

**Developing Standardized Metrics for Assessing
Use and User Services for Primary Sources**

**Report of a Meeting Held in Chapel Hill, NC
Funded by the Andrew W. Mellon Foundation**

**Helen R. Tibbo
Brian Dietz**

July 2004

The “Developing Standardized Metrics for Assessing Use and User Services for Primary Resources” work group meeting was held on the Campus of the University of North Carolina at Chapel Hill on June 3-6, 2004. Funded by the Andrew Mellon Foundation, this meeting brought together project research staff: Dr. Helen R. Tibbo and graduate archival student, Brian Dietz from the University of North Carolina at Chapel Hill; Dr. Wendy Duff and graduate assistant Sarah Carson from the University of Toronto; and Dr. Elizabeth Yakel and graduate assistant Elizabeth Goldman from the University of Michigan with archivists and curators from across the United States and Canada. These academics and practitioner partners began work toward the development of standardized tools and metrics to assess use and user interactions within archives and other primary resource repositories at this meeting.

The 3-day event started with a dinner on Thursday evening where project partners informally gathered, sharing introductions and conversation. Dr. Elizabeth Yakel opened the working meeting on Friday morning with an introduction to team participants. She next addressed the absence of use and user-based statistics throughout the archival literature and profession, noting that such statistics could support a deeper understanding of the experience users have when researching primary sources and that they could provide a data-rich foundation for the administration of cultural heritage agencies that provide access to primary documents.

Dr. Yakel presented the two fundamental questions that would continue to be discussed throughout the conference: 1) What should we measure? and 2) how should we conceptualize measurement, both locally and in a broader sense? Yakel discussed four key goals that would need to be met on the way to answering the first set of questions. These goals are:

- the development of measurement support tools;
- the balance between global and local perspectives;
- the need, if any, to educate practitioners within the profession; and
- the desire to increase discussion of measurement within the archival community.

Dr. Yakel next proposed four desired outcomes of the meeting. The investigators planned to identify:

- the critical areas of measurement that might include assessments of archival reference desks, descriptive tools, and exhibits;
- the specific elements that archivists should measure;
- the best methods of collecting data; and
- the barriers to access from the perspective of practitioners and academics.

Wendy Duff concluded the opening remarks by inviting all conference participants to provide input, share their expertise, understand their roles as representatives of both the larger archival community and their own unique communities, and provide direct and honest feedback about the project, during and after the conference.

Dr. Tibbo extended a welcome from the School of Information and Library Science Information (SILS) and the University and participants introduced themselves to the group. Dr. Joanne Marshall, outgoing dean of the School of Information and Library Science at the University of North Carolina at Chapel Hill then provided a second welcome from SILS. After

introductions, Drs. Gary Marchionini and Barbara Wildemuth, professors from SILS with extensive expertise in evaluation and assessment methodologies, and Dr. Paul Conway, Director of Information Technology at Duke University Libraries and author of a seminal call for user studies within the archival community, gave keynote addresses that provided context for the remainder of the meeting. The program for the day's events with the slides from the presentations can be found at: <http://ils.unc.edu/metrics/intro.html>.

In his presentation, "Modeling Use and Evaluation," Gary Marchionini introduced user evaluation with the suggestion that while such studies are difficult, they are enjoyable, important, and rewarding. He pointed out that evaluation can satisfy a variety of needs, including providing a basis for service assessment, decision-making, and a data-rich understanding of a problem, conceived as the exploration of some hypothesis. Marchionini asserted that evaluation should be situated within this last point in order for it to provide for long-term understanding and action.

After noting that researchers needed to determine what to evaluate (e.g., products, processes, outcomes, satisfaction), why they were conducting the evaluation (e.g., basic research, performance quality, or designing guidance), and to select a method of evaluation (e.g., product testing, controlled comparisons), Dr. Marchionini highlighted six variables to be considered in designing and conducting user evaluations. The first and most central variable is the people involved, both individually and institutionally. The second element is the context for the service being evaluated. Context includes elements such as institutional mission, the physical or virtual setting, and the flow of work within the institution. His third variable is the information system architecture including backend configuration, middleware, and the user interface. Knowledge domain is the fourth variable. Evaluation design should conform to the structure of knowledge within a discipline (spatial knowledge of the geographer compared to chronological knowledge

of the historian), the rate of change of the discipline's knowledge base, and the scope, complexity, and interdisciplinarity of the knowledge. The next variable is the resources available to conduct the evaluation. How much time is the user willing to commit? How much money is available for the evaluation? The goal of evaluation is the final, and perhaps the most important, variable to consider in designing user studies. Designers should consider their stakeholders, including patrons, the knowledge domain, and the scope of evaluation, or the amount of time the measurement will cover—immediate, annual, mid-term, or generational. Marchionini proposed that archival evaluations be designed as generational studies in order to capture the effects of technology on our patrons, and that AX-SNet's evaluations extend to a longer term of fifty to one-hundred years.

In concluding his presentation, Marchionini outlined some alternative strategies of evaluation. AX-SNet could focus on the information-seeking context, follow a usability testing approach, conduct longitudinal studies via systematic case studies, or take the epidemiological approach whereby we start with the outcome and trace the influences (in this final approach Marchionini suggested the development of an information-retrieval interaction model that engages the user in the process). In each model, he stressed that information retrieval be considered a process of information seeking, rather than the matching of documents to queries. Dr. Marchionini provided several examples of evaluation studies from his work with the Bureau of Labor Statistics, the Library of Congress, and the Perseus Project.

SILS Professor, Barbara Wildemuth, presented "Overview of Methods," an explanation of designing evaluation studies. Wildemuth started her presentation by asking who of the participants had experience conducting user studies. Excluding the investigators, five of the participants raised their hands, but no published reports followed their studies.

According to Wildemuth, the basis of a successful evaluation is a clear question. No evaluation will be successful without a clear picture of what is being measured. Questions that might be asked include those about the current or potential users of a collection. Either question would likely follow or require a corresponding set of purposes of study, methods of measurement, and impact of evaluation. Wildemuth, in emphasizing the focus of a study, distinguished the evaluation study, which has a local purpose, from a research study, which tends to have broader purposes; each study occupies either end of a “generalizability of results” continuum. Evaluation studies need only be internally valid, while research studies need only be externally valid. The scope of a study can focus on a range of scale, from a particular activity within an organization, e.g., the functionality of a web site, to the full workings of the archives.

Another term Wildemuth defined was evaluation criteria. As she explained it, evaluation criteria are the manner in which you define what “good” means, which could include effectiveness at meeting the needs of users, utility/usefulness/impact or value of the collection, and user satisfaction with an archive’s collections and services. Criteria can be compared either to a pre-defined standard (criterion-referenced evaluation) or to some other alternative (norm-referenced evaluation). In developing criteria we conceptualize questions and data collection methods by which we will find the answers to our ultimate concerns, of whatever scope or focus.

Wildemuth suggested the two ways to collect data is by watching users and by asking questions. Methods of watching users include direct observation during use of materials, think-aloud protocols, and through analysis of transaction logs. Methods of asking questions include questionnaires (surveys or measures), interviews, and focus groups.

Dr. Wildemuth discussed the considerations in drawing a study sample that includes defining a population of interest of appropriate size and developing a sampling plan. The study

sample should be closely related to the questions asked. Examples could include remote users or potential users that are not current users. Sample size is a factor in how the study is conducted and how generalizable the outcome is; intensive studies will have a smaller sample, extensive studies a larger sample. Capturing the targeted population, or sample, requires a sampling plan, and includes probabilistic and non-probabilistic sampling options. A fundamental question is how representative of the entire population of users the sample must be. Like the study sample, the sampling plan should correspond to the phenomenon one is attempting to understand.

The next concern is the design of the research plan. In this phase, researchers consider when and how data collection is to take place, and attempt to control for the factors that could lead one to draw false or flawed conclusions. The research design must control for extraneous variables, the “external” factors a sampled individual brings to the study that influence how he or she responds during observation or to the questionnaire. Closely related to this is the desire for internal validity, the soundness of conclusions drawn from the study’s results. Threats to internal validity include chronology of events taking place during the study, the maturation of subjects, instrumentation, and subjects’ expectations of the study. Each of these can affect a subject’s responses to the study. Finally, researchers may need their design to achieve external validity if they aim to have results be generalizable across user populations.

Analysis of the data follows design and data collection and is either quantitative or qualitative. Quantitative methods of analysis are usually descriptive statistics or inferential statistics. Two examples of quantitative data Wildemuth provided were the number of users that have accessed a collection during a set time period and ratings of user satisfaction with an archive’s reference services. Before drawing conclusions from quantitative results the researcher must check for statistical significance and meaningfulness of results.

The steps in a qualitative method such as content analysis of textual data are specifying a unit of analysis, theme recognition, and the search for negative evidence. Qualitative analysis will usually yield themes that imply actions to be taken and positive and negative evidence that can be weighed against one another.

In 1986, in issue 49 of *American Archivist*, Paul Conway published the paper “Facts and Frameworks: An Approach to Studying the Users of Archives,” in which he explored a potential framework to be used in assessing how well archives were meeting the needs of their users.¹ Conway, the third keynote speaker at the meeting, presented “Facts and Frameworks Revisited,” in which he appraised his original article and offered potential new directions for the framework he set out nearly twenty years ago.

In disclosing his current assumptions about the nature of archival holdings in an age of digital access, the advent of which occurred after the appearance of his original article, he resituated his framework for evaluation considering the role of digital resources and tech savvy users. In considering the development of metrics for digital resources, Conway posed the question, “What do we need to know about digital use to help manage digital assets?” In asking if the payoffs of metrics are sufficient to engage in a standardization evaluation plan, he suggested that, as an administrator, he found cost-benefit analysis more important than output numbers as a means of justifying the existence of archives.

Conway discussed what from his original framework still resonates. The criteria of quality, integrity, and value, which could be conceptualized across his five methodological stages of measurement, could be integrated into a new metrics framework. These five stages—

¹ Conway, Paul L. “Facts and Frameworks: An Approach to Studying the Users of Archives.” *American Archivist* 49 (Fall 1986): 393-407.

registration, orientation, follow up, survey, and experiments—were found to have continuing relevance, as does his proposal to fit method of measurement to stage of access.

The presentation then turned to what was missed from the original article. Conway admitted his 1986 perspective was limited by a linear model of behavior, and thus measurement, and suggested the superiority of iterative assessment. Related to this was his over-reliance on the reference process rather than integrating the complexity of user behaviors. Conway felt that the 1986 framework is not conducive to program evaluation; nor is it sensitive to time.

Conway focused the rest of his presentation on what he saw as the importance of metrics, or what metrics could be used to measure. Metrics could be designed with the following five categories in mind: Quality of Access, Integrity of Assets, Value of Archives, User Collaboration, and Managing Archival Functions. Research and evaluation questions relating to quality of access include such areas as:

- Usability testing of websites, finding aids, and interfaces;
- Fundamental critique of EAD as an appropriate access mechanism; and
- Connection between use of artifacts and use of digital surrogates.

Studies concerning integrity of assets could focus on issues including:

- Credibility and trust;
- Effectiveness of navigation and juxtaposition;
- Context and its value for access;
- Impact of transformed functionality; and
- End-user assembled collections

Conway gave several examples of research that might focus on the value of archives including local impact, cost/benefit analysis, and the impact of intellectual property restrictions on scholarship and creativity.

Conway concluded his presentation with a discussion of lingering doubts. Asking whether standardization is worth the effort, he wondered if researchers within the field could utilize metrics consistently and if content could be made consistent. He also speculated if quantitative data is useful at the local and national levels, and if metrics contribute to the administrative justification of archives. He considered the nature of research agendas, asking if they were for academics only. This question is fundamental to the present project that has the goal of uniting academics and practitioners in the development of a sound data-driven base for use and user research for the archival profession.

A general discussion of the three presentations followed Paul Conway's talk. Discussion focused initially on the role of archives in preserving authenticity and the likelihood that this mattered little to users who only wanted seamless and on-demand access to information. It was suggested that we make an attempt to understand the archive from the perspective of the user rather than imposing our definitions and mappings.

Some participants questioned the relevance of use data for justifying the existence of archives. Gate counts matter little to an administration that already knows its facility is being utilized. Archives may have a monopoly, and if so, justifying its role is less important than understanding the user experience. Still, it was warned, archives are easy targets when budgets are being cut, and that data speaking to an archive's relevance might help save it from fiscal reduction. Participants agreed that quantitative data coupled with qualitative data, such as narrative stories, would have a positive impact. Data can be used to justify the archival role as

well as improve services. The findings of evaluations and user-based studies can add to an archive's self-awareness and security.

The discussion touched on the challenge of collecting information about users who are not being reached. There was a desire to know who these potential users are, why they are not coming, and how they can be enticed to use archival resources. A 360-degree feedback model was offered as potentially useful to this end, as it relies on small study samples. There was also an impression that archivists tended to ask only if they have the materials researchers are requesting rather than seeking to understand those who might not be served.

The group discussed the need to understand why use and user-based studies were conducted before developing a standardized metrics set. It may be unnecessary and less valuable to develop standardized metrics for use locally. "Big questions," wherein researchers wish to understand a phenomenon across the profession, require standardization. Reliable, rigorous methodological work will be required for such research studies to be valid.

After transitioning toward the nature of research with physical and digital artifacts and the difficulty users have when working with finding aids, the discussion returned to standardization. Some participants expressed the need to have access to raw user data that could be manipulated for various purposes, e.g., for grants, for administration. This was seen as an argument for developing a sophisticated set of tools and questions, with individuals advocating for collecting smarter data, not more data. Other points made included the idea that much data is currently being collected that is not being put to use, and that there is a need for more sophisticated data management systems. Partners asked how this project could facilitate or implement standards, and how we could anticipate what the best tool for a particular purpose would be. There was general support for the idea that if there was a toolset that assisted with

collecting a basic suite of use and user-based data it would be both well-received and widely used. The discussion ended with a comment on the importance of clearly defining methodology so as to make it repeatable and reusable, and there was an emphasis placed on the need to share findings.

Friday's proceedings ended with break-out group discussions. The investigators suggested four questions for the four groups to consider:

- If you had unlimited resources available, what don't you know that you would like to know about the use of your archives?
- Why do you want to know this (what value would this have)?
- How would you use this information if you had it? and
- What barriers are there to knowing what you would like to know?

Partners discussed wanting to know what categories of users exist and how use patterns break down by category; what records are used and why; why users thought a particular collection or archives would hold the information they sought; the lifecycle of searching, and information seeking behavior in general. Also discussed was the nature of the relationship between their physical and virtual collections and the patterns of use of online resources; the usefulness of archival descriptive tools; the effectiveness of outreach and instruction, including online instruction; how users come to trust, or view as authentic, archival holdings; and finding the products that result from archival research. Reasons given for wanting to know the above information include self-assessment of collections and services, being able to anticipate complementary materials for researchers, and being able to provide guidance to administrators in terms of resource allocation.

With this information, partners said they would be able to correct finding aids and online information to match users' needs in order to create more effective search processes; develop more efficient user training; identify potential resource streams; and retool outreach in order to lure potential users, which could include matching online presence with natural searching terminology.

The primary barriers participants saw between archives and the collection and analysis of user-based information were competing and increasing demands placed on archivists; limited time and resources; the fear of annoying users with requests for feedback; archivists' belief that they know best what users need; the conflict between how users and archivists see the advantages of physical and digital resources; and the threat that responses would be problematic or that demands could not be met. Other barriers mentioned included privacy concerns, the anonymity of online sessions, as well as the physical gulf between archivist and remote users; the difficulty of finding scholarly output based on research done in a particular archive; lack of skills and knowledge in developing instruments and interpreting data; users' desire for self-sufficiency; changing priorities of resource allocators; infrastructural support for data; and the murkiness in identifying the value of discrete archival functions.

Saturday's proceedings began with reflections on the first day of the conference and reports of the break-out sessions. The bulk of the discussion centered on the process of developing standardized metrics. Consideration was given to the diversity of archives, services offered, pattern of usage, and users—scholarly, administrative, genealogical, and recreational. Would different data gathering techniques depend on type of institution? Partners agreed that any survey or questionnaire would need to include a minimum number of core questions so that comparison across institution could occur. The training component of the project was

considered. A requirement of this project would include designing guidelines on how to take a number of questions from a bank and create workable surveys.

Partners discussed the manner in which use and user-based data collection modules would be developed, including the idea that work may begin with surveys of archivists to discover what is already known about users. It was proposed that partners and the institutions they were representing would volunteer to create component modules. Modules could then be presented to other institutions of similar mission, and they could suggest improvements. Updated modules could then be tested at institutions that had have not participated in the process. Testing could be carried out according to two methods, either testing the same data gathering tool at different types of archives or testing different types of tools in similar type archives.

The academic value of this project is the ability to collect data from different institutions and collate it in order to get a larger view of archival use, especially if measurements are crafted to be employed longitudinally. Participants voiced the need for an institutional infrastructure that would support data collection to be in place to ensure long-term viability of any metrics project and comparative studies. Others discussed the potentially important role of professional societies in developing standards and infrastructure. A final problem considered was working studies into the existing work-flow. There was a feeling that reports have to be easily produced; otherwise, such work is an addition and it may not get done.

Reports from the previous day's group sessions followed the opening discussion. Wendy Duff then reviewed the existing literature on user studies in archives, during which the partners interjected comments. Comments touched on the reluctance of institutions to publish any findings based on collected data and the needed change in institutional culture that could lead to the publicizing of findings. A suggested possible outcome of the conference could be a template

that institutions could use to post technical memos about user-based data to their websites. The value of outcome-based evaluation to funding institutions was asserted.

Beth Yakel followed Duff's literature review with an overview of the current practices of data collection in archives. Comments were made during this presentation, as well. One of the partners reiterated the need for this project to include a strong educational component. The impressionistic nature of current data collection was voiced, as was the problem of institutional memory, which is threatened to diminish with staff turnover. Someone raised a concern about the optimal number of individuals responsible for data collection and analysis within an institution. Finally, partners revisited the issue of the delicacy of user privacy.

The Saturday afternoon session began with further assessment of the conference's progress. The investigators emphasized the cooperative nature of the project; practitioners will provide the questions and goals of measurement around which the investigators will build principles, standards, and methods conforming to and measuring these areas of research. Partners responded by announcing a need for common data elements and a widely recognized way of pooling data. One partner wondered about the usefulness of beginning with the development of data elements and a data dictionary so that some standardization was provided at the outset. This would help partners in handling the range of ways of categorizing users, collections, and services.

A long-term goal of developing and employing standardized metrics would be gaining an understanding of how use patterns, as well as the profession in general, changes over time, especially as our web presence increases. The former could be a valuable tool in assessing appraisal—how does anticipated use match with actual use. Further, by presenting examples of archives that successfully achieve high-impact levels, user studies could serve to establish

benchmarks for peer institutions. In terms of data collection, it was suggested that a small number of institutions could compare what they collect and determine how distinct their efforts are, with special attention to the reasons for any existing gap in how users and resources are categorized.

There was common consensus in the need to standardize registration data, and that this might be a good first effort; but all felt that this was an effort better supported by professional organizations, like SAA and ACA, rather than funding agencies, bodies that might have more interest in knowing the scholarly impact of archives (partially in relation to financial costs). While there was a reiterated concern over how user studies would be worked into archives' workflow, an anticipated outcome was user studies enhancing workflow.

Partners then reformed their groups from Friday, and they were asked to consider three questions: What are the most important things you want to measure, what components of these things would you want to measure, and what would representative data be? Groups' representatives presented their results.

There was considerable overlap of results, as well as unique proposals. One item that most groups expressed an interest in measuring was researcher profile or registration data, including demographic variables and institutional affiliation. Other common values were what materials were getting used (with its preservation and budgetary implications), the path through which users arrived at an archive, and the purpose of research or proposed outcome of research.

One group stated the importance of focusing on the return on investment, or gauging an archive's impact in terms of its budget. Impact of this kind, it was later agreed, was difficult to get at, and the comment was made that, while this knowledge is ultimately crucial to have, we might have to decide if the cost of tracking down this information is worth the benefit of having

it. The same group emphasized receiving feedback on physical and virtual educational tools. Similarly, another group asked about the impact of archival services. Components of this included what services were of use, the satisfaction with services, outcome of research, and the changing view users may have as a result of their use of services. This second group also concentrated on understanding users' information-seeking patterns, which could include comparing strategies and tools on- and off-site visitors employ.

In anticipating how to better deal with the rising numbers of remote users and requests, another group wanted to measure the components of remote access. Who is coming to repositories remotely, why, and what resources are they using? Knowing this could have positive allocation consequences on reference services and web outreach.

Sunday morning brought further clarification of how the investigators conceptualized our topic of study. Wendy Duff presented a mini-lecture, addressing partners' questions, as well as assimilating ideas that had come out of the group sessions.

Duff began her lecture by differentiating user studies from evaluations. Archivists evaluate their performance, services, and systems. Many annual reports include some sort of evaluation. Alternately, a user study can be anything that involves inquiry into how users carry out their work. This, however, is not necessarily an evaluation. The two can be brought together, and Duff asserted that a focus of this grant has been to standardize the evaluation tools archivists use when studying the interaction between user and system.

Duff then spoke about the components—the where, when, how, what, and why—used in building and asking questions. “Where” addresses the collection under study, and can include a distinction between physical and virtual resources. “When” is the point at which one gathers the data, for instance, at registration. Duff related this component to the concern voiced Saturday

about working these studies into existing workflow; the possibility of starting with a registration module was also raised. Answering the “how” identifies method. One could perform a content analysis of transaction logs, or one could conduct a survey and do quantitative analysis. There is no single best method, only methods more appropriate for finding answers to certain questions. “What” is what you are evaluating, such as reference services, descriptive systems, educational and outreach programs, and collection use. Finally, why do you want to measure something? You could be looking for cost benefits, levels of satisfaction, effectiveness of services, or impact of research done in your archive. Each of these “whys” can be conceptualized in multiple ways.

Duff, speaking for the three investigators, envisioned building modules that incorporate the various component combinations. As an example, a module could be designed to evaluate physical reference, using the measures impact and satisfaction, and could be conducted as an exit survey interview.

Duff concluded her thoughts by laying out what she saw as the options for personal and institutional involvement. Individual partners could each volunteer, and they could get their institutions involved. Minimally, partners could offer to monitor developments merely by keeping abreast of progress. Another option would be participating in the design of modules or review modules that have been created. At the highest level of involvement, individual partners, within their professional networks and archives, could advocate the adoption of standardized metrics for evaluation and user study, as well as urge for broader individual and institutional involvement. Partners could also work as part of an education program.

Institutional involvement has an advocacy component, too. Equally important to this, though, is getting institutions to commit to using the instruments that come from these efforts. It is hoped that institutions will assist in the testing of instruments as they are developed. Finally, it

will be crucial for the sake of comparison and understanding that institutions volunteer to share aggregate data, within allowable limits, gathered during evaluations and studies.

A brief discussion followed Duff's presentation. Building off of Saturday's group sessions, conference participants decided on the need to work first on crafting registration modules for on-site and remote users. Issues relating to the unique needs of different types of institutions would be left for a later phase of development. The current focus was placed on the ability to compare data across archives.

The conference concluded with an invitation offered to all present to continue as members of the project; all participants accepted the option of continuing as project partners. In order to get their institutions involved, partners expressed the need for a letter of explanation with a mission statement and mention of time commitment; a listserv and tools site; a web site where technical memos and other documents will reside; and a set of principles. A statement of purpose would be drafted for use at national conference sessions and regional meetings. A page of explanation could appear in the SAA newsletter in order to begin a professional conversation. The possibility of a meeting at the August SAA in Boston was recommended, as was a session at the SAA 2005 meeting.