Evaluating Digital Libraries: A Longitudinal and Multifaceted View

GARY MARCHIONINI

ABSTRACT

THE PERSEUS DIGITAL LIBRARY (PDL) IS ONE OF THE primary digital resources for the humanities. Under continuous development since 1987, the project has included an ongoing evaluation component that aims to understand the effects of access to digitized source materials in the humanities. A summary of the PDL genesis and current status is given and the multifaceted and longitudinal evaluation effort is described. A brief synthesis of results is provided and reflections on the evaluation along with recommendations for DL evaluation are given.

INTRODUCTION

Digital libraries marry the missions, techniques, and cultures of physical libraries with the capabilities and cultures of computing and telecommunications. Evaluating digital libraries is a bit like judging how successful is a marriage. Much depends on how successful the partners are as individuals as well as the emergent conditions made possible by the union. All three entities—the two individuals and the gestalt union—are of course influenced by their context as well. The difficulties arise from the complexity of mutually self-adapting systems interacting in a rich environment. Metrics for success for component parts of a complex system may be distinct from the metrics for success of the marriage (e.g., success for an individual partner is typically necessary but not sufficient to ensure success for the marriage).

Digital libraries (DLs) are extensions and augmentations of physical

Gary Marchionini, School of Information and Library Science, CB #3360, 203 Manning Hall, University of North Carolina, Chapel Hill, NC 27599-3360
LIBRARY TRENDS, Vol. 49, No. 2, Fall 2000, pp. 304-333

libraries (Marchionini & Fox, 1999). As extensions, we might evaluate the individual partners using existing techniques and metrics. Assessing the impacts of libraries on the lives of patrons and the larger social milieu are the ultimate goals of evaluation, but the practical difficulties of assessing such complex and varied impacts cause us to measure the effectiveness and efficiencies of library operations and services as surrogates for these impacts. Metrics such as circulation, collection size and growth rate, patron visits, reference questions answered, patron satisfaction, and financial stability may be used to assess physical library performance in this regard. Clearly, these are metrics that may be points of departure for evaluating digital libraries, but they are not sufficient to characterize the new rapidly emerging entity. Evaluation criteria for digital technologies can also be useful. For example, metrics such as response time, storage capacity, transfer rate, user satisfaction, and cost per operation may be useful in assessing technological components but may not be sufficient to characterize DL performance, let alone impact. As extensions of physical libraries and digital technologies, these metrics are good starting points, but we must look further to consider the effects of DLs as augmentations that provide new services, products, and capabilities.

In assessing new services and products, it is difficult to distinguish novelty effects (both positive and negative) from long-term effects. More importantly, new services and products typically create new effects that cannot be predicted until an "installed base" of practice takes root. Additionally, some of these unanticipated effects are due to the new services and products, and some are due to the marriage of existing services and products to the new ones. It seems certain that assessing these effects will not happen in "Internet time." The effects of DLs will emerge over time as physical libraries, DLs, and people mutually adapt and mature; the problem of evaluation for DLs is thus one of assessing complex adaptive systems.

The goal of this discussion is to provide a view of an important DL that has been evolving for more than a decade. Over this time, both the Perseus Digital Library (PDL) (Crane, n.d.-a) and the related evaluation effort have evolved, guided by central missions to provide and understand the effects of broad access to digitized source materials in the humanities. From the beginning of the evaluation effort in 1987, the primary evaluation aim was to address the impact of this project on users and the humanities community. This article will provide a reflective summary of results for this particular DL, discuss the methodological approaches taken to understand the evolving DL, and argue for multifaceted and longitudinal assessments of DLs in general. The article first describes the genesis, evolution, and current status of the Perseus Digital Library; provides a perspective on evaluation as a research and problem solving endeavor; summarizes how this perspective was applied to the evaluation of the

Perseus DL over a twelve-year-period and what outcomes have emerged; and finally provides some reflections and recommendations for DL evaluation.

The Perseus Digital Library: Access to Primary Resources

As stated on its Web page: "The Perseus Project is an evolving digital library of resources for the study of the ancient world and beyond." The mission statement reads: "Our primary goal is to bring a wide range of source materials to as large an audience as possible" (Crane, n.d.-b). These themes of evolution and wide-scale access to source materials have been constants from the earliest days of the project. A small team of classicists led by Gregory Crane began planning in 1985, and several small grants supported a number of prototypes that led to a large grant from the Annenberg/CPB Project to begin building the "hypertext" in 1989. One component of the plan was an external evaluation effort that has continued until the present. The initial plan was to digitize as many ancient Greek texts and English translations as possible; gather or create images, maps, and video objects related to locations and artifacts; and build tools for searching and manipulating these materials (see Crane, 1988, and Crane & Mylonas, 1988, for early articulations of the Perseus vision). The Apple HyperCard platform was selected since it offered the best hyperlinking and multimedia capabilities in the late 1980s. One objective was to create a CD-ROM package that contained many of the primary readings and resources that students taking classics courses would need and to make this package available for the cost of one or two textbooks. In addition to university students, the Perseus team expected that classical scholars would find the corpus and tools helpful to their research and would also contribute new translations, interpretations, and tools. Because the funding aimed to apply new technologies to improve learning and teaching, the project was characterized as an interactive curriculum.

Elli Mylonas conducted a series of interviews with twenty professors in a variety of humanities fields at twelve institutions in 1987 to discuss how Perseus might be used for instruction (Mylonas, 1987). The results suggested that Perseus could be both a reference and a source of primary materials. Although professors were skeptical about using Perseus for their own research, they saw possibilities for extending and refreshing their own knowledge, especially in small departments where they teach courses outside their main research expertise. They raised issues of using a new technology, navigation in the corpus, and overall quality of the information resources (e.g., accuracy of texts, keeping information up to date, choice of Loeb editions of texts). Most importantly, they raised issues about the relationships between Perseus (digital realm) and the rest of the scholarly universe and whether students would limit themselves to Perseus materials and points of view. This issue of defining an information space and shaping user perspective through what is (and is not) included and how users may work through the materials became an important pedagogical issue for the project. As part of a planning grant before the project began, a workshop, "Assignments in Hypertext," was held at Harvard University in March 1988. At that meeting, the Perseus team clarified their aim to focus on primary source material that scholars and students might use to create their own interpretations rather than instructional materials that explicated meaning didactically. This constructivist philosophy rejected the use of secondary readings and authoritative rigid "paths" through the database[1] and promoted the notion of primary materials as raw materials for student exploration and investigation. To this end, the primary drivers for Perseus became acquiring as much primary source material as possible and developing navigational and analysis tools (e.g., search, hyperlink structures, and morphological summaries).

The first prototypes consisted of HyperCard stacks that presented Greek texts and English translations, graphical site plans for temples and other environments, and basic search and display tools. The staff recognized that rapid changes in technology could render their efforts obsolete unless they chose robust data models. One crucial decision was to use the Standard Generalized Markup Language (SGML) to code texts as they were keyed into digital form. This was an expensive decision at the time because the HyperCard platform did not handle SGML markup and the value added by domain experts marking up structure and semantics in the texts was not usable in the short term. In hindsight, this was a prescient decision as systems more than a decade later are able to take advantage of this coding. Another decision was to use an object-oriented approach to managing the multimedia data. A "catalog card" was developed for each physical object (e.g., vase, coin, architectural object, sculpture, site). This card serves as an entry point or reference for all specific files or screens related to that object. For example, an important vase might have more than 100 image files associated (shots at various details around, inside, and on the bottom) with it but one main entry that includes information on the following: Collection, Summary (text), Ware, Shape, Painter, Potter, Context, Region, Date, Period, Dimensions, Primary Citation, Decoration, Graffiti, Inscriptions, Parallels, Collection History, Condition, Shape Description, Sources Used, Keywords, and Views (links to the actual image files). Catalog templates for the other objects are likewise defined (see the art and archaeology collections at http://www.perseus.tufts.edu/art&arch.html). The Perseus team was thus creating specialized metadata schemes for different objects in the collection long before bibliographic management became known as metadata. These crucial data management decisions were informed by a technical advisory board and an educational advisory board, each made up of internationally prominent researchers who met annually to react to decisions and give advice.

The collection development plan called for amassing a significant portion of the extant Greek text and a large number of images and maps. Selection of Greek textual materials and translations was driven by logic (useful in undergraduate courses) and opportunity (availability). For texts, in addition to Greek text itself, several other types of information were included: apparatus criticuses (commentaries and explanations added to texts), scholia (annotations made by people over time on the manuscripts, sometimes becoming part of the text itself), metrical analysis (for poetry), staging notes (for plays), bibliographies, and English translations. The translations were particularly controversial since there were many translations for the important works, and intellectual property decisions led to using the Loeb translations at Harvard where the project was then based (in a few cases, new translations were commissioned). In addition to these basic files, existing indexes and a lemmatized word index were included. Because the Greek language is highly inflected, finding the lemma (root forms) for a word is crucial to reading and translation; as a side effect, the complexity of the morphology makes word search in Greek potentially more powerful than in English since there is more information packed into the morphology. Crane developed the Morpheus tool (a morphological parser) for this purpose, and it was used to create the word indexes and is an important component of the Perseus text analysis tools in the current version. The team identified texts in ten genres (epic poetry; Elegiac, Iambic and Lyric poetry; tragedy; comedy; historians; orators; mythology; philosophy; inscriptions/papyri; and other). More than ninety primary texts were identified for inclusion in these genres. Greek texts were keyed offshore and then subjected to extensive editorial processing where proofreading, additional notes, and markup were done. In the current PDL, there are almost 300 Greek and Roman texts in Greek, Latin, or English along with eighteen secondary texts and nineteen Renaissance texts.

An important goal of Perseus is to bring text and other media together to add value to scholarship and learning. The original plan called for both purchasing slides from museums and a large-scale original photography effort. To guide collection development, a set of thirteen topics related to art, architecture, and archaeology (AAA) were identified: house, propylaea, stadia, stoa, temples and sanctuaries, invention and refinement of architectural idiom, theaters, topography, town planning, artists and artisans, Greek athletics, daily life, and stylistic development in Greek art. The artifacts that carry the meaning characterized by these topics include: architecture, vase painting, architectural sculpture, other relief sculpture, and freestanding sculpture (coins were added later). The representations for these artifacts are in the form of slides and drawings that were digitized. Over the years, in addition to rights to existing slides, thousands of original slides were taken at museums and sites in the United States and Europe. Original slides were copied and archived remotely, and the acquisition and preservation experience is documented for others to use.[2] Today there are more than 33,000 images available through the image browser, representing more than 500 coins, 1,500 vases, 1,400 sculptures, 180 sites, and 380 buildings—each object having a catalog card entry point.[3]

In addition to the art, architechture, and archeology visual media, there were three meta collections created that cut across the texts and AAA—an encyclopedia, a narrative overview, and an atlas. The encyclopedia is accessible via hyperlink or word search from any view of the PDL. The overview is a substantial essay (an electronic book) by Thomas Martin that introduces the ancient Greek world and includes hyperlinks to items in the DL. The current PDL includes a number of additional secondary treatments (e.g., vase painting, Greek and Latin syntax). The atlas has gone through many changes as it moved from the CD-ROM version that included LandSat imagery and maps for pre-determined regions to the WWW version that is built upon a full geospatial database. The current WWW atlas provides access to more than 1,000 physical places in the ancient world at multiple levels of resolution, ranging from a global view that allows a user to label bodies of water, populated places, and modern borders, to a zoomed in resolution that allows the user to display contour lines, spot elevations, and rivers (Chavez, 2000).

The Perseus project, from its earliest days, was situated in an academic region (Cambridge, Massachusetts) that supplied a wealth of technical and content talent. The project team was led by a philologist who articulated the mission and assembled an interdisciplinary humanities team that included people with specializations in ancient history, archaeology, art history, and Greek and Latin language and, over the years, drew graduate student assistants from many departments in the Boston-area universities. Importantly, the Perseus team shared a belief that information technology is a powerful medium for advancing the study and appreciation of the fruits of humanistic thought and facilitating new levels of expression by students and scholars alike. This point of view was somewhat radical in the mid 1980s, and the original team members who were in untenured faculty positions or about to become assistant professors in classics departments openly discussed the dangers of working to change how classics is practiced and taught. In the short term, these dangers were instantiated, as all three of the tenure-track central team members were not offered tenure in their first faculty position. In the long term, each of these three are tenured faculty leading their departments and the field in leveraging technology to advance classics and the humanities in general. There are two lessons here. First, the importance of leadership, tenacity, and commitment and a ready talent pool all contributed to the persistence and evolution of the DL. Second, it is important to assess impact over a long period of time.

The ambitious plan to create Perseus emerged over a three-year period (1985-88) and was supported first by small equipment and planning grants before the four-year grant from the Annenberg CPB/Project began in 1989. The main evaluation plan was developed as part of the proposal for the four-year cycle. Over the 1989-1993 period, the system was developed as a set of HyperCard stacks with a variety of database backend supports on Unix and Macintosh platforms. A CD-ROM version (Perseus 1.0) was produced and published by Yale University Press (now out of print). A second Macintosh version (Perseus 2.0) has been available for several years, and a platform-independent CD-ROM version is now available. Funding to extend the system to the platform-independent version, to add materials related to the history of science, to add materials related to ancient Rome, and to create and evaluate instructional models was obtained from a variety of sources, including the National Science Foundation, the National Endowment for the Humanities, and the Fund for Improvement of Post-Secondary Education (FIPSE). These grants extended the hypertext corpus and tools to a more diverse library and made the project an ideal candidate for the Digital Library Initiative Phase Two program, which provides support for the 1999-2004 period. Evaluation has been included in all of these efforts at various levels of support to ensure persistent and longitudinal assessment feedback to the project team. Securing a steady stream of funding cannot be overestimated when examining the overall impact of the Perseus DL.

Over the past twelve years, the corpus migrated from a HyperCard driven CD-ROM to the World Wide Web while adding new materials and tools. The central mission of providing access to large amounts of source materials has been carried out by the project director (Greg Crane) and many original Perseus team members and they have continued to guide the emerging DL. Although in the early days the project did not refer to itself as a library, the library metaphor was explicitly captured in the catalog card metadata records and in providing cheap and easy access to large volumes of primary source materials. Today the Perseus DL includes more than 225 gigabytes of texts, images, maps, and indexes and garners 300,000 http requests per day mainly at the Tufts site but also at European mirror sites at Oxford and Berlin. Commercial encyclopedias, as well as hundreds of syllabi at universities and K-12 institutions around the globe, link to it. A spin-off organization, Stoa (www.stoa.org), has been created to support research and electronic publication for humanities scholars, and Tufts university has begun to support the Perseus DL as part of its overall infrastructure. The DL funding in coming years promises to extend the scope of materials and tools greatly. In the following sections, the evolution of the evaluation effort and key results are detailed.

Evaluation as a Research Activity

Evaluation has many connotations ranging from highly focused and well-defined product testing to the highest form of cognitive reflection.[4] Classical program evaluation aims to identify causal models that link well-specified variables to dependent outcomes. Suchman (1967) indicates the difficulties in actually executing such evaluations in social science settings but gives guidelines for systematically collecting and using quantitative methods for large-scale program evaluation. Inspired by anthropological research, a range of qualitative methods for evaluations have been developed that do not pose hypotheses or presuppose causal models. Williams (1986) provides a set of readings that support such qualitative approaches to evaluation. Many theorists propose combining methods through triangulation. Cook and Reichardt (1979) offer a collection of papers that describe qualitative and quantitative approaches to evaluation with an eye toward synthesis, including Campbell's (1979) recommendations for systematic case studies. Rossman and Wilson (1985) provide an example where data from multiple methods are synthesized. In contrast, Bednarz (1985) provides a theoretical overview of the different paradigms and argues that, although it is effective to synthesize multiple data within a paradigm, synthesizing across paradigms ultimately fails as one approach inevitably dominates. Clearly, evaluation research continues to be an active area of methodological research in its own right.

It is important to distinguish evaluation as a research process from evaluation in the product testing and system efficiency sense. Many specific measures applied to product testing may very well be used as evidence in evaluation research. However, evaluation research considers the interactions of complex phenomena—including people—and reaches conclusions through chains of inferences supported by data rather than direct measurement. As noted earlier, the evaluation literature bristles with debates over basic approaches to evaluation, especially with respect to qualitative versus quantitative methods and rationalistic versus hermeneutic philosophies. Collecting multiple data sets and triangulating the results is advocated in most paradigms, and the PDL evaluation takes this approach by systematically collecting data using statistical techniques for summarizing data where appropriate but not using inferential statistics to test pre-conceived hypotheses. Rather, triangulation is used to make inferences and develop arguments about PDL meaning and impact. This approach is based on the belief that *evaluation* is a research process that aims to understand the meaning of some phenomenon situated in a context and the changes that take place as the phenomenon and the context interact. This definition implies that evaluation specify what is the research process (metrics and procedures), what is the phenomenon (its mission and salient characteristics), and the context(s) in which the phenomenon occurs. Of course, when developing evaluation plans and carrying out the work, the primary emphasis is on the research process because the phenomenon and context are so omnipresent to instigators of the evaluation. This last point is the reason outside evaluators are often used, although this implies more expense since the outsiders must "come up to speed" on the phenomenon and context at hand. The Perseus experience suggests that this very process of learning about the phenomenon and the context was itself an important part of the overall evaluation.

Evaluation has both theoretical and practical impact in information science. Theoretical constructs, such as information needs, relevance, and information transfer, are debated and assessed regularly, and metrics for assessing system development and operation are crucial to continued progress in practice. In the case of information retrieval, evaluation is often focused on the effectiveness of a result set in a specific search, or aggregations of results across many searches, to assess and compare different search systems (see Harter & Hert, 1997, for a recent review of IR evaluation; see Voorhees & Harmon, 2000, for an overview of the recent Text Retrieval Conference [TREC] results). Metrics such as recall and precision are typically used, and a standard set of procedures that includes test questions and pooled relevance judgments are used to ensure comparability across systems. Usability testing is another type of evaluation that focuses on the effects obtained when individuals apply an information processing system to accomplish tasks (Nielsen, 1993). Usability testing adopts metrics such as time to completion, accuracy, satisfaction, and errors. A variety of procedures for situating the tasks (laboratory/field setting, assigned/open tasks, and so on) are used in usability tests. Another branch of information science, bibliometrics, aims to assess the impact of individuals or communities (e.g., journals) on research progress through citations and other bibliographic relationships (see White & McCain, 1989, for a review). Citation and co-citation counts (including hypertext links in the WWW) serve as the basic metrics upon which new indicators, such as impact value, are derived. Evaluations are also conducted to determine how effective libraries are in carrying out their missions. Griffiths and King (1991) use a model for evaluating information centers that includes measures in four classes: input cost, outputs (quantities, quality, timeliness, availability, and accessibility), effectiveness (e.g., amount of use, user perceptions of services, user satisfaction), and domain (e.g., patrons, staff, information need types, user behaviors). Saxton (1997) provides a meta-analysis of reference service effectiveness that considers nine variables (expenditures, total collection size, reference collection size, collection size per patron, volumes added per year, volumes discarded, overall change in collection size, proportion of change to total size, and number of hours open). System engineering evaluations judge the effectiveness and efficiencies of hardware and software using such metrics as access and transfer latencies, mean time to failure, and development and maintenance costs. The PDL evaluation depends mainly on educational evaluation but draws heavily upon each of these information science subfields for metrics and techniques.

The Perseus Evaluation Plan and Evolution

As part of the proposal to the Annenberg/CPB Project in 1988, an evaluation plan was outlined. The plan was detailed during the first six months of the project and served as a guide for activities throughout the four years of funding and beyond to subsequent funding cycles (Marchionini, Neuman, & Morrell, 1989). The plan presented a multifaceted approach to evaluation as a research process that included multiple metrics and methodologies and aimed to understand Perseus as a new electronic phenomenon with impact in multiple contexts. Two contextual factors were strongly influential in slanting the evaluation effort toward educational contexts. First, the funding was aimed at educational applications of electronic materials—the project was titled "An Interactive Curriculum on the Ancient Greek World." Second, the evaluation team's background and experience were both rooted in education and instructional technology. Steve Ehrmann, the Annenberg/CPB Project program director, asked us to consider the definitional question "What is Perseus?" and, over the years, we posed various explications of the Perseus phenomenon. Crane, in an interview in April 1989, noted that Perseus was "a laboratory" to study heterogeneous information tied together to focus on one subject. To Neuman and me (both faculty in an information science program), it seemed from the start that Perseus was a library that extended the possibilities for self-directed learning.

In the introductory paragraph, the evaluation plan was characterized as a roadmap that would guide decision-making over the years rather than a detailed blueprint specifying all details of the evaluation. This was so for three reasons: technology changes rapidly; variables and metrics related to educational and scholarly processes are complex and difficult to quantify; and we aimed to assess the interactive nature of learning, teaching, and scholarly production. This last point is an important one because assessing interactivity had few precedents at the time and remains a significant challenge today as we struggle to assess browsing and other interactive behaviors in the WWW environment.

The architecture of the evaluation plan was characterized by crossing goals with objects of evaluation to define a set of research questions and then mapping a variety of data collection and analysis methods onto these questions. Three high level goals this DL offered to learners and scholars were access (to large volumes of multiple media source material), freedom (self-directed access and use), and collaboration (among learners and teachers). Four classes of evaluation objects were defined: learners, teachers, the technical system, and the content. These generated a hierarchical set of ninety-four research questions. The learner questions had four main categories[5]: specific tactics and strategies with six specific subquestions; overall patterns of use with five specific subquestions; changes of behavior and perception with ten subquestions, one of which also had two subsubquestions; and Perseus use compared to other approaches with three subquestions. Three instructor questions[6] had nine associated subquestions; three system questions[7] had thirty-eight subquestions and subsubquestions; and three content questions[8] had twelve subquestions associated with them. Methods deemed appropriate and applicable were then associated with each of these questions. The basic methodology aimed to collect data using a variety of methods and then triangulating results to answer the research questions. Four classes of data collection methods were defined: observations, interviews, document analysis, and learning analysis.

Five kinds of observations were identified. *Baseline observations* were made to situate the evaluators and build relationships with individuals involved in the observations. These were semi-structured where we sat unobtrusively in classrooms or labs and made notes during lectures, discussions, and lab sessions. *Structured observations* were defined to follow a specific protocol in a classroom or lab—e.g., systematically observe the behavior and record notes for a purposive sample of individuals. For example, select five students and alternate observations every three minutes to record whether they were taking notes, looking at the instructor, and so on. Although we conducted a few such observations, this technique was used less often than we expected due to the difficulty in collecting such fine grained data in a classroom environment or laboratory with so many other pertinent activities underway. *Participant observations* involved the evaluator with students and were audiotaped. The observer is guided by a semi-structured protocol and may ask or answer questions (participate) according to the situation. This technique was used heavily in one site where a graduate assistant worked intensively with a class over an entire semester (Evans, 1993). *Think-aloud observations* aim to determine what cognitive activity underlies behavior and are used widely in psychology and education research (Ericsson & Simon, 1984). Subjects were asked to think aloud while they worked on various tasks, and the entire session was audiotaped. Both participant observation and think aloud could have been included in the interview category, but we classified them as observations because they are less dependent on self-report on the part of the subject and more focused on the observed activity in which the subject is engaged. The final observation method is *automatic screen journaling* (transaction log analysis). This technique automatically captures user actions such as keystrokes or mouse clicks, adds time stamps to the record, and may include "snapshots" of the screen at critical junctures of the interaction. Routines for capturing user actions were developed and used extensively in one site over a semester (Evans, 1993). Evans provided graphical displays for sessions that demonstrated both systematic and opportunistic study strategies. Patterns such as clear demarcations between persistent use of texts and images contrast with regular and/or random alternations between media. We have also used transaction analysis to study usage in WWW sites such as the Bureau of Labor Statistics (Hert & Marchionini, 1997). Transaction log analyses are used with the Perseus WWW logs to determine gross interaction patterns today (e.g., number of requests for different resources, temporal patterns of access, and so on).

Interviews were initially defined around course schedules with introductory, midsemester, and exit interviews planned. Although we did conduct interviews at different intervals for some classes, a better way to categorize the interviews we eventually used is verbal interviews with individuals or groups and written or online questionnaires. We often conducted interviews with individual instructors and students. These interviews were guided by semi-structured protocols and were typically audio taped. Such interviews were conducted at ten different universities, in some cases over several years, and have proven to be one of the most valuable evidence sources for our findings and reports over the years. Group interviews were also conducted at several universities. These are likewise guided by general questions and audio taped (one session was videotaped). Most of these sessions were with groups of students and yielded candid commentary on how instructors use Perseus and what students thought about the DL. A written questionnaire was developed for use in classrooms and, over the years, almost 1,000 students at several universities completed the questionnaire (the questionnaire was modified several times over the years to reflect changes in the technology and content). The questionnaire has three main sets of questions: demographics, including computing and Perseus usage; system features; and impact on learning. Check lists, Likert scales, and a few open-ended questions are included. A somewhat surprising result that recurred over the years was the lack of correlation between demographics and learning impact and a positive correlation between system interface features and learning impact. Perceptions about ease of use are closely related to perceived learning effects.

In recent years, as attention has shifted to Perseus impact on teaching and research, a number of verbal protocols and e-mail questionnaires have been used with instructors. In 1999, a short online questionnaire eliciting general information about Perseus use via the WWW was used to collect data. After a pilot test with voluntary responses, the questionnaire was automatically given to every tenth unique visitor to the Perseus home page (IP addresses were recorded so that no single address received the questionnaire a second time). There were 20,701 responses. Of those completing the questionnaire, 7 percent were professors, 24 percent were undergraduates, 11 percent were graduate students, 16 percent were K-12 students, 5 percent were continuing education students, 4 percent were K-12 teachers, and 25 percent placed themselves in the "other" category.

Clearly, the PDL is finding substantial usage in educational settings as well as by the public outside the classroom. The majority (66 percent) of respondents said that they were using the PDL for the first time, and 11 percent said they used the PDL once a week or more often. The PDL (and other DLs) must consider serving first-time or casual users for the foreseeable future. Of those responding, 54 percent said they were using the PDL from home, 16 percent at school, 14 percent at the office, 3 percent from a library, 3 percent from other places, and 9 percent did not respond. The significant home access data has interface design and system performance implications for digital librarians since home infrastructure support will tend to lag behind institutional infrastructure (e.g., bandwidth, latest client software). It is interesting to note that 37 percent of the respondents said they were using the PDL for personal interest, followed by 23 percent for research, 21 percent for homework, 9 percent for class work, and 9 percent had no answer. In many ways, the PDL is used like a public library in this regard. Of those responding, 36 percent reported that they learned about the PDL from a search engine and 27 percent followed a link from another site. These huge numbers of visits based on searches or links distinguish the PDL from physical libraries. Another 13 percent learned about the PDL from a teacher, 7 percent from a friend, 3 percent from a publication, 6 percent from other sources, and 9 percent provided no answer. These data represent a large sample of DL users and bear reflection as digital librarians develop and upgrade their systems (see Marchionini, Scaife, & Crane, 2000, for the details of this survey).

Document analysis uses the content of a variety of objects to understand goals, outcomes, and processes. In another venue, we used document analysis extensively in assessing user needs for the Library of Congress Digital Library Program (Marchionini, Plaisant, & Komlodi, in press). In the Perseus evaluation, we first focused attention on products produced by the project such as documentation and project reports. Later, with funding from FIPSE, we focused on instructional documents such as syllabi, assignments, and instructional materials such as structured paths through the materials. Other documents analyzed include: the number and range of research papers and conference presentations, link patterns (e.g., WWW links [citations] to the Perseus home page), electronic list messages (number of messages, who participates, and content categories), and original materials contributed by scholars (the Stoa).

Learning analysis was the fourth general class of methodology defined in the evaluation plan and included baseline data such as student reading and translating rates, assignments and syllabi characteristics, and estimates of student performance. Grades and other instructor assessments and structured self-assessments were also proposed. In hindsight, we were able to obtain far less of this type of data than we hoped. Access to student grades is problematic and gathering representative samples of skill levels was logistically impossible. Bill McGrath reported at the Perseus evaluation meeting in 1996 on his systematic analysis of student course evaluations for his classes over several years while Perseus was introduced and used. In this case, the evaluation averages went down. Student interview data we collected with McGrath's classes indicated that students very much enjoyed the lively lectures and discussions, were attracted to his courses for this reason, and were somewhat resentful of time taken away from class discussions as electronic resources were used to augment the lectures. Although student course evaluations are attractive potential sources of evaluation data, the many factors that go into final evaluations and the logistical problems of obtaining access make them less useful in practice. It is important that instructors be prepared to accept negative effects (course evaluations being only one indicator) as well as positive effects when they make significant changes to their courses and teaching styles.

Based on the evaluation plan, we conducted site visits to ten universities from 1989 to 1993 and produced reports for each site visit at the end of each year. In one case, we were able to conduct a controlled comparison of four sections of a large class in which two sections used Perseus and two did not. Other data-collection activities were undertaken opportunistically. For example, Perseus was used in the Fogg Museum exhibit as an adjunct to a classics course, and several Perseus workstations were incorporated into the "Greek Miracle" exhibit at the National Museum of Art (NMA). In both cases, patron questionnaire data were collected and interviews were conducted at the NMA. Marchionini and Crane (1994) report the results of the evaluation to that time with emphasis on the comparative study. Neuman (1991); Morrell, Marchionini, and Neuman (1993); and Marchionini, Neuman, and Morrell (1994) report details for different aspects of the evaluation through the first half-decade.

In 1995, funding to evaluate the educational and scholarly productivity aspects of Perseus was obtained from FIPSE. Although the overall evaluation plan continued to serve as a rubric, this funding marked an important juncture in the evaluation focus as the main efforts shifted to teaching and scholarship as the primary emphasis, with learning as an indicator of teaching effects. The first three-year cycle focused on building and assessing instructional materials that incorporated Perseus resources, and the second three-year cycle focused on how Perseus influenced scholarship in support of teaching. The first three years essentially extended the original evaluation plan, although with substantially less support, and the second three years extended the evaluation to the research and scholarship goals of Perseus. Annual reports for these evaluations were produced, and several are available in the Perseus DL under the teaching with Perseus division (http://www.perseus.tufts.edu/FIPSE/) (see Marchionini, Scaife, & Crane, 2000, for a recent report).

RESULTS SYNTHESIS

Many of the detailed results of the evaluation are available in published reports or WWW sites, so the main findings are briefly summarized here. The results are organized into five categories: physical infrastructure; conceptual infrastructure; mechanical advantage; augmentations; and community development/systemic change.

PHYSICAL INFRASTRUCTURE

The results from the early years were strongly influenced by technology. In every interview and observation, issues of using the physical components of the PDL arose. Whether working with a standalone HyperCard application on a dedicated workstation in a lab, or a network sharing the PDL from magnetic or CD-ROM stores, there were recurring problems with hardware and software reliability. These problems were organized under a category labeled "physical infrastructure." These problems were due to a variety of factors, including creating and delivering commercial-grade systems with personal computer platforms and development environments; primitive computing support for faculty offices and laboratories in most universities, especially for the humanities; lack of technical support and training staff; low levels of computer literacy on the part of humanities faculty and students; complex interfaces due to wide ranging content and tools; and poor mass delivery technologies (CD-ROM and networks). In the mid-1990s, moving Perseus to the WWW and increasing experience with digital technology by students, faculty, and universities substantially reduced the number of complaints and comments related to physical infrastructure. Although the interface is still complex due to the variety of material and tools for search and analysis, Perseus is generally accessible twenty-four hours a day, seven days a week to anyone in the world who has an Internet connection. Thus, substantial progress has been realized toward the general mission of providing widespread access to humanities source materials. In hindsight, it is important to note that, if the evaluation of the PDL had ceased after the first four years, the many problems related to physical infrastructure would have likely dominated the results and stymied continued development and decision making. This point again illustrates the importance of taking a longitudinal approach to evaluation.

Conceptual Infrastructure

Along with the need for physical infrastructure, the evaluation quickly made it apparent that there is little guidance and support for using DLs. Teachers do not know how to integrate these resources into their instruction; students do not know how to best use these resources when guided by assignments, let alone direct their own learning with digital resources; scholars do not know how to leverage DLs and incorporate their scholarship into them; and organizations such as universities, museums, and publishers do not know how to value, reward, and adopt DLs and the people who work with them. The FIPSE grants aimed to address some of these issues, and we have some good examples of instructors who are integrating their research and teaching through the DL. For example, early in the project, Neel Smith guided students through data collection activities with geospatial data in Perseus to discover new relationships in the altitudes of cities that minted coins in ancient Greece. Today there are several new examples at the Stoa site (www.stoa.org). For example, Nick Cahill is using his research on Olynthus as part of his course materials, supplementing his scholarly book with online materials, and Christopher Blackwell and a group of collaborators are creating online resources and a public forum related to Athenian democracy. The Perseus DL section on teaching includes information (e.g., syllabi, links, assignments) on fifty-one courses offered by twenty-seven different instructors at twenty-three different institutions. Although there is yet much progress to be made on creating and using digital resources in learning, the Perseus DL and its related projects are providing raw materials as well as pedagogical models for conceptual infrastructure.

Mechanical Advantage

A class of anticipated results showed up repeatedly over the years. The PDL provides people with more information more quickly than otherwise possible. These mechanical advantages provided by the digital medium were evidenced in several ways. First and foremost, access to large amounts of information is available with a few mouse clicks rather than grabbing a book (or, as many students pointed out, walking to the library and finding the book) and finding information manually. Such access is especially important in smaller schools where the library collection does not include broad ranges of texts nor multiple copies of common texts; where slide collections are not extensive and may be highly restricted to faculty and graduate students; and where the artifact collection does not have many examples of vases, sculptures, or other objects. In the WWW environment, this applies beyond the PDL as links in and out lead to much broader arrays of resources. This access is even more crucial from homes; referrer logs show many distance education courses that link to the PDL. Second, selective access is improved and faster for electronic materials.

Students were able to do word lookups in the Greek lexicon faster in their translation courses, although some students did express concern that the "ease" of lookup might diminish their translation skills. More common examples were the advantages of doing word searches in Greek or English, which facilitated finding relevant passages when writing essays or constructing arguments.

The results of the comparative study were statistically significant: students found more unique citations when using Perseus than not using it. However, there was no significant improvement in the overall quality of the resulting essays. Third, instructors are able to present more varied examples in class and use these examples in a more facile manner. The range of images and texts in the PDL exceed what might be shown with slides, and moving between images, texts, and maps is easier with a computer than multiple slide projectors, video recorders, and overhead transparencies. It is important to qualify this with the observation that setting up and using a computer and projector and mastering the PDL interface is a requisite for such usage that has become less onerous today than it was ten years ago. Fourth, instructors can create directed paths through the materials rather easily. The early versions of Perseus included a path tool that allowed users to record selected portions of their traversal of the database. In the WWW environment, instructors can easily provide sets of URLs interspersed with commentary or questions as part of student assignments. Likewise, students can easily create their own paths/electronic presentations or add URLs in their word-processed papers.

In all these cases, it is important to note that mechanical advantage alone is not sufficient to improve learning or critical thinking. In fact, mechanical advantage raises many issues about learning in electronic environments. What to do with the time that might be saved? How to deal with possible information overload? How to integrate results from multiple data sources such as texts and vases? Students and instructors sometimes worried about how easy it is to focus on searching and examining results rather than the more challenging activities of reflecting on meaning and creating one's own interpretation of the evidence. As Tom Martin, a Perseus advisor and early user, noted in an interview: "Collecting data comes more easily than interpreting it." Clearly, broad fast access to source materials has been made possible by the PDL, thus achieving one of the guiding missions of the project. The evaluation results answer the questions of what people do with this access and how they manage the new challenges brought by such access. Many examples of exasperation and rebellion were found—e.g., students who strongly preferred reading assigned secondary works and writing essays rather than conducting investigations in masses of data to discover relationships and make interpretations that might be presented as Perseus paths or Web pages. On the other hand, other students reported being inspired by the self-directed exploration and multiple media. Clearly, individual differences, such as motivation, learning style preferences, and domain knowledge as well as the classroom and university setting, influenced these reactions for students as well as instructors. Rather than focusing on these differences, we early on began to look for examples of ways that the PDL empowered new types of learning and teaching, going beyond the amplifications of mechanical advantage to new augmentations made possible by the PDL.

Augmentations

Clark (1983) long ago warned researchers to avoid media comparison studies when assessing educational technology impacts since the many variables cannot be controlled. Kozma (1991) and Salomon (1979) have argued that the symbol systems of different media do in fact strongly influence learning. All agree that it is extremely difficult to do comparative studies of learning effects. As part of the PDL evaluation, several comparative assessments were made, yielding no definitive effects but rather reinforcing the "it depends" conclusion. Moreover, identifying new effects using existing treatments and metrics is unlikely, and the interviews and observations yielded interesting anecdotes which suggested that new kinds of teaching and learning were emerging. Thus, we began to look for specific examples that would demonstrate how the PDL augmented learning and teaching. Four classes of augmentations are briefly summarized here.

First, students who had no Greek language were able to apply the philological tools in the PDL to investigate the meanings and nuances of Greek words and associated concepts. Different instructors have used variations of this activity, but the main idea is to explore an important cultural concept (e.g., concepts such as wealth and honor) by: (1) first looking for all occurrences of the term in the Greek to English lexicon (a simple search in the lexicon), thus locating all the Greek words that have this term in their definitions, (2) locating all occurrences of those Greek words in the Greek text corpus (a set of simple searches in the texts), and (3) reading the English translation of the section of text containing the Greek term (users can display Greek or English versions). By doing so, students were able to see that concepts such as "wealth" carried modern connotations (gold, animals, etc.) but also that the "house" in the sense of family and lineage was an important facet of wealth in ancient Greek culture. Such investigations reflect the kind of work scholars do to build interpretations about ancient cultures. These investigations would have been impossible without the electronic corpus of Greek text and translations and the associated indexes and lexicons. In interviews with students, both strongly positive and negative reactions were voiced—the positive centered on the exploratory investigation, the negative on the time-consuming nature of the searches when a treatise on the topic could have been read more quickly. Another student wrote on the questionnaire: "The amount of information initially overwhelmed our research group. It took many hours to glean any meaning from the text references Perseus provided. But those many hours would have been many weeks if it was not for Perseus. In the end, Perseus gave us the ability to come up with an intelligent view of what Herodotus thought about freedom." Clearly, the motivations and styles students bring to tasks affect the effects any DL will have.

Second, students were able to leverage the ready access to volumes of text and visuals to make discoveries and amass evidence to support their discoveries—the PDL is a laboratory for humanities research. One student spent the semester studying vase paintings to determine how women were depicted and gathered evidence that mortal women, except deities and hepatia, were always depicted in subservient positions to men. Another student discovered anomalies in vase paintings depicting hoplites without sandals but line drawings in books showing them with sandals. A treatment of the veracity of the historical record became a scholarly theme that grew out of simple curiosity about why, on a vase painting, Sciptians wore sandals but heavily armored Greek soldiers did not. This particular student noted that he spent about fifteen hours looking at images before he made his "discovery." Without traveling to many museums to study these vases, neither of these students would have melded visual thinking into their written work. Other students were able to use text tools to do first rate research. One student noticed that Herodotus uses the concept of "catastrophe" as if it were an infectious disease. She then investigated the usage in Lysias to compare how a historian and an orator used the concept and discussed the overlaps and distinctions of a "crippling agent" in human affairs. Students preparing for a summer course trip to Greece used Perseus to prepare tours of specific sites. Each student was responsible for leading the tour of a site and used the site maps, site photographs, as well as background information in Perseus, in preparing the tour. Clearly, not all students make discoveries, and significant portions of time are spent "surfing" for interesting connections but, as these examples pile up, the value of easy access to large volumes of data begins to emerge.

Third, there were instances of teams of students or students and professors collaborating around the PDL as "electronic campfire." In addition to the example of the coinage and altitude correlation noted earlier, other instances of spontaneous, as well as forced, collaboration were cited. In one case, a graduate student and professor using word analysis tools discovered that the morphological variation Antigone uses to refer to herself is distinctly different from how others in the play refer to themselves. They argued that Euripides used this as a lexical device to reinforce her alienation from the rest of society. Another professor was elated by a discussion that took place in class as the group explored the use of terms for freedom in Greek democracy. He noted new insights he had during class (and shared his excitement with the class) in spite of being a seasoned authority on the subject. Several instructors used group projects and "labs" in their courses. Not surprisingly, the results were mixed. One instructor noted that, although in his class Perseus tended to encourage more group work which, in turn, led to more idea sharing and clarity in expressing those ideas, there were also cases of homogenization of thinking and abandonment of responsibility. He summed the collaboration nicely as: "You get brand new highs and brand new lows." The message here regarding a wider range of diversity in behavior has important implications for instruction, DLs in general, and evaluation.

Fourth, the PDL provides the material and tools for new forms of creative expression. In the early days, students used the path tool to create paths through the corpus that represented their interpretation of assignments or ideas. In the WWW environment, students create Web pages that integrate PDL materials or word-processed documents that have images, texts, linguistic analysis results, and live links to the primary materials that support their arguments. Instructors increasingly require students to create such expressions rather than traditional essays. The PDL has become an especially useful resource for humanities scholars as they do research and incorporate their results into instruction. Scholars investigating word senses and uses as part of their translations and interpretations of Greek texts have used the philological analysis tools in Perseus. To explicitly support scholars in many classics subfields, the Stoa consortium was founded in 1997 to create a venue for scholarly research and instructional support. Stoa (www.stoa.org) provides tools and advice for scholars creating electronic documents, develops standards for tagging and displaying these products, and offers an electronic publishing platform for sharing their work and eliciting scholarly feedback. To date, thirteen ongoing research projects are included in the Stoa, and many of these projects incorporate PDL materials and serve to extend the PDL.

Community Development/Systemic Change

Perhaps the most important long-term developments are changes at organizational levels, such as departments and schools, and the emergence of a community of practice that leverages and advances the PDL. The original project depended on advisory committees that served as liaisons to the larger technical and educational communities. The team also gave talks at regional and national conferences and published papers in journals to inform the broader community about PDL and generate interest and reaction. In addition, the evaluation team conducted site visits to a number of universities which, in addition to the main objective of gathering evidence of use, caused local self-reflection on PDL practice. Instructors and students who participated in interviews were surely cognizant of what they said and likely reflected afterward on the interview and the

PDL experience. In addition, member checks (asking interviewees to review the written summaries and interpretations of their comments) facilitated reflections days or weeks later as participants reviewed the summaries. All these efforts served to alert classicists and instructional technologists to the goals and progress of the PDL.

The first tangible example of system change outside of the project staff settings was at Ball State University (BSU) in the mid-1990s. At site visits, we noted that four professors were using Perseus in their courses, and the department had committed the bulk of its materials and equipment resources to acquiring equipment to deliver the PDL. The faculty leveraged PDL use within the university to garner support for technical innovations that were strongly encouraged by the administration (e.g., faculty served on university-wide committees devoted to educational technology and were profiled in campus publications). In interviews with classics majors, students noted how they had used PDL in their assignments and projects and expected that they would have access to PDL in their graduate programs at other universities (which at the time was highly unlikely). The faculty discussed ways to build PDL into introductory courses so that students in advanced courses could be expected to take advantage of PDL without special instruction. In effect, the PDL was becoming institutionalized—part of the educational culture—in this department even before the WWW version was widely available.

These developments at BSU signaled new developments in classics departments and the field itself. By the late 1990s, job postings for faculty in several classics departments began to include requirements or preferences for technological skills. The core Perseus faculty were obtaining tenure and promotion at the schools to which they had moved. Yale University Press sold out of the first run of CD-ROMs, and the second edition served as an alternative to the WWW-based PDL. New textbooks began to include supplemental course materials that incorporated Perseus, and some online encyclopedias began to link users to the PDL. Another indication of community acceptance is the continued success of the project in receiving funding in highly competitive research initiatives. Peer-review funding sustained over multiple funding cycles demonstrates a level of prestige and usefulness within scholarly communities. The PDL has attracted funding in competitions judged by the humanities, information and computer science, and educational communities. Likewise, both the CD-ROM and WWW versions of the PDL have received many awards (e.g., more than four dozen awards, reviews, and certifications are listed on the WWW site in the summer of 2000).

As the WWW version of the PDL continued to evolve, the evidence of its impact on the field grew as more and more people accessed the corpus. Figure 1 summarizes WWW requests over a four-year period. Note that these numbers represent page requests rather than all transfers on a page (e.g., a page with five GIFs counts as one request even though the transaction logs contain six http requests). Note that spikes in usage recur during academic periods. In spring 2000, the PDL was responding to as many as 250,000 requests per day. The AltaVista portal listed almost 30,000 links to the PDL home page in mid June 2000 compared to almost 56,000 for the Library of Congress home page. Thus, the transaction log data provide another powerful indication that the PDL has become an important part of the humanities infrastructure.

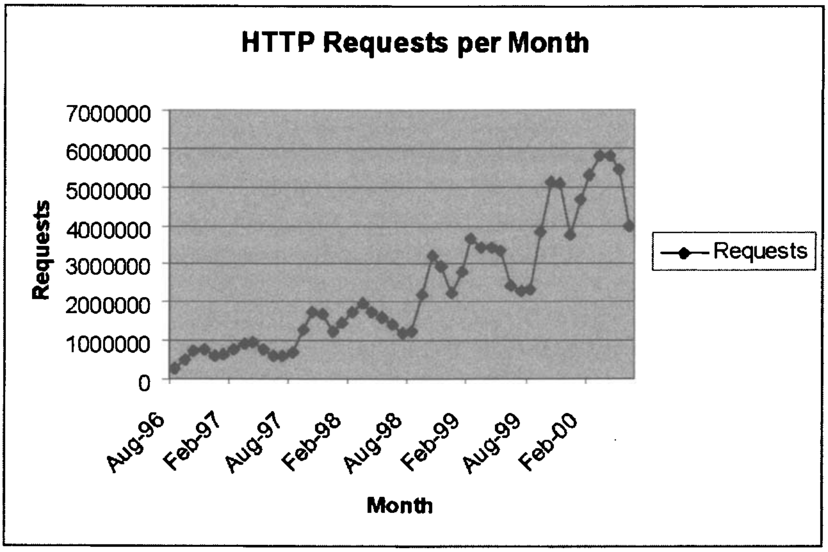

Figure 1. Perseus HTTP Requests.

Yet another indication of the influence of the PDL in education and the scholarly community is its expansion from Greek and Roman culture to other humanities areas as holders of important intellectual property are drawn to the technical and editorial expertise that has accrued at the PDL. The issues of editing and managing large corpora of source materials have led to demand for new roles and new skill sets for scholars in the humanities. The need to explicitly address the challenges of training and promoting scholars who demonstrate both domain and technical excellence was made explicit in a recent paper (Crane & Rydberg-Cox, 2000) that called for post doctoral positions in the humanities to support corpus editors.

Reflections and Recommendations

The evaluation effort has explicated many of the outcomes and challenges related to the PDL over more than a decade. The results above provide guidance to other DLs and shed light on questions at many levels. Is the impact worth some $10 million of investment? How has the PDL influenced learning? Teaching? Scholarly research? Is the PDL still viable? Sustainable? How does the evaluation inform continued development? The evaluation offers partial answers to all these questions as discussed in the results synthesis above. In this final section, attention is focused on the evaluation process itself and on recommendations for digital library evaluation in other settings. Three main points about DL evaluation with associated corollaries are offered:

1. *Evaluation efforts must explicate goals on a continuum ranging from evaluation research to product/system testing.*

At the research end of the spectrum, the goals are related to understanding complex phenomena through inference and chains of evidence. At the product/system-testing end, the goals are related to direct measurement of well-specified criteria that inform practical decision making. Most evaluation efforts in academic settings fall somewhere in between, using direct measurement and inference chains to build arguments and cases that inform decision making and continued development. In the PDL evaluation plan, formative and summative components were originally specified to distinguish this continuum.[9] What is important for evaluation research is to gather and integrate as many specific measures as feasible without depending too heavily on any single measure. Metrics such as number of HTTP requests, number of objects digitized, response rates, server down (or busy) time, interface feature sets, error rates, number of abandonments,[10] satisfaction ratings, interview comments, e-list traffic, cost per request,[11] and interesting anecdotes provide important glimpses into the DL phenomenon and context but individually do not provide a full view. It is surely possible and necessary to measure these results, but they are only threads in the more substantive evaluation questions asked in a research vein. Evaluation research that incorporates multiple data threads yields a complex fabric of effects that itself changes shape and meaning depending on the light and angle of view. Clearly, explicating evaluation goals on the research side of the continuum implies more cost and time commitments—factors that must be taken into account as evaluation is planned. If limited resources are available, focusing on a small set of well-defined system effectiveness questions may be prudent. However, because DLs are emergent systems, more ambitious evaluation efforts that gather baseline data and track changes over time are encouraged because they will more strongly benefit the DL community in the long run.

2. *Digital libraries are emergent complex systems.*

Digital libraries meld electronic tools and procedures with the entire range of forms of human expression. This includes new forms of expression made possible by the technology associated with DLs. The resulting complexity may yield effects that are greater than the sum of the parts—i.e., emergent properties (Kauffman,1995). The PDL offers some hints of this in the ways that students are empowered to do research and investigations traditionally reserved for graduate students, professors, and other scholars; in the implications of the PDL as a humanities laboratory; in the incorporation of PDL sources into the basic infrastructure of classical studies; and in the need for new skill sets for humanities scholars who leverage technical tools to create new interpretations and expressions. If DLs are emergent phenomena, DL evaluation must surely be designed to seek unexpected outcomes. Two characteristics that support such an evaluation are longitudinality and flexibility.

The original evaluation plan aimed to address a number of general questions over time by using a variety of data collection and analysis techniques. The plan was designed to be flexible in that techniques and questions could be adapted as the PDL itself and the technological and cultural contexts changed. This flexibility and attention to the interactions between the PDL and its environment is a defining characteristic of what might be termed "emergent evaluation." Emergent phenomena are driven by a small set of rules that control how systems interact with the environment (Clark, 1997). In the case of evaluation, the rules are determined by a high-level mission and data collection techniques. For the PDL, evaluation of the emergent phenomena of electronic resources in the humanities is controlled by the mission of broad access to source materials and a set of techniques that are adapted to the environmental conditions in which the mission operates. Over a dozen years, the environments included many physical sites (various instructional and research settings), a range of physical infrastructure developments (from single stand-alone PCs in a department office running primitive hypertext software to high-end workstations in dorms, homes, and offices linked through the WWW delivering software and content supported by 24/7 campus system administrators), and a growing range of conceptual infrastructure (from novice students and teachers with no experience using technology in learning and teaching to highly computer literate students and faculty). If the evaluation had concluded after five years, the effects of physical infrastructure would have dominated the results. The longer view illustrates some of the more substantive effects of the PDL as a critical infrastructure for the humanities and an important focal point for computing in the humanities.

A rationalistic approach to evaluation would compare the effects of a DL with the effects of a physical library on library effectiveness metrics. Various components effects (e.g., technical, content, and individual user) would ideally be separated out and weighted to produce some predictive model that explains performance and informs subsequent design. Unfortunately, such an approach oversimplifies both the components and metrics and, more importantly, does not take into account the higher-order interactions that emerge when dynamic systems operate in the real world. Positive outcomes on one dimension lead to unpredictable side effects on another dimension. In contrast, the longitudinal and multifaceted approach taken in the PDL evaluation was able to tolerate contradictions and side effects by looking at long-term effects and juxtapositioning data from multiple sources. The many specific questions under the four classes of main questions served to guide questionnaire and structured protocol developments and guide observations but, because the intention was to develop a high-level road map rather than a detailed blueprint, a richer process that captured some of the emergent properties of the PDL was possible.

3. *Integrate statistical data and narratives to assess impact as well as performance and usage*

Like circulation and holdings data in physical libraries, transaction log summaries and other performance data demonstrate operational effects of the library but do not explain how this usage impacts stakeholders. Marchionini has argued (Marchionini, 1995; Marchionini, Plaisant, & Komlodi, in press) that impacts change over time and vary by stakeholder (e.g., individuals, groups, organizations, society). Operational data are powerful components in a chain of inferences that address impact, but the PDL evaluation illustrates the value of anecdotes and "stories" that illustrate new effects—i.e., how DLs augment existing capabilities with new ones. These augmentations garner public support for a DL and should not be underestimated in assessing impact.

In evaluating DLs, it is important to consider the changes that DLs bring and only some of these are explainable through deltas in statistical data. Some of these changes are positive, but others will be controversial. Because human attention is a finite resource, new or additional capabilities can displace or reprioritize existing capabilities. Seeking and documenting these changes can be uncomfortable, especially when the evaluation funding is tied to the DL. Integrating multiple views is more naturally done with narratives rather than summary statistics, and integrating these forms of evidence can aid in assessing complex change.

Similarly, it is evident that DLs will lead to more diversity. Beyond the obvious broader access by global populations of users, the ranges of materials and new tools for access and use lead to new "highs and lows" of human application. Evaluation that treats these ranges may not make funding bodies or traditional user communities happy, but they will prepare the way for the changes that are inevitable as new information phenomena take their place in the realms of education and scholarship. Narrative explanations are crucial here as well.

The PDL evaluation reveals some of the complex interactions among information resources, stakeholders, and technology. Several observations about the success of the PDL may inform the development and evaluation of other DLs. Digital library success is aided by:

- clear missions;
- strong leadership and a strong talent pool;
- good technical vision and decisions;
- quality content and data management;
- giving users multiple access alternatives; and
- ongoing evaluation effort.

As we worked to understand the mission of Perseus in the early days, many metaphors were used by the staff to describe the project vision. One constant was the mission to maximize access to source materials. This mission immediately led to many benefits that garnered support and understanding. These include: allow people to make their own interpretations rather than learning accepted dogma (facilitating critical thinking); critical mass of content facilitates new discoveries; extending appreciation of classical culture makes classics more viable and sustainable in university curricula. This clear and populist mission served the PDL well over the years, and other DLs can benefit from a clear and crisply articulated mission statement.

Leadership is important to organizational success and, although DLs may exist in virtual space, the resources behind the scenes are real and must be assembled, inspired, and managed. The PDL has had a single chief and highly stable steering group throughout its history. Additionally, the talent pool offered by several major universities supplied the manpower to build the PDL. Continuous leadership works with a clear mission to attract and inspire such talent. Knowing the needs of stakeholders is important to leadership—this includes potential funders and discipline leaders as well as staff and end users. Tenacity and commitment lead to ongoing dissemination and evangelizing that garner support and usage. Although this is hardly surprising in any organizational setting, it is worth pointing out that it clearly extends to DLs in cyberspace.

The PDL benefited from staff and advisory boards that recognized important trajectories in technical development. Good decisions about storage and dissemination (e.g., CD-ROM for standalone and LAN use, early adoption of WWW), system architecture (e.g., object-oriented design, database lookups rather than hard-coded hyperlinks, and current emphasis on open source tools), and multi-platform delivery (e.g., standalone and WWW) all allowed the PDL to evolve while technology changed dramatically. It is easy to say that good technical decisions are required for DL success but harder to put into practice. Excellent advisory groups that represent different points of view and are willing to give regular attention to progress are difficult to assemble. They should not be political bodies—

once they are assembled, the DL staff must be prepared to take full advantage of their advice.

Likewise, decisions about content and data models are crucial to success. The PDL collection development policies were informed by an advisory board and are also opportunistic. It was aided in some sense by a finite set of ancient Greek texts, but there were many arguments over which texts to include and which museums and sites to approach for images. Good decisions about data models (e.g., catalog card metadata for art objects), data descriptions (e.g., adoption of SGML), and a general focus on content rather than technology may have been aided by the fact that classicists understand the importance of persistence of data and were prepared to plan for a DL that would outlive the latest technical solution.

The PDL grew out of the early hypertext research of the 1980s and thus was rooted in the notion of giving users control through multiple links and access points. The philological tradition of concordances and systematic word searches brought advanced search capabilities to the corpus from the earliest days. This combination of support for browsing and analytical search supports diverse usage and is an important lesson for other DLs. Giving people control over how they access and use the DL satisfies a broader range of users and gives rise to wider ranges of applications.

Finally, an ongoing evaluation effort serves multiple purposes. Evaluation serves a political/administrative role by providing the reports and data upon which decisions about funding and development may be based. Evaluation results also inform the ongoing development of the DL both technically and conceptually. Evaluation activities also serve to involve staff and users in the work of the DL at reflective levels that may improve usage and support. Evaluation serves to document the evolution of a particular DL. Most importantly, evaluation with a research focus helps to explain the effects of a specific DL and relate it to the larger issues of DL evolution and impact.

Perseus has always been a library. Although it was not fashionable (or fundable?) as a digital library in the mid to late 1980s when libraries were written off as anachronisms and library and information science programs were closing, the aim of making source materials widely available, the emphasis on self-directed learning, the organizational schemes applied for preserving and ensuring scholarly access, and the access and analysis tools created all reflect information science theory and practice. Perhaps better than the other answers about hypertextual systems, digital objects, interactive curricula, and communities of practice, the best answer to Ehrmann's question is that Perseus is a digital library. It has continued to evolve and stands as a significant city in a cyberspace that is now being defined by a vast network of linked digital libraries with tangible connections to the physical artifacts that make up our world.

Acknowledgments

Over the years, it has been my great fortune to be able to draw upon a diverse range of talented people who participated in the evaluation. Kenny Morrell was an original member of the Perseus team and helped shape the original evaluation plan and conduct the early studies. Delia Neuman was my original research partner and worked with Morrell and me to develop the plan and was instrumental in collecting data in the early years and creating analysis strategies. Steve Erhmann, who served as program director at Annenberg CPB in the early days, prodded us with excellent questions and supported us as we elaborated on our approach to longitudinal formative evaluation. Peter Evans and Jennifer Powers served as multiple-year graduate assistants for the evaluation. Greg Crane, the project director, recognized the value of evaluation early on in the project and he, Elli Mylonas, Tom Martin, Maria Daniels, Neel Smith, Ross Scaife, Nick Cahill, David Smith, Sebastian Heath, Lisa Cerrato, and many other "Perseids" gave many hours of their time helping us understand what Perseus means to them. Likewise, many instructors at different institutions allowed us to come into their classes and tolerated our interviews over the years. Special thanks in this regard is due to Walter Moskalew, Chris Shea, and Bill McGrath at Ball State University who hosted several visits and showed us the first indicators of systemic change when multiple faculty members in a department begin to adopt a new paradigm for teaching. Partial funding for the evaluation was provided by the Annenberg/CPB Project and the Fund for Improvement of Post Secondary Education.

Notes

[1] A path mechanism that teachers and students could use to record sequences through the corpus was provided as a tool.

[2] See Daniels & Chavez (1999) for instructions for contributing to the Perseus Digital Library collection.

[3] The current PDL offers CD-ROM and WWW versions. Because some museums restricted image use to CD-ROM, the image collections differ according to intellectual property rights agreements. Likewise, full Greek lexicon versions and some materials beyond the original Greek culture corpus are only available in the WWW version. Additionally, the technical capabilities differ slightly as WWW display and transfer capabilities differ from what is possible with locally mounted CD-ROM. See http://www.perseus.tufts.edu/order.html for a comparison chart of the two instantiations of the PDL.

[4] Bloom, in his taxonomy of educational objectives, poses evaluation as the highest of six cognitive goals for learners (Bloom, 1956).

[5] An example of a comparative question in the learner group is 4.2. In different courses with similar objectives, are there differences in the amounts and kinds of achievement for courses supported by Perseus and those not supported by the system?

[6] An example of a usage question in the instructor group is 2.1. What tactics, strategies, and patterns do instructors develop with Perseus in connection with their teaching? Research?

[7] An example of a system question is 3.4. How are student-defined paths used? Shared?

[8] An example of a content question is 3.2. How well can Perseus help the student or scholar clearly delineate fact from scholarly conjecture?

[9] Flagg (1990) provides a good set of formative evaluation case studies related to educational technology projects.

[10] Hert and Marchionini (1997) report 50% abandonments in 1996 from non .gov, .com. and .edu domains at the BLS Web site but half as many abandonments a year later. Thus, these data must be considered over time.

[11] It costs the government $14 if a person requests federal tax forms by phone, $7 by mail, and $3.50 by walking into a post office to pick them up in person. Online requests for those forms cost the government 13 cents (Trimble, 2000). The cost of the citizen's time to obtain the form is an additional factor beyond the government savings.

References

Bednarz, D. (1985). Quantity and quality in evaluation research: A divergent view. *Evaluation and Program Planning, 8*(4), 289-306.

Bloom, B., & Krathwahl D. (Ed.). (1956). *Taxonomy of educational objectives: The classification of educational goals, by a committee of college and university examiners: Handbook I: Cognitive domain.* New York: Longmans Green.

Campbell, D. (1979). "Degrees of freedom" and the case study. In T. Cook & C. Reichardt (Eds.), *Qualitative and quantitative methods in evaluation research* (pp. 49-67). Beverly Hills, CA: Sage.

Chavez, R. (2000). Generating and reintegrating geospatial data. In R. Furuta (Ed.), *Proceedings of the Fifth ACM Conference on Digital Libraries* (San Antonio, TX, June 2-7, 2000) (pp. 250-251). New York: ACM Press.

Clark, A. (1997). *Being there: Putting brain, body, and the world together again.* Cambridge, MA: MIT Press.

Clark, R. (1983). Reconsidering research on learning from media. *Review of Educational Research, 53*(4), 445-459.

Cook, T., & Reichardt, C. (Eds.). (1979). *Qualitative and quantitative methods in evaluation research.* Beverly Hills, CA: Sage.

Crane, G. (1988). Redefining the book: Some preliminary problems. *Academic Computing, 2*(5), 6-11, 36-41.

Crane, G. (Ed.). (n.d.-a). *The Perseus Project.* Retrieved August 29, 2000 from the World Wide Web: http://www.perseus.tufts.edu.

Crane, G. (Ed.). (n.d.-b). Information about Perseus. *The Perseus Project.* Retrieved August 29, 2000 from the World Wide Web: http://www.perseus.tufts.edu/PerseusInfo.html.

Crane, G., & Mylonas, E. (1988). The Perseus Project: An interactive curriculum on classical Greek civilization. *Educational Technology, 28*(11), 25-32.

Crane, G., & Rydberg-Cox, J. (2000). New technology and new roles: The need for "corpus editors." In R. Furuta (Ed.), *Proceedings of the Fifth ACM Conference on Digital Libraries* (San Antonio, TX, June 2-7, 2000) (pp. 252-253). New York: ACM Press.

Daniels, M., & Chavez, R. (1999). *A guide to photographing architecture, monuments, sites, and topography.* Retrieved August 29, 2000 from the World Wide Web: http://www.stoa.org/guides/sitestds.shtml.

Ericsson, K., & Simon, H. (1984). *Protocol analysis: Verbal reports as data.* Cambridge: MIT Press.

Evans, P. (1993). *The enabling and disabling effects of a hypermedia information environment on information seeking and use in an undergraduate course.* Doctoral dissertation, University of Maryland at College Park. *Dissertation Abstracts International, 54*(10), 3621A.

Flagg, B. (1990). *Formative evaluation for educational technologies.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Griffiths, J-M., & King, D. (1991). *A manual for the evaluation of information centers and services* (AGARDograph No. 310). New York: American Institute of Aeronautics and Astronautics Technical Information Service.

Harter, S., & Hert, C. (1997). Evaluation of information retrieval systems: Approaches, issues, and methods. *Annual Review of Information Science and Technology, 32,* 3-94.

Hert, C., & Marchionini, G. (1997). *Seeking statistical information in federal Websites: Users, tasks, strategies, and design recommendations. Final report to the Bureau of Labor Statistics, July 1997.* Retrieved August 29, 2000 from the World Wide Web: http://ils.unc.edu/~march/blsreport/mainbls.html.

Kauffman, S. (1995). *At home in the universe: The search for laws of self-organization and complexity.* New York: Oxford University Press.

Kozma, R. (1991). Learning with media. *Review of Educational Research, 61*(2), 179-211.

Marchionini, G. (1995). The costs of educational technology: A framework for assessing change. In H. Maurer (Ed.), *Educational multimedia and hypermedia, 1995* (Proceedings of ED-MEDIA '95—World Conference on Educational Multimedia and Hypermedia, Graz, Austria, June 20, 1995) (pp. 33-38). Charlottesville, VA: Association for the Advancement of Computing in Education. Also available on the World Wide Web at: http://ils.unc.edu/~march/costet/costet.html.

Marchionini, G., & Crane, G. (1994). Evaluating hypermedia and learning: Methods and results from the Perseus Project. *ACM Transactions on Information Systems, 12*(1), 5-34.

Marchionini, G., & Fox, E. (1999). Progress toward digital libraries: Augmentation through integration. *Information Processing & Management, 35*(3), 219-225.

Marchionini, G.; Neuman, D.; & Morrell, K. (1994). Directed and undirected tasks in hypermedia: Is variety the spice of learning? In T. Ottman & I. Tomek (Eds.), *Educational multimedia and hypermedia, 1994* (Proceedings of ED-MEDIA '94—World Conference on Educational Multimedia and Hypermedia, Vancouver, BC, June 25-30, 1994) (pp. 373-378). Charlottesville, VA: Association for the Advancement of Computing in Education.

Marchionini, G.; Neuman, D.; & Morrell, K. (1989). *Perseus evaluation plan.* Unpublished Perseus Project Working Paper Number 5, Dec. 1989.

Marchionini, G.; Plaisant, C.; & Komlodi, A. (forthcoming). The people in digital libraries: Multifaceted approaches to assessing needs and impact. In A. Bishop, B. Buttenfield, & N. VanHouse (Eds.), *Digital library use: Social practice in design and evaluation.* Cambridge, MA: MIT Press.

Marchionini, G.; Scaife, R.; & Crane, G. (2000). *Final evaluation on the Perseus Project publication model, 1997-2000.* Retrieved September 12, 2000 from the World Wide Web: http://www.ils.unc.edu/~march/perseus/final_report.pdf.

Morrell, K.; Marchionini, G.; & Neuman, D. (1993). Sailing Perseus: Instructional strategies for hypermedia in the classics. *Journal of Educational Multimedia and Hypermedia, 2*(4), 337-353.

Mylonas, E. (1987). *Using Perseus in a variety of educational settings.* (Perseus Working Papers, 2), Cambridge, MA, December, 1987.

Neuman, D. (1991). Evaluating evolution: Naturalistic inquiry and the Perseus Project. *Computers and the Humanities, 25*(4), 239-246.

Nielsen, J. (1993). *Usability engineering.* Boston: Academic Press.

Rossman. G., & Wilson, B. (1985). Numbers and words: Combining quantitative and qualitative methods in a single large-scale evaluation study. *Evaluation Review, 9*(5), 627-643.

Salomon, G. (1979). *Interaction of media, cognition, and learning: An exploration of how symbolic forms cultivate mental skills and affect knowledge acquisition.* San Francisco, CA: Jossey-Bass.

Saxton, M. (1997). Reference service evaluation and meta-analysis: Findings and methodological issues. *Library Quarterly, 67*(3), 267-289.

Suchman, E. (1967). *Evaluative research: Principles and practice in public service and social action programs.* New York: Russell Sage Foundation.

Timble, P. S. (2000). A question of trust. *Federal Computer Week.* Retrieved August 29, 2000 from the World Wide Web: http://www.fcw.com/fcw/articles/2000/0522/pol-trust-05-22-00.asp.

Voorhees, E., & Harman, D. (2000). Overview of the sixth Text REtrieval Conference (TREC-6). *Information Processing & Management, 36*(1), 3-35.

White, H., & McCain, K. (1989). Bibliometrics. *Annual Review of Information Science and Technology, 24,* 119-186.

Williams, D. (Ed.) (1986). *Naturalistic evaluation* (New directions for program evaluation, No. 30). San Francisco, CA: Jossey-Bass.