requests sent
- number of requests received
- % of requests successful
- time between date of request and date sent
- time between date sent and date received
- average turnaround time

Reference: Possible measurements

- Number of questions answered (quantitative)
- Number of questions answered correctly (qualitative)

Types of measurement

- self-reporting
- peer observation
- supervisor observation
- unobtrusive "planted"questions
- objective-testing
- user feedback

Conclusion

Four areas of measurment of traditional libraries may be useful for digital libraries:

1. measurement of a collection (both quality and quantity)
2. measurement of use of a collection
3. measurement of user satisfaction
4. measurement of service (both quality and quantity)

Although many studies have been done of patron use of computers, librarians haven't gone beyond Information Retrieval (IR) measurements. Very little has been published on system performance. Digital library collections are not yet well established in traditional libraries, so little has been done on use of a

digital collection or measurement of a digital collection.

Baker, Sharon L. and F. Wilfrid Lancaster, "The Measurement and Evaluation of Library Services" second edition, Information Resources Press, 1991.

Blagden, John and John Harrington, "How Good is Your Library?: A Review of Approaches to the Evaluation of Library and Information Services" Aslib, The Association for Information Management, 1990.

Buckland, Michael K., "Concepts of Library Goodness" Canadian Library Journal, April 1982, 63-66.

Cooper, C. Crys, "A Qualitative Study of Novice Users of FIRSTSEARCH on the IBM" ERIC No.: ED360980, May 1993.

Cronin, Mary J, "Performance Measurement for Public Services in Academic and Research Libraries." Occasional Paper Number #9 February 1985, Office of Management Studies.

Hernon, Peter and Charles R. McClure, "Evaluation and Library Decision Making" Ablex Publishing, 1990, page 146.

Kantor, Paul B., "Objective Performance Measures for Academic and Research Libraries" Association of Research Libraries, 1984.

Orr, R. H., "Measuring the Goodness of Library Services: A General Framework for Considering Quantitative Measures" Journal of Documentation 29, no.3 (Sept 1973), 315-32.

D-Lib Working Group on Digital Library Metrics
Open Meeting
October 29, 1998
Kahuku, HI

Minutes (Draft)

Attendees:

Name		*Organization*	*E-Mail*	*Telecon Participant*
Paul	Kantor	Rutgers	Kantor@scils.rutgers.edu	
Larry	Lannom	CNRI	LLannom@cnri.reston.va.us	
Barry	Leiner	CNRI	BLeiner@cnri.reston.va.us	
Rik	Littlefield	Battelle	Rik.Littlefield@pnl.gov	
Ed	Fox	UVA	Fox@vt.edu	X
Ramash	Krishnamurthy	OSU	Krishnar@ucs.orst.edu	X
Bill	Pottenger	UIUC	BillP@cs.uiuc.edu	X

Agenda

0830	Welcome and Introductions
0840	Discussion of Abstracted Metrics
1030	Adjourn

Meeting Summary:

A small and brief meeting (2 hours) of the Metrics WG was held during a breakout session at the DARPA IC&V/IM PI meeting. Four people participated in person, and three by teleconference.

After a brief introductory discussion, the primary topic discussed was the initial draft of the abstracted metrics, based on the challenge problems submitted to date. A brief discussion took place on the relevance of the initial

metrics to the issue of distributed digital libraries. In particular, the question raised was whether unique metrics were needed to deal with the fact that the digital library is distributed. We concluded that from the perspective of a single user of the digital library, the fact that the library is distributed is relatively irrelevant. However, scenarios where there are multiple users of the digital library should probably capture some essential characteristics of use in those contexts, such as the consistency of information or the impact of having multiple users on the performance as perceived by any user.

The bulk of the discussion was on the metrics abstractions themselves. Leiner noted that even metrics such as precision and recall have an assumed context or generic scenario, and presented a slide with the scenario for precision and recall. Paul Kantor proposed the term "frames" to capture the context associated with specific metrics. We used that notion to explore some example scenarios and metrics, and concluded that the notion would be helpful. The idea is that metrics assume a particular operational context, and that such operational contexts may have associated with them multiple metrics. Specific scenarios, or challenge problems, would then be associated with particular relevant frames, which would then indicate the relevant metrics for those scenarios. The frames would act as "filters" to determine the which metrics were appropriate for the particular scenario.

We concluded that it would be useful to continue in this direction of discussion, trying to define some relevant frames and associated metrics, and then test them against specific scenarios.

Powerpoint slides summarizing this discussion were used during the meeting, and are posted for downloading.

Back to Working Group Private Page

Prepared by Barry Leiner
Last modified 11/6/98

D-Lib Working Group on Digital Library Metrics: Prior Meetings

(Private).

This page contains information on prior meetings of the Working Group.

Meetings 24-27 June 1998

The Working Group sponsored three meetings during DL98, 24-27 June:

- An open meeting of the Working Group on 24 June
  - Minutes
  - Slides used during meeting
  - Unedited notes by Rebecca Wesley
- A Birds of a Feather session on 26 June
  - Minutes
  - Unedited notes by Rebecca Wesley
- A Workshop on Digital Library Metrics organized by William M. Pottenger, UIUC/NCSA and Robert McGrath, UIUC/CANIS

Working Group Kickoff Meeting

The kickoff meeting for the Working Group was held on 7-8 January 1998 at Stanford University

- Meeting Minutes (DRAFT)
- Informal notes of kickoff meeting prepared by Bill Arms
- Notes of kickoff meeting prepared by Bea Oshika
- Slides used during the meeting
- Position Papers submitted in preparation for the meeting

prepared by *Barry Leiner*
last modified 11/9/98