The Seamless Cyberinfrastructure: The Challenges of Studying Users of Mass Digitization and Institutional Repositories

Dawn Schmitz

April 2008

Council on Library and Information Resources

Contents

1 Introduction

In recent years, academic libraries have launched major initiatives to make their resources more easily available to users. But with this increasingly sophisticated infrastructure comes a user environment that is challenging for libraries to assess because it can often appear seamless from the user's perspective, making it difficult for users to report back on their experiences in a meaningful way. This creates the conundrum: How can we learn who is using these new resources and how well are they meeting users' needs?

While the goal of optimizing the end-user experience may not always be the most prominent driver of new library initiatives such as institutional repositories (IRs) and mass digitization, this report is based on the premise that the most successful projects are those that are most widely used. Following from this is the belief that understanding how resources are used, and by whom, will lead to more-sustainable initiatives that will earn a secure place among funding priorities. As Harley (forthcoming) writes with respect primarily to online educational resources:

> Even if we all agree that open content should be made freely available for the public good, some entity—be it federally, state, or privately funded—will ultimately need to pay for it. A demonstration of robust demand from a set of relevant constituents will be undoubtedly needed to justify such investment.

Keeping in mind this focus on understanding use as a pathway to sustainability, this report will discuss who is, or may be, using IRs and mass-digitized collections and what steps academic and research libraries can take to learn more about their use. The intent is to suggest strategies that libraries may use to enhance the long-term planning and design of these projects.

As will be discussed in more detail in the following section, IRs and mass-digitized materials are part of the *cyberinfrastructure*. Ribes and Finholt (2007) write that in planning cyberinfrastructure for the long term, *designing for use* is a key element that implies a concern about how to develop resources that will get used and to facilitate the work of research. "This concern," they note, "is rooted in an acknowledgment that an infrastructure without users is not an infrastructure at all" (231).

1.1 Scope of this Report

It may be somewhat unusual to consider mass digitization and IRs together, since they are two distinct types of initiatives from the point of view of library administration. They have different aims, distinct technologies, and separate implementations. However, users do not consider library administration, strategic plans, or the technology that is employed in making resources available when they search for information. Faculty, students, and other members of college and university communities use a wide variety of electronic resources in the course of their work and academic pursuits, and only a portion of these resources is available directly through their own academic library. The resources available through IRs and mass digitization fall on a spectrum that includes student products, working papers, images, formal pre- and postprints, electronic journals, e-books, and digitized monographs, as well as news sources, Wikipedia, Amazon.com, and other proprietary and nonproprietary information sources found on the Web.

Thus, mass digitization and IRs are considered together here because users commonly experience them as integrated components of *cyberinfrastructure,* defined as the information, data, technologies, expertise, best practices, standards, tools, retrieval systems, and institutions that make research possible in the digital age (Unsworth 2006, 6). In practical terms, cyberinfrastructure can be seen as an analog to the roads, bridges, power grids, telephone systems, and other structures that have been termed "infrastructure" since the 1920s. The newer term *cyberinfrastructure*, writes Atkins, "refers to infrastructure based upon distributed computer, information, and communication technology. If infrastructure is required for an industrial economy, then we could say that cyberinfrastructure is required for a knowledge economy" (2003, 5). The use of the term i*nfrastructure* captures the invisibility as well as the necessity of these enabling structures. "One characteristic of infrastructure is that it is deeply *embedded* in the way we do our work. When it works efficiently, it is invisible: we use it without really thinking about it" (Unsworth 2006, 6).

The fact that this report focuses on IRs and mass digitization does not mean that these are the only components of cyberinfrastructure worth examining through user studies. However, for the purposes of this report, the objects of study were narrowed to these two types of initiatives because, as compared with others, such as disciplinary repositories, IRs and mass-digitization projects have a less well-defined end user base and therefore potentially offer a greater number of user-related questions to investigate. This report reviews the literature relating to mass digitization and IRs as a first step toward learning who is using these resources. The study then considers which methodologies may be helpful to researchers who want to learn more about their use. An effort was made to review all literature published since 2003 that relates to the identification of user groups for these resources. User studies that focused on usability (i.e., interface design) or any other aspect of implementation were not reviewed. The literature review also included user studies that focus on use of the Web for academic research. Using the Web in this way prefigures the behavior associated with mass-digitized collections and IRs, and such studies may therefore provide insight about who uses or would use these two emerging resources.

1.2 Rationale for this Report

Studying the use of cyberinfrastructure is challenging for several reasons, not the least of which is the range of uses and users involved. Scholars are both producers and consumers of knowledge, and they can be publishers of information as well as users of it. For example, many scholars may use IRs primarily as a way to informally publish their own work and secondarily to access the work of others. With respect to mass digitization, libraries, publishers, and their partners make formally published information available, and scholars act primarily as end users searching for information. While scholars have written much of the literature that is digitized, they are not directly involved in its digitization and dissemination.

Perhaps a bigger challenge to studying use is the invisibility of the cyberinfrastructure to the end users. As an illustration of this invisibility, one need only consider how poorly its individual components are recognized by users. For example, a user who conducts research using an Internet search engine may be presented with metadata about a particular document or digitized book in her list of results, but she would not understand the intricate mechanisms by which this information has been made available to her. This situation is increasingly common, as libraries, striving to improve access, have worked with each other as well as with corporate and nonprofit entities to make metadata searchable with popular Internet search engines. These efforts have resulted in a user environment in which various components of the cyberinfrastructure interact in an apparently seamless way, such as when results of an Internet search engine connect a user directly to digitized books or to documents in IRs.

Many librarians point out that there are miles to go with regard to interoperability, open access, and aggregation of information in digital libraries, and that Internet search engines are often not the best tools to use to access the most relevant scholarly information. Indeed, to librarians' chagrin, some users consistently use Google as a search engine, regardless of whether it is the method that produces the best results.[1] Users are often confused by the myriad ways of finding information; in this situation, they will choose what is, from their perspective, the simplest mechanism. For students and scholars without access to good academic library collections, Internet search engines are a necessity.

The complexity of the cyberinfrastructure can make it difficult for users to identify the provenance of the materials they gather. Users conduct research without having to authenticate or navigate through a library Web site to find a particular vendor database, and they may not be able to report, if asked later, what resource other than Google they used to find information. If a user retrieves a paper from an IR or reads a book that was made available to her through a mass-digitization project, she may have all the information she needs to use the book or paper for scholarly purposes (i.e., author, date, publisher of the book, whether the paper has been peer reviewed), but she may not realize or remember exactly how the book or paper was made available to her. This lack of recall or understanding creates difficulty for librarians who wish to study who uses IRs or digitized books, since they cannot simply ask users what types of resources they have used. Methods that are available for studies of the use of specific proprietary databases, such as automatic statistics tracking, do not reveal

[1] This is not an argument either for giving up on information literacy instruction or for abandoning efforts to develop better ways of aggregating information than is provided by Google and other search engine companies. The point is to the contrary: understanding who uses these tools and how they do so will facilitate both endeavors.

whether users have accessed IRs or mass-digitized books as general categories. One of the aims of this report is to help librarians understand the unique challenges of studying users in this environment and to help them develop ways to assess the impact of their mass- digitization and IR projects.

2 Background on Mass Digitization

This section begins by presenting a working definition of *mass digitization,* a process that is admittedly often difficult to distinguish from other types of digitization. It then presents an overview of two key issues relating to mass digitization: selection and copyright. These are discussed together since public domain status (or ability to obtain permissions for in-copyright works) is sometimes a selection criterion. The section ends with a discussion of what is known about users of mass-digitized collections.

2.1 Definition of Mass Digitization

Libraries have engaged in digital conversion of large collections of print materials since the mid-1990s, and large-scale projects, notably the Million Book Project led by Carnegie-Mellon University, Zhejiang University in China, the Indian Institute of Science in India, and the Library at Alexandria in Egypt, have been initiated in this decade. However, digitization shifted into a higher gear in late 2004, when five libraries announced they were joining with Google to digitize their books on a mass scale.[2] By February 2008, the Google Book for Libraries project had expanded to include 18 library partners in the United States, Germany, Belgium, Japan, Spain, England, and Switzerland, in addition to the Committee on Institutional Cooperation (CIC).[3] The Open Content Alliance (OCA), established in 2005 as an alternative to Google Book, now has dozens of library partners, including consortia, as well as corporate partners including Microsoft and Yahoo. In addition, Microsoft launched its own search engine, Live Book Search, in 2006. Also in part as a response to Google's project, the European Digital Library (EDL) was formed. In 2006, the EDL announced a project to make mass-digitized works from European countries available via a "single, multi-lingual access point."[4]

Despite this flurry of activity, the term *mass digitization* still has no universally accepted definition. "Mass" digitization cannot be cleanly separated from "large-scale" digitization. Nonetheless, the two characteristics most commonly associated with mass digitization are (1) the relative lack of selectivity of

[2] The Google 5, as they are sometimes referred to, included Harvard University, the University of Michigan, Stanford University, the University of Oxford, and the New York Public Library.

[3] The inclusion of the 12-member CIC added 10 libraries. Google had already been working with two CIC libraries: the University of Wisconsin–Madison and the University of Michigan. See the Google Book Search Web site at http://books.google.com/googlebooks/partners.html.

[4] French national library (Bibliothèque nationale de France, or BNF) Web site: http://www.bnf.fr/pages/version_anglaise/europeana/europeana_eng.htm. Press release on the European Union Web site: http://europa.eu/rapid/pressReleasesAction.do?reference=IP/06/253&format=HTML&aged=1&language=EN&guiLanguage=en

materials as compared with to smaller-scale digitization projects, as discussed in Section 2.1.1), and (2) the high speed and high volume of the process in terms of both digital conversion and metadata creation, which are made possible through a high level of automation. Coyle (2006) offers this definition of mass digitization:

> Mass digitization is more than just a large-scale project. It is the conversion of materials on an industrial scale. That is, conversion of whole libraries without making a selection of individual materials. This is the opposite of the discrete digital collections that we see in online archives like the Library of Congress's Making of America, or the Online Archive of California. The goal of mass digitization is not to create collections but to digitize everything, or in this case, every book ever printed. To do this economically and with some speed, mass digitization is based on the efficient photographing of books, page-by-page, and subjecting those images to optical character recognition (OCR) software to produce searchable text. Human intervention is reduced to a minimum, so the OCR output is generally used without undergoing additional revision. Also, only limited structural markup, such as page numbers, tables of contents, and indices, are included because these cannot be detected automatically by the OCR software and therefore require human intervention in the scanning process (641).

The high-speed, highly automated, and efficient processes described here for scanning and metadata creation are employed by Google, OCA, and EDL, and they arguably stand as the hallmark that separates mass digitization from other digitization projects. The number of books included in these projects is indeed staggering. For example, the University of Michigan predicts its seven million volumes will take six years to digitize; at the rate they had been going before joining with Google, this feat would have taken 1,000 years.[5] Google indicates that it expects to digitize up to 10 million volumes with the CIC project. The OCA runs eight scanning centers in three countries; each center has 10 scanners working 16 hours per day and scanning 12,000 books per month (Albanese 2007).[6] EDL's goal is to digitize at least six million books over several years. OCA estimates the cost of mass digitization at about 10 cents per page, or $30 per book, and the French national library (a participant in EDL) estimated the cost at about €0.09 per page in 2006. These costs cover only scanning and optical character recognition (OCR); they do not include metadata creation, book retrieval, book repair, and reshelving, costs that are largely incurred directly by libraries, even those working with corporate partners.

Mass digitization is sometimes implicitly defined by the unique challenges that are brought to bear when large bodies of printed text are digitized. For example, research is under way to improve the quality of OCR, which can suffer when

[5] University of Michigan Web site: http://www.umich.edu/news/index.html?BG/google/index.

[6] See also video available on OCA's Web site: http://www.opencontentalliance.org/index.html.

large quantities of old books with faded, discolored, damaged, bled-through, or otherwise problematic paper and text are scanned very rapidly. "To create large digital libraries with reasonable costs and in a reasonable amount of time requires rapid scanning which can cause blurred, cropped, or skewed pages as well as missed or duplicated pages," write Feng and Manmatha (2006). Another major challenge is the creation of automated indexing methods that will bring the full benefits of large, full-text collections to readers (Mimno and McCallum 2007).

2.1.1 Selection and Copyright

Level of selectivity has also been offered as a way to distinguish mass digitization from other types of digitization projects. Coyle (2006) goes as far as to define mass digitization as the "conversion of whole libraries without making a selection of individual materials" (641). While this statement may be too strong, it may also be the case that earlier library digitization projects had a strong thematic element that conveyed coherence to the resulting digital collection, and therefore that a relative lack of selection of content may mark a digitization project as "mass." However, it is important to emphasize that selectivity varies among projects and participating libraries. Although Google has expressed its intention to digitize all the world's books and the University of Michigan went on record in 2004 as having a goal of digitizing all of the seven million books in its library, recent announcements of digitization projects indicate participating libraries have chosen to include a significant portion of their collections rather than every single book.[7] This implies that decisions are being made at some level about what to digitize and what to retain in analog form only. For example, attention to content and to reducing overall redundancies is evidenced in the recent announcement by the CIC of its digitization agreement with Google, which indicated that each of the 12 libraries in the consortium would focus on its own unique collections. These include, among others, Northwestern University's Africana collection, the University of Minnesota's Scandinavian collection, Indiana University's folklore collection, and the University of Illinois at Chicago's history and culture of Chicago collection.[8]

No discussion of selection would be complete without a mention of a contentious issue related to Google Book: copyright. A number of publishers and writers groups have sued Google for digitizing in-copyright works. Publishers charge the company with violating copyright law by forcing publishers to opt out of the project if they do not want their books digitized. Google counters that its

[7] In a November 2005 press release, Google stated, "Google is working directly with publishers through the Google Print Publisher Program and libraries through the Google Print Library Project to digitize the world's books." (The title of the library project has since been changed to Google Book for Libraries.) http://www.google.com/press/pressrel/print_publicdomain.html. See also University of Michigan press release at http://www.umich.edu/news/index.html?Releases/2004/Dec04/library/index.

[8] Press release available at: http://www.cic.uiuc.edu/programs/CenterForLibraryInitiatives/Archive/PressRelease/LibraryDigitization/index.shtml

approach is legal because it is primarily an indexing system that provides only snippets of in-copyright works and presents users who wish to access the entire book with links to booksellers and libraries. Some libraries partnering with Google have decided to digitize only books published before 1923, which are safely in the public domain. OCA focuses on books in the public domain, but also digitizes some books not in the public domain with permission of the copyright holder. Similarly, the EDL digitizes both public domain and copyrighted works with permissions.

Perhaps an even more sensitive issue than copyright in the library community is to whom the digitized content and metadata are made available. OCA was established as an open-access alternative to Google, which generally covers the entire cost of digital conversion as part of its agreement with libraries. In practice, however, it has followed a more proprietary approach to making its files and metadata available to other search services. Each agreement between a mass-digitization sponsor or partner (such as Google, Microsoft, or OCA) and a library or consortium has different stipulations with regard to how digitized content and metadata will be accessible to participating libraries/consortia and to commercial or noncommercial entities.

This paper makes no attempt to provide all the details necessary for a robust discussion of access issues. To understand the use (or potential use) of mass-digitized book content, it must suffice to explain that the mass-digitized content of a number of libraries is being aggregated by different services. Users can discover the availability of digitized books across many libraries by using a number of search services, including Google Book Search (the metadata for which are integrated into Google's general search), the Yahoo search engine, Microsoft's Live Book Search, and the Internet Archive search engine, although the latter restricts full-text searching to an individual book rather than all the books it holds. In addition, there are search sites that are (or will be) run by participating libraries/consortia. However, at present, there is no single search service–commercial, library, or nonprofit–that is able to aggregate *all* digitized book content.[9]

2.2 Anticipated Users of Mass Digitization

Libraries and sponsors that participate in mass-digitization projects consistently describe anticipated users for these collections in the widest possible terms. Although "students and scholars" at the participating institutions are often mentioned, project participants generally do not seem to envision long-term access as limited to members of the specific university communities. Most often, they use phrases such as "scholars and the public" or "scholars worldwide" to describe potential users. A comment by Andrew Herkovic, director of communications and development for Stanford Libraries, is typical: "This project

[9] Both the Open Library project and WorldCat are hoping to do this. Also, federated search systems that can search/aggregate search results from multiple sources and search engines are starting to address this issue; of course, this requires that the user knows about the existence of such services at their library.

marks a significant contribution toward our vision of a digitized library in order to provide scholars at Stanford—and beyond—with unprecedented access to scholarly information" (Carwile 2007).

Library spokespersons often state that they anticipate users to access content from across the globe. The head of library communications at the University of Wisconsin, Don Johnson, anticipated the university's digitized resources would be used by "scholars from across the world" (Clinton 2006). The idea that a university's collections would be made available to scholars not just in the United States but also around the world—particularly in developing countries—is often mentioned in the context of democratizing access to library collections. Participants in a symposium at the University of Michigan in 2006 reiterated the importance of using mass digitization to facilitate the work of scholars in developing countries (Fitzsimmons 2006). Although in many cases the term *scholars* is used, potential users have also been described in more general terms. For example, Jennifer Ward, senior communication analyst for the University of California, said, "One of the really great things about this project is it's just broadening accessibility to these resources. . . . Once they're online they're going to be accessible to anyone around the world" (Chen 2005).[10]

2.3 User Issues Related to Mass Digitization

As Coyle (2006) remarks, among the many questions that will need to be answered about mass digitization are "Whom does this digitized library serve?" and "How does it serve users?" Questions such as these are difficult to answer. Harley et al. (2006) note the difficulty of gathering specific data about users of online educational resources when site registration is not required (6-5). The extensive search undertaken for this report found no published studies that examine users of mass-digitized collections specifically. Several articles do discuss the broader issues surrounding access to such collections, and some authors compare and contrast the approaches taken by Google and OCA. While most authors are generally positive about mass-digitization projects and express an understanding of the benefits to libraries of allowing Google to take on the costs of digitally converting entire library collections, several others have praised OCA's open approach, contrasting it with the secrecy of Google's plans. These writers have tended to question the uncertain outcomes of arrangements under which libraries partner with for-profit businesses that have different goals for access, preservation, scholarship, and public service than do libraries. (Tennant 2005, Waters 2006, Johnson 2007). Google has also been criticized for taking advantage of the excellent reputation libraries and publishers have enjoyed for quality assurance in matters relating to book accessibility and readability, only to deliver a product, Google Book Search, that suffers from inadequate usability because it fails to respect, and reflect, the nature of books (Duguid 2007).

3 Background on Institutional Repositories

[10] These remarks should be considered with due caution. Since many of them are taken from press releases and news articles, they necessarily reflect the most ambitious goals possible.

This section discusses how various researchers and institutions define the purpose and content of IRs as well as the relationship of IRs to the open-access movement. The prevalence of IRs is outlined. A review of the literature relating to two levels of IR users—depositors of content and end users who use IRs to find information—lays the groundwork for a discussion of how users could be studied in more depth.

3.1 Definition of Institutional Repository

It may be even more difficult to define *institutional repository* than to define *mass digitization*, since each college or university that develops an IR has a different purpose and vision for the IR, and implements it in a different way. One commonly cited definition is the one offered by Lynch (2003):

> In my view, a university-based institutional repository is a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members. It is most essentially an organizational commitment to the stewardship of these digital materials, including long-term preservation where appropriate, as well as organization and access or distribution (328).

Crow (2002), in a report for the Association of Research Libraries' (ARL) Scholarly Publishing and Academic Resources Coalition (SPARC), departs from Lynch somewhat, providing a definition that emphasizes the IR's role in open access to scholarly literature, Crow defines the IR as "a digital archive of the intellectual product created by the faculty, research staff, and students of an institution and accessible to end users both within and outside of the institution, with few if any barriers to access." Crow suggests four qualities of an IR for SPARC's purposes: institutionally defined; scholarly; cumulative and perpetual; and open and interoperable (16).

Acknowledging that this definition reflects SPARC's focus on addressing issues in scholarly communication, Crow indicates that other types of content may be included depending upon how a given institution defines the goals of its IR. He mentions the potential overlap between the roles of university archives and IRs. Lynch and Lippincott (2005) similarly remark that a more comprehensive view of an IR would not focus solely on open access to scholarly literature. Instead, it would strive to document the intellectual and cultural life of the institution. Such an IR might include the deposit of "datasets, video, learning objects, software, and other materials."

McDowell (2007) supplies a definition that is similar to Crow's in its focus on scholarly output but that leaves the door open for the inclusion of other types of objects. In this definition, an IR must be "an institution-wide service" that is open to faculty in every department and "intended to collect, preserve, and provide access to, among other things, faculty scholarly output in multiple formats." Excluded from McDowell's study were "repositories of student work or digital libraries of archival-only materials" and "format-specific repositories meant only to collect one type of work such as learning objects, electronic theses and/or dissertations (ETDs), or images." In this view, an IR must include a range of object types and these must include the scholarly output of faculty.

While these authors differ over whether IRs should focus primarily on promoting open access to scholarly literature, they generally concur that IRs should be defined institutionally rather than, say, by scholarly discipline. Allard, Mack, and Feltner-Reichert (2005) offer a definition that emphasizes the institutional or consortial nature of many IRs: "a digital collection that captures and preserves the intellectual output of an institution whether it represents a single or multi-university community." (327). Therefore, while IRs are defined institutionally, a college or university need not act alone. It may create an IR based upon existing or new library or institutional consortia.

To put things most simply, the definition of IR remains in flux. Allard, Mack, and Feltner-Reichert (2005) write, "Among the 30 articles analyzed only three did not include at least a brief discussion about the definition of an IR. This suggests that the concept of an IR is generally regarded as not being common to the profession like other concepts, such as collection development." (330). In a similar vein, McDowell (2007) suggests that it is becoming increasingly difficult for institutions to agree on a definition for IRs, given that "increasingly varied contents, platforms, purposes and policies are also emerging." This lack of agreement is illustrated in a survey of academic librarians intended to measure the implementation of IRs in the United States (Lynch and Lippincott 2005). The authors note that respondents did not easily distinguish between IRs and digital library collections:

> The responses to our survey also underscore the confusing relationships at many institutions among digital libraries, digital research collections and collections of materials in institutional repositories, and the ways in which all of these relate to the scholarly communications process. A number of respondents identified materials being accessioned into the institutional repository that we would have thought of as digital library collections.

This disagreement about the differences between digital libraries and IRs underscores the lack of clear boundaries among elements of the cyberinfrastructure even among library professionals.

Just as there is a range of definitions for what constitutes an IR, there is a range of platforms upon which IRs can be built, including DSpace, Fedora, Digital Commons, and Eprints. However, from the perspective of end users, one of the most important attributes of IRs is what they share—metadata, which enable discovery of their resources and therefore potentially make them visible to a wider audience. The Open Archives Initiative (OAI) establishes a protocol for metadata interoperability that allows information about content in IRs to be harvested, aggregated, and shared. OAIster, a union catalog of digital resources that uses the OAI's protocol for metadata harvesting (OAI-PMH), allows users to search for information in IRs. In addition, users are able to discover some content in IRs through commercial search services such as Google Scholar.

3.2 Prevalence of Institutional Repositories

One reason for learning more about the uses of IRs is that they are becoming fairly common in many types of institutions of higher education. Almost half of such institutions in the United States that responded to a 2006 survey reported that they had implemented, or were in the planning stages of implementing, an institutional repository. (Not every planned project, however, results in a full implementation [Markey, Rieh, St. Jean, Kim, and Yakel 2007].) A study by McDowell (2007) found that in March 2006, 91 institutions in the United States had IRs. Several of these institutions were participating in consortial IRs. McDowell (2007) also found that the types of institutions adopting IRs may fit a different pattern than commonly believed. "Although in popular rhetoric IR implementation was occasionally still portrayed as a trend primarily at larger schools with more faculty and graduate scholars," it is not just large research universities that have IRs. Most schools that were found to have IRs, or to participate in consortial IRs, have populations of under 15,000; less than 16 percent had more than 30,000 students. An international survey of IRs published in 1995 (van Westrienen and Lynch 2005) revealed the existence of IRs in Australia, Canada, and 10 countries in Europe, with the percentage of universities in those countries having IRs ranging from 100 percent (Norway and Germany) to 5 percent (Finland).

3.3 User Issues Related to Institutional Repositories

There are three levels of users with respect to IRs: implementers, depositors, and end users. Aside from the challenge of creating a good interface for end users, user issues surrounding IRs have so far been most concerned with how to engage more faculty in using the IR for self-archiving, namely, depositing digital objects such as articles, preprints, and data.

3.3.1 Who is depositing content into IRs?

Thomas and McDonald (2007) found it difficult to determine the precise number of IR depositors, claiming that "responsibility for name consistency in most repositories seemed to rest with the depositors themselves." Thus the author "Edward Smith" might be the same person as "E. Smith," but in the absence of explicit name-authority protocols, the system would count the names as separate people. The authors found that "most repositories do a poor job of maintaining standard forms of names for contributing authors, so the same author may be listed under multiple name variants and treated as separate people," a problem that was exacerbated because of the high number of coauthored papers and the inability of the reporting functions for the software (they studied only those repositories using Southampton's E-Print software) to count the total number of multiauthor papers.

Student work seems to be better represented in IRs than that of faculty, and the difficulties of persuading faculty to deposit their content is of concern to many. McDowell (2007) determined that the median annual increase of objects in U.S. repositories from November 2005 to November 2006 was only about one per day, with student work accounting for the highest percentage of items (41.5%). Such work consisted largely of ETDs and similar work such as senior honors theses. While noting that problems of categorization often make it difficult to determine the types of content deposited, she found that about 37 percent is faculty scholarly output, with around 13 percent being peer-reviewed works (preprints, postprints, articles in e-journals, and e-books).

McDowell also found that depositors from just a few universities accounted for a large percentage of faculty work, thus skewing the findings. When five of the schools with the largest IRs were excluded, only 14 percent of the items in IRs had been deposited by faculty and only 7 percent were refereed works. She concludes, "This assessment of repository size, as measured by total item count, confirms other studies, both anecdotal and data-driven, that content recruitment continues to be difficult at U.S. academic institutions."

These findings are consistent with those of other studies of IRs in the United States. Lynch and Lippincott (2005) also noted that IRs contain a large amount of student-generated content such as ETDs, and concluded similarly that this is caused by the difficulty of persuading faculty to deposit their work:

> Because the outreach to faculty can be a slow, incremental, somewhat piecemeal process, some institutions begin populating their institutional repositories with the work of their students, rather than their faculty, as a quick means of acquiring a substantial body of a specific type of content. An electronic theses and dissertations (ETD) program is one such approach.

Davis and Connolly (2007) studied the rate at which different collections at Cornell's IR were increasing in size and the Internet provider (IP) addresses from which items were downloaded. They concluded that "there is little evidence to suggest that individual faculty are making significant contributions of regular scholarly output to the repository. Although the breakdown of submissions by IP address is not conclusive, it is echoed by the growth patterns exhibited by the majority of collections." They found this pattern generally repeated at other institutions using the DSpace platform.

A key issue is voluntary versus mandatory deposit. Sale (2006) attempted to determine faculty behavior in three institutions in Australia and the United Kingdom that have some type of mandatory deposit policy. They found that at institutions with mandatory-deposit policies, researchers deposit their work soon after publication rather than complying with the six-month embargo placed by publishers to expire. (The repository managers can ensure the documents aren't available for open access until legally permissible.) Sale (2007) advocates departmental mandate of deposits, contending that even if only a small percentage of departments cooperate, the result will still be greater than the 15 percent to 30 percent rates of deposit under voluntary policies. Lynch and Lippincott (2005) found that in the United States, participation in IRs is more likely to be voluntary, leading them to conclude that success depends on outreach to individual faculty members. "Those institutions that have made a concerted effort to understand their faculty needs and to reach out systematically to their faculty seem to have been more successful in attracting content for their repositories."

Davis and Connolly (2007) examined what motivates faculty to participate–or not to participate–in IRs. They interviewed 11 faculty from a range of departments at Cornell University and concluded that primary reasons for not using digital repositories were the learning curve required to learn any new technology; concerns about copyright to deposit published literature; concerns about what constitutes a "publication" and whether depositing a preprint might undercut the ability to get a work published in a journal; unwillingness to associate one's work with work of others that is perceived to be of lesser quality; fear of plagiarism; an unwillingness to release research results prior to formal publication; and a reluctance to release work that has not been through peer review and may include mistakes. Reasons given for failure to use Cornell's IR in particular included preference for the use of subject repositories; perceived lack of DSpace functionality; closer identification with one's discipline than one's institution as pertains to scholarly work; the perception that IRs are "islands"[11]; and the general lack of knowledge about cross-searching and shared metadata. In their international survey of IR implementation, van Westrienen and Lynch (2005) found similar reasons for faculty nonparticipation. The reasons these authors report were difficulty informing faculty of the value of IRs, confusion about copyright and plagiarism, impact factors and scholarly credit, the perception that materials in IRs were of low quality, and cumbersome deposit procedures.

Foster and Gibbons (2005) studied faculty work practices as part of an overall effort to better market the institution's IR and adapt it to faculty needs. They found the terminology they used to promote the IR had little meaning or interest for faculty, and they decided to stop using terms such as *institutional repository*, *open-source software*, and *metadata*. They then focused on concerns expressed by faculty, namely, making it easier to access their work through Google as well as the IR itself; keeping digital items preserved and safe from damage or loss; making it possible to give out links to their work rather than

[11] Mandatory deposit policies for various constituencies are being implemented, discussed, and revised at universities. Harvard University's Faculty of Arts and Sciences has implemented a mandatory deposit policy, although faculty members can choose to opt out upon written request. See posting in *Library Journal Academic Newswire*, 14 Feb. 2008, available at: http://www.libraryjournal.com/info/CA6532658.html?nid=2673#news1. Protests by graduate students in the University of Iowa's writing programs over a policy to automatically make dissertations and theses, including creative works, available online as open access documents (Foster 2008) was followed by an announcement by the provost rescinding the policy on March 17, 2008. See the provosts's statement at http://news-releases.uiowa.edu/2008/march/031708mfa.html

sending it by e-mail; maintaining ownership of their work and control over who sees it; and eliminating the need to maintain a server or do anything complicated. Their findings echo the view of Johnson (2002), who noted that academic faculty "publish for professional recognition and career advancement, as well as to contribute to scholarship in their discipline. Accommodating these faculty needs and perceptions—and demonstrating the relevance of an institutional repository in achieving them—must be central to content policies and implementation plans."

Finally, a 2004 online survey about scholarly communities conducted at Curtin University of Technology in Perth, Western Australia, stands out as relevant to the present discussion since it included a question that asked about IRs specifically (Genoni, Merrick, Willson 2006). In response to a question inquiring about their familiarity with the "concept of an electronic (open access) institutional repository," 36 percent of the respondents replied that they were familiar the concept and 64 percent indicated that they were not (740). (This university had established an IR a year earlier.) The survey asked what types of materials should be available in an IR, with the following results (more than one response to this question was invited): peer-reviewed published articles, 83.7 percent; preprints (not yet published articles/conference papers), 72.1 percent; teaching materials (e.g., lecture notes), 64 percent; and unpublished research material/data, 52.3 percent. The 246 respondents comprised academic staff and postgraduate students in roughly equal numbers. They were from all divisions of the university: humanities (39.6 %), sciences (34.1%), and social sciences (21.7 %). Twenty-nine respondents failed to report a disciplinary affiliation.

3.3.2 End Users of Institutional Repositories

In contrast to the number of studies that examine depositors of IR content, there are very few studies that shed light on the end users of IRs, i.e., individuals searching for and using IR content. One study conducted by Ithaka (2006), a nonprofit organization devoted to issues relating to information technology in higher education, asked faculty about institution-based digital repositories without distinguishing between those focused mainly on images or on rare/unique special collection and those focused on preprints and postprints. Most faculty surveyed were unsure of whether their institution had any such repository, and fewer than a third from institutions with IRs had contributed content to them. Still, faculty generally viewed the objective of institution-based repositories as primarily to preserve their own work: more than four-fifths looked to outside electronic resources, rather than to repositories of information, to support their own research and teaching.

A U.K. study by the Joint Information Systems Committee (JISC) (Zuccala et al. 2006) examined the use of several types of repositories, including the IR, at the University of Southampton (E-Prints SOTON). On the basis of Web inlink statistics found by tracking the number and placement of links to documents in the IR, the study authors concluded that the repository was situated in a primarily academic context on the Web. An associated user survey was sent by e-mail primarily to members of a JISC-related listserv and to people on the private e-mail lists of various repository managers for the purpose of gathering information for several types of digital repositories. The IR-specific survey results included the following: 69 percent of survey respondents used the E-Prints SOTON, and 43 percent had learned about the IR from a colleague or a friend. Approximately 10 percent or less learned of it from a journal, and only a negligible number found it using a search engine. Fifty-seven percent found the IR easy to use and 51 percent found the material in the IR usually relevant to their needs.

While these studies provide some information about end users of IRs, the JISC survey developers acknowledged the difficulty of reaching users and nonusers of repositories. This may be one of the main reasons there have been so few studies attempting to learn more about IR end users. McKay (2007) noted:

> This dearth of usage data means we do not know: whether typical IR users are local or from outside the hosting institution; whether they find the IR via the institutional homepage or via search engine referrals; we do not know what kind of information they look for and use; nor how they use the functionality offered by IRs.

Other information about users that is not known includes who they are, not just in terms of institutional affiliation but in terms of status (graduate/undergraduate students, high school students or teachers, university or college faculty or instructors, members of the general public); to what use, if any, they are putting the information retrieved; and how satisfied they are with the quality of the information.

Hagedorn (2003) reported on a user survey conducted at the University of Michigan in the course of developing the user interface for OAIster. The intention of the survey was to gather information about the expectations of users who report going online to find information. While it predated many U.S. IRs and was not intended specifically to measure interest in IRs, OAIster collects metadata from all OAI-PMH data providers with digital content and provides data that hold relevance to a discussion of potential IR users. The survey revealed that a 88.8 percent of respondents were interested in finding journal articles, 80.1 percent were interested in finding reference materials, 56.2 percent were interested in data, 48.1 percent in full-text books, 42.7 percent in gray literature, 25 in audio files, 53.2 percent in images, and 17.5 percent in movies.[12] These results suggest that, for this subset of users, the key role of IRs may be to provide access to surrogates for published journal literature. When the survey was conducted, the service had not been widely promoted to end users, resulting in a high number of responses from members of the digital library community but a relatively low level of responses from faculty, students, and researchers. Of the 538 respondents, 42.4 percent were librarians, library staff, or information professionals; 11.9 percent were graduate students; 9 percent were faculty

[12] Results of the survey are available at http://www.oaister.org/o/oaister/surveyreport.html

members; 5.7 percent were undergraduate students; and 4.8 percent were research scientists.

4 Studies Suggesting the Use of Mass Digitization and Institutional Repositories

In the absence of studies that specifically examine patterns of use of IRs and mass digitization, one must look more broadly at the types of search strategies commonly employed by various constituencies. Studies that look at how different members of academic communities use the Internet to search for information for their coursework, teaching, and research are useful in this regard, since individuals who conduct general Internet searches may eventually find their way to IRs and mass-digitized books. Although standard Internet search engines search only metadata in mass-digitized collections and not the full texts, the fact that Internet search engines do return results from these collections–whether or not users are specifically searching for this type of resource–makes the extent to which students and scholars are using search engines for information retrieval of interest.

While the Internet-usage studies discussed in this section lack the level of granularity needed to determine whether users are finding their way to mass-digitized books using search engines, one could assume that a certain percentage of Internet searches on academic research topics direct the user to a digitized book relevant to his or her research. For example, a Google search for "Henry Ward Beecher" may result in several first-page hits that link the user directly to digitized books by and about this historical figure in Google Book Search. Once linked to Google Book Search, the user can do further searching. This example illustrates how serendipitous discovery in Web-based research may lead users to digitized books.

The importance of Internet search services for resource discovery in IRs has begun to be documented through examination of referral URLs and download statistics. Organ (2006) found that of the 51.1 percent of referral URLs for full-text downloads to an IR that were known, 95.8 percent were from Google, with a slightly smaller percentage for cover-page downloads, owing, presumably, to the way in which Google ranks and displays results. Organ (2006) concluded that even though the results are based upon one particular IR operating on the Digital Commons platform, "the important role Google plays in the research and discovery process has become apparent . . . . the dominance of Google is most likely universal" for other repositories on other platforms. While Google may not continue to be the dominant search engine, the importance to IRs of making their content findable by Internet search engines is suggested not only by this study but also by many other studies that indicate a heavy reliance on search engines by researchers at all levels.

4.1 Undergraduate Use of Internet Search Engines

Given that the cyberinfrastructure is seamless to the point where users can find their way to both IRs and mass-digitized books by starting with general Internet searches, examining the prevalence of the use of Internet search engines by various constituencies can be a starting point for formulating research that examines more specifically the users of IRs and mass-digitized books. A great deal of literature has documented that among members of higher education communities, undergraduates rely on Internet search engines more than on any other source for their academic work. In its literature review of the library use and preferences of the millennial generation, a Pew Internet and American Life Project survey found that 73 percent of students use the Internet more than the library for information searching; only 9 percent use the library more than the Internet (Jones 2002).

Similarly, in a study using a controlled environment, Griffiths and Brophy (2005) found 45 percent of undergraduates used Google first when locating information to complete assigned tasks, followed by their library OPAC (10%), Yahoo (9%), Lycos (6%), and other sources (4% combined). A follow-up study revealed that 27 out of 38 participants (representing 34 subjects at Manchester Metropolitan University in the U.K..) chose Google or a combination of Google and Yahoo, leading the authors to conclude that "it is clear that the majority of participants use a search engine in the first instance" (545). Griffiths and Brophy note that this reported preference for starting with Internet search engines mirrors results of earlier studies examining the information-seeking behavior of this user group. They summarize students' reasons for preferring to use search engines over other sources:

> Search engines are liked for their familiarity and because they have provided successful results on previous occasions. Individual search engines were frequently described by students as "my personal favourite," and phrases such as "tried and tested," "my usual search engine," and "trusted" were frequently given by students when asked why they chose this source first.

The authors note that Google in particular garnered many positive remarks from participants. Reasons for its popularity include the perception that it is straightforward, simple, bright, and eye-catching, and it corrects misspellings. One student wrote, "I find the site very helpful. It seems to have whatever I want. I'm happy with it. It is simple but complete" (546).

4.2 Internet Use among Graduate Students and Faculty

Although few studies have endeavored to examine the role played by mass digitization and IRs in the research practices of scholars, reviewing general studies of how scholars conduct research can provide a basis for a consideration of how more-targeted studies can provide greater insight into scholarly uses of mass digitization and IRs. This subsection provides an overview of how scholars conduct research, then explores how scholars use the Web to find information, highlighting findings that point in the direction for further research into the use of mass digitization and IRs.

After performing an integrated analysis of four studies into research activities, Palmer (2005) found significant differences in how scholars in different fields conduct research: The process of inquiry for humanities scholars follows a long, meandering, and often unpredictable path that often involves physical travel, aided by digital resources such as online finding aids. For scientists, the process tends to be more routine and time-specific; it is directed toward solving a particular problem or testing an idea. Humanities scholars use general Internet searches alongside online indexes and databases, as well as print sources, to confirm and refresh their understanding of the research that has been done on a general topic. Scientists rely even more heavily on online sources for such confirmation searching, though they are usually concerned with a particular question or problem. Interdisciplinary scholars and scientists also conduct general Internet searches, along with bibliographic searches, with the goal of finding information outside their core area.

Palmer (2005) also found that in the humanities, research is conducted through interaction with primary and secondary texts, which serve as key sources of evidence. Personal collections are highly valued, as are digital or physical library collections and databases. Digital access to texts is considered increasingly valuable, although "local digital library collections tend to be perceived along with the Internet as one big digital blur of information" (1144). With respect to material accessed on the Internet, questions of provenance, surrounding collections and cognitive authority create an environment in which scholars are less informed than they are with material accessed in the library or in their own collections. Still, with respect to mass digitization it is notable that "in the humanities, there is a growing appreciation of the ability to interrogate the full text of large corpora, especially in literary, linguistic, and cultural fields of inquiry" (1144). Scientists value the ability to find and collect papers online, and many also value the ability to share and federate datasets, a finding that seems to relate directly to online repositories such as IRs.

Studies focusing on the extent to which faculty and graduate students use the Internet in general are important for understanding the use of particular Internet-based resources because they offer clues that may help direct further research that would target the use of IRs and mass digitization. Studies of Internet use among graduate students and faculty take different views into the topic: academic status (differences among undergraduates, graduate students, and faculty behaviors); type of resource (proprietary versus free); and discipline. In general, the research indicates that graduate students and faculty members are less likely than undergraduate students to use Internet search engines for their research but that they still rely on them to a significant extent.

George et al. (2006) studied use of the nonlibrary Internet among graduate students at Carnegie Mellon University. Interviews with 100 master's and Ph.D. students representing all colleges and departments revealed that 97 percent use the nonlibrary Internet; of these, 73 percent mentioned using Google. The breakdown of Google users by discipline shows computer science with the highest percentage (93%) and the humanities with the lowest (50%), with other fields falling somewhere in between: art and architecture (56%), business and policy (91%), engineering (85%), and the sciences (69%). The authors also found that half of all graduate students search for papers and articles online, noting that, "though varying widely across disciplines (35% in humanities to 64% in computer sciences), half of all graduate students (50%) use the Internet to search for online papers or articles: research papers, white papers, journal articles and/or working papers." These data are not specific enough to allow one to draw conclusions about which types of Web sites are being visited to find these materials, but they do suggest that further research into the extent to which graduate students are finding this information in IRs is warranted.

Barrett (2005) studied graduate students in the humanities and found that several study participants reported using Google, as well as methods more commonly associated with humanities research (e.g., chasing down citations and browsing shelves), to find information. Participants relied upon the OPAC, discipline-specific CD-ROMs, Internet search engines, and Web sites to find information, and all except one "strongly disagreed with the stereotype that humanists dislike information technology." Participants reported that they appreciated the efficiency of electronic databases and the convenience of remote access to full-text journals; nonetheless, "the most common complaint participants had about electronic information technology was the lack of available primary sources" (326). Several participants reported that they were required to travel in order to access primary sources, which included "contemporary journals, recordings, individual recollections, museum artifacts, original manuscripts, and books" in archives and special collections (327). Taken together, these data suggest that graduate students in the humanities are, or could be, heavy users of relevant mass-digitized content—given that they are avid users of information technology and have a strong interest in primary sources such as books.

Harley et al. (2006) used discussion groups and a combined online/paper survey to find out what kinds of online resources are used in undergraduate teaching in the humanities and social sciences. They found that faculty and instructors at universities, liberal arts colleges, and community colleges in California[13] used a wide range of online digital resources in their teaching. The most common resource categories were images or visual materials (75%), news or other media sources and archives (64%), portals that provide links or URLs relevant to particular disciplinary topics (63%), online reference sources (62%), digital film or video (62%), maps (53%,), online or digitized documents (50%), and audio materials (46%). Online curricular materials created by faculty at other institutions—those provided by such sites as MIT OpenCourseWare, World Lecture Hall, and Merlot—were used by 35 percent of respondents overall and by 43 percent of community college faculty. The most common method for

[13] A similar survey of a broader audience, college and university faculty in the humanities and social sciences in the United States and abroad, returned strikingly similar overall results to the California-based survey. However, the response rate from non–U.S. institutions limited the researchers' ability to examine the effects of national or cultural differences (Harley et al. 2006).

finding digital teaching materials was Google searches (81%), followed by respondents' personal collections (69%).[14]

Perhaps the most illuminating of Harley et al.'s (2006) findings were the factors that prevent the use of online resources for teaching. The most common reasons given for not using digital resources were their failure to substitute for current teaching approaches (75%), lack of time to use them (66%), and distraction from core teaching goals (47%). Among the obstacles cited to the use of digital resources were lack of classroom technology (53%), difficulty organizing distributed resources (45%), feeling overwhelmed by the quantity of resources (44%), and not having the time to verify the credibility of resources (43%). A large number of respondents indicated a need for assistance in a number of aspects of using digital resources, including help setting up a technical infrastructure (82%), creating Web sites (82%), digitizing (80%), learning how to use a learning management system (79%), importing resources (79%), gathering, organizing and maintaining digital materials (78%), and integrating resources into a learning management system (78%). Overall, these findings suggest that further research could be done to gauge interest in using mass-digitized collections and IRs for educational purposes and to determine whether more could be done to make such resources more useful for teaching.

In their study of the information-seeking behavior of agricultural and biological scientists, Kuruppu and Gruber (2006) used semistructured interviews and focus groups to gather qualitative data. The authors found that these scholars located information on Web sites resulting from Internet searches, with several participants reporting that they used Google. While the authors reported that the graduate students and faculty members often used library databases and indexes to find articles for their research, "increasingly, scholars contact authors and researchers directly (most often by e-mail) or *investigate institution Web sites* to gather more information about a research area, rather than using library services such as Interlibrary Loan" (613, emphasis added). This reference to "institution Web sites" could be an indication of the use of IRs.

A Web-based survey of faculty at the University of Idaho (UI) conducted by Jankowska (2004) had somewhat inconclusive results with regard to faculty use of the Internet for research purposes. The survey results indicated that faculty used information and communication technologies (ICT) for a variety of purposes. The most popular format of ICT used by faculty was e-mail and document exchange (86%). Electronic journals, books, texts, and forms were second (71%). Online library services offered by the library took third place (65%), followed by comprehensive Web sites (53%). The term *comprehensive Web*

[14] Harley (forthcoming) notes the significance of the number of faculty who use their personal collections in their teaching, and she comments on the inadequacy of resources available to them to manage these collections: "More than 70 percent of faculty said they maintain their own collections, although relatively few of them make their resources available to others on the Web. It was clear from our discussions and from comments on the surveys that many faculty want the ability to build their own collections, which are often composed of a variety of materials, including those that are copyright protected. How to manage this potpourri of resources and integrate them into teaching practice is the challenge."

sites was apparently not defined (55). In response to an open-ended question about how they conducted searches for literature and data, the authors noted "almost 26 percent of the UI college professors stated that they used the Web as their source for literature and data searching." A smaller percentage used commercial databases, refereed journals, books, and government publications (57). However, it is not clear what "the Web" means. Although many faculty members mentioned using the Internet search engines Ask Jeeves, Yahoo, Northern Light, and Google, responses to questions about the use of databases confirmed the researchers' belief that "some college professors did not recognize the difference between Web sites and commercial databases available via the Internet" (57). This illustrates the difficulty of using surveys to ascertain which electronic resources are used by particular constituencies.

An online survey of business faculty at Penn State found that respondents reported using resources available on the Web at no charge more than they used subscription databases (Dewald and Silvius 2005): "Web use for research was quite high; overall, 74 percent reported using the Web either most of the time or almost always, and only 2 percent reported almost never" (317). [15] Respondents were, however, not necessarily more satisfied with the Web than with databases. The authors note that "those who reported using subscription databases were less satisfied with databases' ease of use than the Web and equally satisfied with timeliness of the Web and databases. However, these respondents were more satisfied with the factors of accuracy, content, and format for databases than for the free Web" (325). Respondents evidently experienced no confusion regarding whether they were using subscription databases or the Web, probably because the survey instrument for this study included definitions for key terms. The *Web* was defined as "sites that are available to anyone searching or browsing the Web without paying any fees," while *library databases* was defined as "those subscription databases available through Penn State University Libraries *or* other libraries or businesses, including article databases and business information," and several examples were provided (316).

A Web-based questionnaire survey of information-use patterns of 97 British academics in computer and information sciences, business/management, and English literature found widespread use of Internet search engines for research, with no significant differences in use of search engines among disciplines: 89 percent of computer and information science (CIS) and of English academics, and 78 percent of business/management academics used them more than once a week (Gardiner, McMenemy, and Chowdhury, 2006, 347). Although the survey did not use the term *institutional repository*, it did question respondents as to their use of "higher education institution" (HEI) Web sites and found differences among disciplines with CIS academics using HEI Web sites more than twice as much as English academics do (51% and 24%, respectively) in a week.

[15] Before assuming that business faculty's high use of the free Web is an indicator that they may be using IRs or mass-digitized books, one should note that there may be more information of relevance to the field of business than many other fields on the Web. The authors of this study note that the U.S. government, as well as professional business organizations and associations, provide much useful information and data on Web sites, and that a number of books have been written to guide business researchers how to find information on the Web.

Business/management academics fell in the middle of this range with 32 percent of these respondents using them more than once a week (347). It is not known whether respondents were reporting usage of IRs when they reported using HEI Web sites. Nonetheless, the types of disciplinary differences suggested here are intriguing and may provide the basis for a hypothesis and further study about who are the most common end users of IRs.

This survey by Gardiner, McMenemy, and Chowdhury (2006) also questioned respondents about their use of "digital libraries." In hindsight, the researchers admitted this term was vague and returned questionable results (349). While the term *digital library* was specifically questioned—presumably because the results of this question flew in the face of anticipated responses, with more academics in English (41%) than in CIS (31%) or business/management (14%) reporting using digital libraries at least monthly— it is likely the case that this and other studies suffer from the use of many terms that seem to reflect researchers' assumptions about respondents' knowledge of library terminology.

4.3 Academic Uses of the Internet in Developing Countries

For many, one of the great appeals of information technologies generally and of mass digitization and IRs specifically is their potential to democratize access to information, particularly from the wealthy nations to developing countries. Unfortunately, little research has been done that specifically examines the extent to which scholars and students in developing countries use, or would benefit from access to, the products of mass digitization and IRs. Much of the research that has been published on the potential of open-access publishing (including that made possible by IRs) in developing countries has focused as much on scholars' ability to publish their own work as on their ability to access that of others (Kirsop and Chan 2005, Papin-Ramcharan and Dawe 2006). Similarly, with the preservation of indigenous languages and culture a key objective, discussions of library digitization in relation to developing countries tend to focus on the digitization of local materials rather than on the use of U.S. or European mass-digitized materials by scholars in developing countries (Witten et al. 2002, Mujoo-Munshi 2003, Jeevan 2004). Still, given the assumptions discussed in Section 2.2 of this report about the potential worldwide use of books mass-digitized in U.S. libraries, it would be worth attempting to learn more about how scholars and students in developing countries use, or don't use, IRs and mass-digitized materials.

Given the lack of data specifically relating to this topic, the most promising starting point is the extent to which students and scholars in developing countries use the Internet. In developing nations, use of the Web and Internet search services among students and faculty at academic institutions is now common, despite technological and infrastructure barriers that exist in many places. This significant reliance on the Internet for research, which is similar to patterns in developed countries, leaves open the possibility that scholars and students may be using the Internet to find mass-digitized books and materials in IRs. Many studies report that significant proportions of university undergraduate students in a range of countries use the Internet for academic studies, including 98.2 percent of students in computer science and information technology surveyed at the University of Malaya in Kuala Lumpur (Saad and Zainab, 2004); 89.1 percent of undergraduate students surveyed at the Federal University of Technology, Akure, Nigeria (Ojokoh and Asaolu, 2005); and almost half of undergraduates surveyed at Obafemi Awolowo University, Ile-Ife, Osun State, Nigeria (Olufemi Omotayo, 2006).

Studies documenting the use of the Internet for academic work among graduate students and faculty tell a similar story: 34.1 percent of academic staff and 39.7 percent of postgraduate students surveyed at the University of Ghana used the Internet specifically as a source of information for research (Badu and Markwei, 2005); 92.9 percent of academic staff surveyed at Delta State University in Abraka, Nigeria "have benefited from the use of the Internet through downloading of related information materials for research, etc." (Adogbeji and Toyo 2006, 7); 83.9 percent of faculty surveyed at Kuwait University had used Internet search engines; of these users, 89.5 percent said a main purpose was "to look for information for my research" (Al-Ansari, 2006, 798). A study conducted at Panjab University, Chandigarh, India (Mahajan, 2006) broke down Internet use by academics by discipline: 80 percent of researchers in the sciences surveyed used the Internet for three to four hours per week and 20 percent used it eight to ten hours per week. All these researchers "have a positive attitude toward the Internet and feel comfortable gaining information through it for academic and personal purposes," the authors note (2). While 70 percent of researchers in the social sciences shared this positive attitude and 85 percent of them used the Internet for two to three hours per week, only 20 percent of humanities researchers shared this positive attitude toward the Internet. The majority of these also used the Internet for two to three hours per week.

Although none of these studies indicates what kinds of information these scholars are finding, in virtually every study, researchers and students reported heavy use of Internet search engines. Given that mass-digitized books and material in IRs can increasingly be found through general Internet searches, it would seem beneficial to examine in more detail what specific kinds of resources these students and scholars are finding and using on the Internet

5 Possible Directions for Further Research

Because its seamlessness may make it difficult for users to recognize its discrete elements, the cyberinfrastructure provides unique challenges to designers of user studies. This section begins with a discussion of why effort and resources should be spent on studying users of IRs and mass digitization and then provides an overview of methodologies that could be employed to study users in this online environment.

5.1 Importance of Studying Users

Among the reasons given by Harley (2007) for studying users of digital collections in more depth is the need to "address questions of strategic planning and investments in digital resource provision and use." While the major costs of mass-digitization projects are sometimes borne by outside entities such as Google or Microsoft, and open-source software is available for building IRs, both mass digitization and IRs nevertheless require a significant investment for libraries. It can be difficult to determine these costs because in many cases libraries do not budget them separately.[16] At the very least, the investment includes staff time and energy that could be spent on other endeavors. But it can also involve a significant allocation of resources to manage the long-term preservation of digital content, both in IRs and in those cases when mass-digitized content is retained for local use.

To put into perspective the massive amount of storage space required for a mass- digitization project, Rieger (2008) gives the example of Cornell University, where the digitization of 100,000 volumes in a year now requires 60 terabytes of storage, which is 12 times the amount required for Cornell's previous digitization activities over a 15-year period. Rieger found that there are currently no metrics or methodologies for estimating the resources required for preserving digital content, but she references the 2006 Life Cycle Information for E-Literature (LIFE) study, a collaboration between University College London and the British Library, that attempted to estimate the total life cycle costs of digital asset preservation. The study found the cost of preserving a single e-monograph for 10 years to be £30 (about US $51), which includes costs for acquisition, ingest, basic metadata, access, storage, and preservation, but not for creation. Rieger cautions that the study was based on a file size of only 1.6 megabytes per e-book, while a digital book created by Cornell partnering with Microsoft requires 700 megabytes of storage space. These data underscore the substantial resource investment needed for the preservation-related aspects of a mass-digitization project.

Learning more about the use of IRs and mass-digitized content, particularly about who uses them and for what purposes—and who does not use them—would enable libraries to assess the overall cost-effectiveness of these projects. Unfortunately, only a small amount of research has been conducted into the use of these resources. While Harley's (2007) study focuses on digital resources used by faculty for teaching in the humanities and social sciences, her observations about users apply to all aspects of cyberinfrastructure: "We know very little . . . about how digital resources, such as those produced at research universities, are actually being used by the different tiers of higher education institutions both in the U.S. and abroad." She further notes that the library community must think strategically about how to study users in order to facilitate better project planning:

[16] Markey et al. (2007) asked planners and implementers of IRs about costs. While they overwhelmingly cited the library as the source of funding, many could not cite specific budget figures for IRs. A typical strategy was to absorb IR-funding costs into routine library operating expenses.

> Lack of a clear picture about users of these resources makes coordination of user studies (methods, findings, business models, strategic planning) across projects challenging. What is the overall value of "user" studies? How can we begin to assess overall user demand, and what analytic methods are useful for the various phases of decision-making (e.g., start-up, site design, dissemination, maintenance, scaling, new audiences)?

One place to begin this coordination of user studies is to determine what kinds of studies could accurately identify who currently uses mass-digitized resources and IRs, for what purposes, and with what results.

5.2 Methodologies for Studying Users

The key to designing user studies is to determine how best to gather meaningful data and to analyze it effectively. This section outlines some methodologies for studying electronic-resource user behaviors and preferences and discusses how these methods could be employed to better understand users of IRs and mass digitization. The choice of which methodology to use for a given research study must depend on the types of questions at hand. Are questions related to the attributes of users, for example, geographic location or academic status? Do questions revolve around how users are finding their way to the resource? The types of information they are accessing? How they are using the information? How valuable they find the resource? How well the use of the resource contributes to the formation and maintenance of online communities? In each case, a different methodology, or combination of methodologies, would bring the best results. A comprehensive discussion of research methodologies may be found in Covey (2002).

5.2.1 Interviews and Questionnaires

Many types of interviews and questionnaires are retrospective. They rely on the respondents' recall and articulation of their research practices. Although this type of survey methodology has been widely used by libraries in the past, it will most likely need to be modified as the information architecture for library research becomes increasingly sophisticated. With respect to mass digitization, a user may recall that he often accesses historical documents online, but would not necessarily remember or know whether he had benefited from a mass-digitization project. In a similar vein, a user can be linked directly from a search engine results page to a document in portable document format held in an IR, thereby bypassing the IR interface altogether and making it unlikely that he would know that he had used an IR (Organ 2006). Even when users do find their way directly to an IR, its user interface is not likely to include the term *institutional repository*, because each IR has its own name or acronym that reflects its parent institution(s) and its own purpose. This makes a user-survey question that uses the term *IR* or *institutional repository* fairly meaningless.

Students and faculty in higher education around the world commonly use search engines. They are the primary means by which users can discover and access

both mass-digitized materials and digital objects in IRs. It is therefore probable that a significant portion of users of these research resources are finding them through Internet search engines. Yet search engine use introduces unique challenges to studies of users of academic libraries that are compounded by the efforts libraries and information technology professionals make to create a user-friendly environment. For example, a 2005 study found that 24 percent of ARL libraries had included Google Scholar on their alphabetical lists of indexes and databases and that Google Scholar appeared on instructional guides or workshop information on the Web sites of 20 percent of these institutions (Mullen and Hartman 2006). It would not be surprising, then, if users of this resource at these libraries were unable to distinguish, in retrospect, whether they used a search engine available on the open Web or through the library.
This problem has not gone unnoticed. As Griffiths and Brophy (2005) note:

> Search engine usage is difficult to measure because search engines—and the Internet in general—are not controlled environments, such as a library home page or a specific information database. As such, it has been difficult to apply the traditional model of recall and precision used in evaluating information retrieval (IR) systems to Internet search engines (SEs). (540)

For their study of searching behavior on the Web, Griffiths and Brophy gave undergraduate students a set of 15 tasks related to academic information seeking and asked them fill out a questionnaire immediately after each task. The questionnaire asked them, among other things, where they had tried to find the information to complete the task. Even in this controlled environment, in which little time passed between the use of the resource and the reporting of its use, the researchers noted that "some students exhibited confusion regarding services, listing the library catalogue and the BBC as search engines they had used" (545). This suggests the source of the problem is not simply memory failure but that users do not place online information resources into the same categories as librarians do.

Given that users often have difficulty relating their information-seeking experiences to researchers, interviews may be more effective than questionnaires. Urquhart et al. (2003) argue that it is difficult for users to report their experiences on questionnaires. For example, in one study that used both questionnaires and interviews, there were variations in how respondents answered similar questions about the same information-seeking incident. Again, the seamless nature of the online information environment led to the confusion:

> Although it was the interviewers wanted (*sic*) to find out which services students had used, and the routes taken, it was apparent talking with some of the students that Internet use was seamless, in the sense that students did not differentiate between specific services and they found it difficult to identify what they had used.

Urquhart et al. suggest that the back-and-forth exchange between questioner and respondent that is possible with certain advanced-interviewing techniques can enable researchers to gather the data they are seeking. They conclude that from the perspective of the researchers, who are "concerned with detecting trends in the use and uptake of more formal library and information services, the blurring of boundaries of information provision on the Internet meant that there was often some 'unpicking' to be done when discussing what has happened in a search" (78).

Interviews are much more labor-intensive than are questionnaires. A much larger and potentially more representative sample of users can be gained with a questionnaire than with interviews. But if questionnaires are used, it would be to the researchers' advantage to consider the difficulty users have in recalling and identifying the online research resources they have used. To make survey questionnaires more meaningful, it is imperative to define terms such as *search engine* or *institutional repository.* In addition, it may be effective to include descriptions or visual cues to facilitate respondent recall. These cues could take the form of screen shots of the interface of the resource in question. While a user may not be able to say whether she accessed a digitized book through the Internet Archive, she may recognize and recall using it if she is reminded of its unique interface.

Point-of-use surveys, in which the user of a particular resource is presented with the opportunity to complete a questionnaire while using the resource, eliminate the recall and identification problem. One disadvantage of point-of-use surveys, as noted by Harley and Henke (2007), is that they may elicit poor response rates. Inaccurate sampling is a second disadvantage. Harley and Henke (2007) suggest that such surveys provide the most useful data about the use of online research resources when used in tandem with transaction log analysis. For this reason, they will be discussed in greater detail in the following section.

5.2.2 Transaction Log Analysis and Link Analysis

This section focuses on two techniques that employ quantitative measures to understand how Web-based information is used and interrelated.[17] Transaction log analysis (TLA) can be used to identify users and usage patterns on individual Web sites and repositories that make up the cyberinfrastructure. Link analysis

[17] These methods are often associated with Webometrics, "the quantitative study of Web-related phenomena." Webometrics is a subfield of informetrics, which has been defined as the "quantitative study of information production, storage, retrieval, dissemination, and utilization" (Wolfram 2000, 78) covering both scholarly and nonscholarly communication, and "based on the combination of advanced information retrieval, data and text mining, and quantitative studies of information flows" (Wormell 2000, 132). Webometrics was developed when "methods originally designed for bibliometric analysis of scientific journal article citation patterns" were applied to the Web. (Thelwall et al. 2005, 81). Bibliometrics predates the broader field of informatics but has been defined similarly as "the study of the quantitative aspects of the production, dissemination, and use of recorded information," (Tague-Sutcliffe 1992, 1). However, "traditionally, bibliometrics has dealt with the study of print-based literatures" (Wolfram 2000, 78).

can help paint the larger picture of how individual digitized documents relate to other information on the Web, and particularly of how users themselves have formed these relationships by creating these links.

5.2.2.1 Transaction Log Analysis

Transaction log analysis, or Web server log file analysis, can be employed to understand more about the use of many types of digital repositories (Zuccala et al. 2007). A range of data on end user search and access activities can be captured in standard Web logs:

> This includes data about the most heavily used collections, frequently accessed objects, number of items viewed during a typical session, activity by time and date, search strategies, referring URLs and search engines, geographic location of user-based IP addresses, and file downloads (Goddard 2007, 77).

In addition to providing insight into which items in the repository are used most often, TLA can reveal how users have discovered the items. An examination of referring pages can reveal the context in which links to items in the repository are provided—for example, whether it is a scholarly citation, a blog entry, or a Wikipedia page. If the referring page is a search engine, the search terms used can often be identified by parsing the URL of the search results page. Most search engines embed the user's search terms in the URL of the results page because this is one of two standard ways to store user-submitted Web queries, notes Zuccala (2007, 563).

Log file analysis may be more useful for revealing *how* users search for information than it is for learning *who* uses a particular resource and his or her attitudes toward it. Nevertheless, researchers can get information about users' IP addresses and Internet domains, which can provide a limited amount of information about users. For example, an examination of IP addresses may reveal the number of users that are using computers in the library, a dormitory, or perhaps even in a particular department of the university, as well as general geographic information.

The main benefit of TLA is that it provides actual data, recorded automatically, and does not rely on reports by users. Harley and Henke (2007) suggest that point-of-use surveys and transaction log analysis are complementary:

> Among their strengths, surveys can be used to develop a profile of the site's visitors and their attitudes, behavior, and motivations. In particular, sites often employ surveys to determine personal information about their users, to discover users' reasons and motivations for visiting the site, and to explore user satisfaction levels. Transaction log analysis (TLA), on the other hand, can describe the actual usage of the site, including the relative usage volume of different resources, the

> details of users' navigation paths, the referring pages that led users to the site, and the search terms used to locate or navigate the site.

The authors advise against using data from point-of-use surveys to make generalizations about a site’s users since response rates tend to be very low. At the very least, response rates should be checked against total traffic to the site during the survey period. Although both point-of-use surveys and TLAs are automated, Harley and Henke (2007) warn that they can be time-consuming. Still, combining the two approaches can produce a better picture of users than can either approach by itself.

Collecting download and other usage statistics from transaction logs is already a common practice of IR administrators, and initiatives are under way to facilitate the sharing and comparison of these data across repositories, such as the JISC project Interoperable Repository Statistics. Log file analysis can be used to dig even deeper; it can provide insight into how users found the site, what they were looking for, what they viewed and downloaded, and how long they spent on the site as well as general information about their physical location.

When combined with qualitative methods such as point-of-use surveys, this type of information can provide IR administrators with useful data on the user experience of IRs. However, care should be taken to ensure user privacy is not compromised when IP addresses are tracked along with usage patterns. Davis and Connolly (2007) maintained user privacy as they analyzed many types of descriptive IR-usage statistics by preserving general IP address data while expunging any specific addresses.

While log file analysis can also be used for in-house digitization projects using software such as CONTENTdm (Goddard 2007), this method would not be directly available to libraries partnering with commercial or nonprofit entities in mass-digitization projects, since Web server log files of the search engines operated by these outside entities would not be accessible without their permission and cooperation. Some researchers have gained permission from commercial search engine providers to gain access to their log file data in order to examine user behavior (Spink et al. 2001), but this cooperation is not necessarily forthcoming from the corporate partners involved in mass-digitization projects.[18] If researchers interested in studying the use of mass-digitized books gained this cooperation, not only statistics regarding numbers of downloads of each item (information that is readily available even to end users of OCA-digitized books on the Internet Archive site) but also data such as the general locations of users, how they were referred to the site, what they are searching for, and how long they spend on the site, could be analyzed. Again, safeguarding privacy is a serious concern with any user log file analysis, and steps should be taken to make sure all data are made anonymous.

[18] For example, Google does not provide usage data from Web server logs to its partner libraries. (E-mail to the author from a Google Book Search communications representative, 4 March 2008.

5.2.2.2 Link Analysis

Some have suggested that log file data be combined with link data to form a more robust analysis of user activity in relation to digital repositories (Zuccala et al. 2006, Zuccala et al. 2007). Link analysis draws on the analogy between links and citations and on the related assumption, drawn from citation analysis, that links to a work are a sign of appreciation for that work (Bar-Ilan 2005). Therefore, tracking links beyond what can be done with log file analysis can provide insight into the value readers place on a work. In other words, even if links aren't directly followed so they show up in log file analysis of referrer pages, the very fact that links have been created in such types of Web pages as blogs, personal Web pages, wikis, and bibliographies of scholarly papers is notable.

Commercial search engines now have applications that can be used to automatically extract URL data from Web pages, making it possible to return URLs of pages that link a Web site or repository (or any of the pages within it). The advanced- search help pages of search engines should be consulted to determine precisely what kinds of search queries are available. Software tools are available to facilitate the data- gathering process, retrieving link data from commercial search engines and calculating summary statistics about retrieved inlinks or URLs (Zuccala et. al. 2006). However, there are many limitations to the efficacy of using commercial search engines for link analysis that make it impossible to know whether results are complete. Queries are generally limited to about 1,000 URL hits. They use unreported methods to find and rank pages, thereby introducing biases into the results, and they do not index the entire Web (Thelwall 2008, Zuccala et al. 2006). Because of these limitations, no link analysis using search engines will return results that are 100 percent complete. However, Thelwall (2008) explains new methods for link analysis research that can maximize results from search engines, such as query splitting, domain searching, and top-level domain searching.

An analysis of links pointing to a repository Web site has been done as part of a JISC study of digital repositories (Zuccala et al. 2006). The JISC report recommended that link analysis be carried out every four to six months for the digital repositories in the JISC study. Link analysis can reveal new links to documents in an IR and can search for lists of links to other similar repositories, seen as "competitor" sites. By analyzing new link patterns, repository managers could gain insight into new or potential user groups. "If more links or different types of links are found to be directed to the site of a similar international resource, then perhaps these links represent previously unrecognized users, or areas for further outreach or cooperation," note Zuccala et al. (2006, 3).

Link analysis could be applied to mass-digitized book collections on certain platforms. For example, a query could be formulated to find inlinks to all pages that are in the domain www.books.google.com, which would return pages that contain links to books digitized by Google. All the limitations of using Internet search engines for link analysis research that were outlined above would also apply in this situation, and with a corpus as large as Google Books, obtaining the first 1,000 inlinked pages, without an understanding of how the link search results are ranked, would have limited utility. Still, a recent search using Yahoo! Site Explorer revealed that Yahoo! currently indexes more than two million Web pages that link to books digitized by Google or to the Google Books home page. Of the 1,000 pages returned, Wikipedia entries dominated. A rough estimate based on an unscientific sample of these results revealed that approximately 76 percent of these 1,000 inlinks were Wikipedia pages and a significant percentage of the remainder were blog entries. The data are downloadable to a spreadsheet, which makes it possible to sort and manipulate results.

5.2.3 Field Experiments

Data about general categories of users gathered from online surveys/transaction log analysis could form the basis for in-depth studies involving direct observations of particular user groups. Direct observation would make it possible to determine which (if any) of the targeted resources particular groups use and to ask them questions about their use of these resources. While in-person observation for long periods might be too labor-intensive, this approach could be automated. Kellar, Watters, and Shepherd (2007), in a study of Web-based information seeking, gained permission from research participants to install a specially designed Web browser on their laptops in order to capture their Web usage and Web browser interactions. Participants also recorded their Web usage in electronic diaries, using predetermined categories. A similar approach could be adapted to gather data about whether and how members of particular user groups employ certain Web-based resources and to gather data about their usage. Follow-up interviews could be conducted with participants who revealed, through the harvesting of Web browser histories and diaries, that they had used the resources in question.

5.2.4 Online Ethnography

The terms *online ethnography*, *netnography,* and *virtual ethnography* describe related research methodologies that adapt the qualitative, interpretive, and participatory methods of ethnographic research to the study of online communities and cultures. In cultural studies, online ethnography can be used to study the sociocultural implications of the Internet as a site of community formation and discourse (Hine 1998). In marketing, the methodology has been used to study consumer attitudes and preferences toward products and services as disparate as coffee and plastic surgery—topics that are not Internet related but that are nevertheless subjects of discussion in online communities (Kozinets 2002, Langer and Beckman 2005). Online ethnography can also be useful when the subject of inquiry is related to products and services offered on the Internet, such as students' experiences with distance-learning programs. Indeed, it has been argued that "when the field to be researched is virtual, conducting the interview online seems consistent with the actual practice of the participants" (Crichton and Kinash 2003). Online ethnography may be particularly well suited to understanding the experience of using the cyberinfrastructure.

Kozinets (2002) delineates a five-step process for conducting netnography for marketing research that could be adapted to learning more about the use of mass-digitized books and IRs:

1. *Make cultural entrée*: Identify an online community that can be studied and then locate one or more relevant online sites where this community meets to interact, such as a blog, an online bulletin board, or a listserv. Kozinets (2006) suggests blogs are particularly adaptable to the methodology but also cites networked game spaces, instant messaging chat windows, and mobile technologies as increasingly attractive places.
2. *Gather and analyze data*: This may include data directly copied from computer-mediated communication as well as observational data by the researcher. Given that so much data are automatically recorded (i.e., a blog entry or an interview conducted by e-mail or instant messaging), it may not be as necessary to use field notes as it would with traditional ethnography. However, by the same token, information overload may be a problem, so it is always necessary to sort data in order to track which pieces of information are most useful.
3. *Ensure trustworthy interpretation*: Keep in mind the limitations of the methodology. "Netnography is based primarily on the observation of textual discourse, an important difference from the balancing of discourse and observed behavior that occurs during in-person ethnography" (Kozinets 2002, 64). In other words, the data being gathered are not the same as those that would be gathered with traditional ethnography. What is being analyzed in online ethnography is the communication about the behaviors. Moreover, the implications of the fact that the communication is computer mediated (e.g., the possibility that identities may be altered in the virtual world) must be kept in mind.
4. *Conduct ethical research*: Kozinets (2002) offers these guidelines: full disclosure to online community members of the researcher's affiliations and intentions; assurance to informants of full confidentiality and anonymity; solicitation and incorporation of feedback from members of the online community in question; and treatment of the online forum as private, making it necessary to seek permission from members of the online community to use any postings that will be quoted in published research. However, no consensus has emerged over these ethical guidelines, given differing views on whether postings on the Web are more like public, mass-mediated forms of communication (analogous to a letter to the editor in a newspaper) or private communications. The question of privacy, in turn, calls into question what constitutes "informed consent" in the online environment (Haggerty 2004). Some researchers have argued that sites of online communication can be considered public if they are not restricted (i.e., password protected), and that, particularly with sensitive topics, it may be beneficial for the researcher to act covertly (Langer and Beckman 2005). Following this approach, some online ethnographers have taken the role of "participant observer" and chosen not to identify themselves as researchers in those cases when they

considered the communicative site open to the public and therefore not private (Sandlin 2007).

5. *Provide opportunities for community members to provide feedback:* As suggested in (4) above, Kozinets (2002) argues that members of the community being studied should be invited to provide *member checks*, "a procedure whereby some or all of a final research report's findings are presented to the people who have been studied in order to solicit their comments" (66). Beyond the ethical imperative, he provides several rationales for this recommendation, including the possibility of maintaining an ongoing relationship between community members and researcher.

Online ethnography can be more efficient than traditional ethnography because the researcher and the research participants do not have to meet in person. Moreover, it can be easier to maintain a record because of the ongoing textual record that can be captured; for example, no transcription of text-based online interviews is necessary. At the same time, online ethnography has many of the benefits of traditional ethnography, namely, the great extent to which it is user oriented, resulting in findings that are less influenced by the preconceived ideas of the researcher than are other qualitative methods. Kozinets (2006) says the following about netnography in relation to consumer market research:

> As a method, netnography is faster, simpler, and much less expensive than traditional ethnography. It can allow almost up-to-the minute assessments of consumers' collective pulse. Because it is unelicited, it is more naturalistic and unobtrusive than focus groups, surveys, or interviews. Unlike surveys, it does not force consumers to choose from predetermined researcher assumptions but provides a wealth of grassroots, bottom-up generated information on the symbolism, meanings, and consumption patterns of online consumer groups. It offers a powerful window into the naturally occurring reality of consumers" (281).

Along with the efficiencies of online ethnography come drawbacks, including the relatively impoverished nature of computer-mediated textual communication as compared with face-to-face interpersonal communication, which includes body language and other nonverbal cues. Moreover, online ethnography is open to the same criticisms of traditional ethnography, such as the implications of the lack of objectivity when the researcher is a participant in the culture being studied.

There are many ways of using online ethnography to study the use of the cyberinfrastructure. One way to start may be to identify blogs or listservs where scholars discuss and share ideas about conducting research and to monitor them to determine whether the participants share perspectives that would be relevant to the specific research question. Once an online community is identified, the researcher can choose whether to fully participate in the community or to observe and conduct a more covert textual analysis of the postings of the community members (Kozinets 2006, Langer and Beckman 2005). Separate from, but related to, this question is the main ethical concern: whether to disclose to the community members that research is being conducted and to obtain their permission to quote them in published research. If disclosure is chosen, interviews with individual community members can be arranged via e-mail or instant messaging. One of the main drawbacks of online ethnography is that, as a qualitative research method, it does not provide data on a large volume of users. To mitigate the effects of small sample size, it can be combined with data gathered automatically, such as clickstream patterns captured on Web server logs (Clark et al. 2006).

6 Summary and Conclusion

Users of mass-digitized collections and IRs do not, nor should they need to, understand the intricate mechanisms that have provided them with access to information on the Web. While this seamless cyberinfrastructure is a boon for users, it creates difficulties for libraries trying to understand those users and better serve their needs. The libraries that have participated in projects to mass-digitize their rich collections have imagined that students and scholars in their own communities and around the world would benefit from the content they are digitizing. Research indicates that, given the great reliance on Web search engines for academic research, it is possible that many users are finding their way to these resources. Similarly, users may be employing these same search engines to access content in IRs.

While research has provided insight into the practices and behaviors of potential and actual depositors of content into IRs, little is known about the users of IRs and mass digitization. Mass digitization and IRs fall on a single continuum of resources, yet they differ in many ways. Most notably, IRs provide scholars an opportunity to add to the body of recorded knowledge through publishing, while mass digitization makes a large existing corpus of printed literature available to scholars for use in their work. The challenge for those undertaking user studies is to understand the complexities of the user experience while still being clear about research goals, the type of users being studied, and type of resource or service under investigation.

Those designing user studies must think carefully about how to capture and analyze data that would shed light on user experiences. This report has cited some reasons why it is more difficult to study Web-based user behavior outside of the library environment than it is to study use of licensed library resources, which require authentication by users and which users might be more likely to identify discretely. Asking users to relate their research experiences retrospectively may not work as well as it has in the past. By keeping in mind that users experience the cyberinfrastructure as a relatively seamless web, it may be possible to design more-targeted surveys that aid in user recall of specific resources. There are many other ways to endeavor to learn about the user experience, such as attempting to capture research behaviors using automated methods while asking users to describe their experiences, analyzing how users have created a web of connections to digital resources, and identifying relevant online user communities and engaging with them virtually to learn more about their experiences.

As the cyberinfrastructure becomes more complex, new ways of learning about its use must be developed. This may not be easy. As Covey (2002) notes, "The methods for assessing new resource delivery evolve at a slower rate than do the resources themselves." Yet it is worth the effort to take up the challenge of devising these new methods. Greater attention to understanding the user experience will benefit academic libraries attempting to manage IRs and mass-digitization projects and will make them sustainable. If they are to produce more value for all their stakeholders, libraries must understand precisely how large-scale, resource-intensive cyberinfrastructure initiatives further institutional missions and user-service goals.

References

Adogbeji, Oghenevwogaga Benson, and Oghenevwogaga David Toyo. 2006. The Impact of the Internet on Research: The Experience of Delta State University, Nigeria. *Library Philosophy and Practice* 8(2). Available at http://www.webpages.uidaho.edu/~mbolin/adogbeji.htm.

Al-Ansari, Husain. 2006. Internet Use by the Faculty Members of Kuwait University. *Electronic Library* 24(6): 791-803.

Albanese, Andrew. 2007. Cornell Joins Google Scan Plan. *Library Journal* (Sept. 1). Available at http://www.libraryjournal.com/article/CA6471093.html.

Allard, Suzie, Thura R. Mack and Melanie Feltner-Reichert. 2005. The Librarian's Role in Institutional Repositories: A Content Analysis of the Literature. *Reference Services Review* 33(3): 325-336.

Atkins, Daniel E. 2003. *Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure*. Arlington, Va.: National Science Foundation. Available at http://www.nsf.gov/publications/pub_summ.jsp?ods_key=cise051203.

Badu, Edwin Ellis, and Evelyn D. Markwei. 2005. Internet Awareness and Use in the University of Ghana. *Information Development* 21(4): 260-268.

Bar-Ilan, Judit. 2005. What Do We Know about Links and Linking? A Framework for Studying Links in Academic Environments. *Information Processing and Management* 41(4): 973-986.

Barrett, Andy. 2005. The Information-Seeking Habits of Graduate Student Researchers in the Humanities. *Journal of Academic Librarianship* 31(4): 324-31.

Carwile, Laura. 2007. Google Project to Digitize Books Grows. *Stanford Daily*, January 25. Available at: http://daily.stanford.edu/article/2007/1/25/googleProjectToDigitizeBooksGrows.

Chen, Brian. 2005. UC to Digitize Out-of-Copyright American Works. *California Aggie*, October 5. Available at http://media.www.californiaaggie.com/media/storage/paper981/news/2005/10/05/FrontPage/Uc.To.Digitize.OutOfCopyright.American.Works-1321222.shtml.

Clark, Lillian, I-Hsien Ting, Chris Kimble, Peter Wright, and Daniel Kudenko. 2006. Combining Ethnographic and Clickstream Data to Identify User Web Browsing Strategies. *Information Research* 11(2). Available at http://informationr.net/ir/11-2/paper249.html

Clinton, Lexie. 2006. U. Wisconsin Joins Google Archive Project. *Daily Cardinal*, October 12. Accessed Lexis Nexis University Wire, 19 October 2007.

Covey, Denise Troll. 2002. *Usage and Usability Assessment: Library Practices and Concerns*. Washington, D.C.: Council on Library and Information Resources. Available at http://www.clir.org/pubs/reports/pub105/contents.html.

Coyle, Karen. 2006. Mass Digitization of Books. *The Journal of Academic Librarianship* 32(6): 641-645.

Crichton, Susan, and Shelley Kinash. 2003. Virtual Ethnography: Interactive Interviewing Online as Method. *Canadian Journal of Learning and Technology* 29(2). Available at http://www.cjlt.ca/content/vol29.2/cjlt29-2_art-5.html

Crow, Raym. The Case for Institutional Repositories: A SPARC Position Paper. *ARL Bimonthly Report* 223 (2002). Available at: http://www.arl.org/sparc/

Davis, Philip M., and Matthew J. L. Connolly. 2007. Institutional Repositories: Evaluating the Reasons for Non-use of Cornell University's Installation of DSpace. *D-Lib Magazine*. 13(3/4). Available at http://www.dlib.org/dlib/march07/davis/03davis.html.

Dewald, Nancy H., and Matthew A. Silvius. 2005. Business Faculty Research: Satisfaction with the Web versus Library Databases. *Portal* 5(3): 313-328.

Duguid, Paul. 2007. Inheritance and Loss? A Brief Survey of Google Books. *First Monday* 12(8). Available at http://www.firstmonday.org/issues/issue12_8/duguid/.

Feng, S., and Manmatha, R. 2006. A Hierarchical, HMM-based Automatic Evaluation of OCR Accuracy for a Digital Library of Books. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries.* Chapel Hill, N.C, June 11–15, 2006). New York, N.Y.: Association for Computing Machinery.http://doi.acm.org/10.1145/1141753.1141776.

Fitzsimmons, Beth C. 2006. *Mass Digitization: Implications for Information Policy.* U.S. National Commission on Libraries and Information Science. Washington, D.C.: NCLIS. Available at www.nclis.gov/digitization/MassDigitizationSymposium-Report.pdf.

Foster, Andrea L. 2008. U. of Iowa Writing Students Revolt Against a Plan They Say Would Give Away Their Work on the Web. *Chronicle of Higher Education*. 13 March. Available at: http://chronicle.com.proxy2.library.uiuc.edu/daily/2008/03/2029n.htm

Foster, Nancy Fried, and Susan Gibbons. 2005. Understanding Faculty to Improve Content Recruitment for Institutional Repositories. *D-Lib Magazine* 11(1). Available at http://www.dlib.org/dlib/january05/foster/01foster.html.

Gardiner, Donna, David McMenemy, and Gobinda Chowdhury. 2006. A Snapshot of Information Use Patterns of Academics in British Universities. *Online Information Review* 30(4): 341-359.

Genoni, Paul, Helen Merrick, and Michele A. Willson. 2006. Scholarly Communities, E-research Literacy and the Academic Librarian. *Electronic Library* 24(6): 734-746.

George, Carole, Alice Bright, Terry Hurlbert, Erika C. Linke, Gloriana St. Clair, and Joan Stein. 2006. Scholarly Use of Information: Graduate Students' Information-Seeking Behaviour. *Information Research* 11(4). Available at http://informationr.net/ir/11-4/paper272.html.

Goddard, Lisa. 2007. Getting to the Source: A Survey of Quantitative Data Sources Available to the Everday Librarian: Part II: Data sources from specific library applications. *Evidence-Based Library and Information Practice*. 2(1): 68-88.

Griffiths, Jillian R., and Peter Brophy. 2005. Student Searching Behavior and the Web: Use of Academic Resources and Google. *Library Trends* 53(4): 539-554.

Hagedorn, Kat. 2003. OAIster: A "No Dead Ends" OAI Service Provider. *Library Hi Tech* 21(2): 170-181.

Haggerty, Kevin D. 2004. Ethics Creep: Governing Social Science Research in the Name of Ethics. *Qualitative Sociology* 27(4): 391-414.

Harley, Diane. 2007. Why Study Users? An Environmental Scan of Use and Users of Digital Resources in Humanities and Social Sciences Undergraduate Education. *First Monday* 12(1). Available at http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1423/1341.

Harley, Diane, and Jonathan Henke. 2007. Toward an Effective Understanding of Website Users: Advantages and Pitfalls of Linking Transaction Log Analyses and Online Surveys. *D-Lib Magazine*. 13(3/4). Available at http://www.dlib.org/dlib/march07/harley/03harley.html.

Harley, Diane, Jonathan Henke, Shannon Lawrence, and Irene Perciali. 2006. Use and Users of Digital Resources: A Focus on Undergraduate Education in the Humanities and Social Sciences. Center for Studies in Higher Education, University of California, Berkeley. Available at http://digitalresourcestudy.berkeley.edu.

Harley, Diane. Forthcoming. Why Understanding Use and Users of Open Matters. In Toru Iiyoshi and M.S. Vija Kumar, eds. *Opening Up Education: The Collective Advancement of Education Through Open Technology, Open Content and Open Knowledge*. Cambridge, Mass.: MIT Press.

Hine, Christine. 1998. Virtual Ethnography. *Internet Research and Information for Social Scientists, International Conference Proceedings*. Proceedings of IRISS '98, Bristol, United Kingdom, March 25-27, 1998. Available at http://www.intute.ac.uk/socialsciences/archive/iriss/papers/paper16.htm.

Holliday, Wendy, and Oin Li. 2004. Understanding the Millennials: Updating Our Knowledge about Students. *Reference Services Review* 32(4): 356-365.

Ithaka. 2006. *Ithaka's 2006 Librarian and Faculty Studies: Overview of Key Findings.* Available at http://www.ithaka.org/research/faculty-studies

Jankowska, Maria Anna. 2004. Identifying University Professors Information Needs in the Challenging Environment of Information and Communication Technologies. *Journal of Academic Librarianship* 30(1): 51-66.

Jeevan, V.K.J. 2004. Digital Library Development: Identifying Sources of Content for Developing Countries with Special Reference to India. *International Information & Library Review*. 36(3): 185-197.

Johnson, Richard K. 2002. Institutional Repositories: Partnering with Faculty to Enhance Scholarly Communication. *D-Lib Magazine* 8(11). Available at http://www.dlib.org/dlib/november02/johnson/11johnson.html.

Johnson, Richard K. 2007. In Google's Broad Wake: Taking Responsibility for Shaping the Global Digital Library. *ARL: A Bimonthly Report* (250): 1-15.

Jones, Steve. 2002. The Internet Goes to College: How Students Are Living in the Future with Today's Technology. Pew Internet and American Life Project. September 15. Available at http://www.pewinternet.org/report_display.asp?r=71.

Kellar, Melanie, Carolyn Watters, and Michael Shepherd. 2007. A Field Study Characterizing Web-based Information-Seeking Tasks. *Journal of the American Society for Information Science and Technology* 58(7): 999-1018.

Kirsop, Barbara, and Leslie Chan. 2005. Transforming Access to Research Literature for Developing Countries. *Serials Review* 31(4): 246-255.

Kozinets, Robert V. 2002. The Field behind the Screen: Using Netnography for Marketing Research in Online Communities. *Journal of Marketing Research* 39(1): 61-72.

Kozinets, Robert V. 2006. Click to Connect: Netnography and Tribal Advertising. *Journal of Advertising Research* 46(3): 279-288.

Kuruppu, Pali U,. and Anne Marie Gruber. 2006. Understanding the Information Needs of Academic Scholars in Agricultural and Biological Sciences. *Journal of Academic Librarianship* 32(6): 609-623.

Langer, Roy, and Suzanne C. Beckman. 2005. Sensitive Research Topics: Netnography Revisited. *Qualitative Market Research* 8(2): 189-203.

Lynch, Clifford A. 2003. Institutional Repositories: Essential Infrastructure for Scholarship in the DigitalAage. *Portal: Libraries and the Academy* 3(2): 327-336.

Lynch, Clifford A., and Joan K. Lippincott. 2005. Institutional Repository Development in the United States as of Early 2005. *D-Lib Magazine* 11(9). Available at http://www.dlib.org/dlib/september05/lynch/09lynch.html.

Mahajan, Preeti. 2006. Internet Use by Researchers: A Study of Panjab University, Chandigarh. *Library Philosophy and Practice* 8(2): Available at http://www.webpages.uidaho.edu/~mbolin/mahajan2.htm.

Markey, Karen, Soo Young Rieh, Beth St. Jean, Jihyun Kim, and Elizabeth Yakel. 2007. *Census of Institutional Repositories in the United States: MIRACLE Project Research Findings*. Washington, D.C.: Council on Library and Information Resources. Available at http://www.clir.org/pubs/abstract/pub140abst.html.

McDowell, Cat S. 2007. Evaluating Institutional Repository Deployment in American Academe Since Early 2005: Repositories by the Numbers, part 2. *D-Lib Magazine* 13(9/10). Available at http://www.dlib.org/dlib/september07/mcdowell/09mcdowell.html

McKay, Dana. 2007. Institutional Repositories and Their 'Other' Users: Usability Beyond Authors. *Ariadne* 52. Available at http://www.ariadne.ac.uk/issue52/mckay/

Mimno, D., and A. McCallum. 2007. Organizing the OCA: Learning Faceted Subjects from a Library of Digital Books. Proceedings of the 2007 Conference on Digital Libraries, Vancouver, British Columbia, Canada, June 18 –23, 2007. New York, N.Y.: Association for Coumputing Machinery. Available at http://doi.acm.org/10.1145/1255175.1255249.

Mujoo-Munshi, Usha. 2003. Building Digital Resources: Creating Facilities at INSA. *International Information & Library Review* 35(2-4): 281-309.

Mullen, Laura Bowering, and Karen A. Hartman. 2006. Google Scholar and the Library Web Site: The Early Response by ARL Libraries. *College & Research Libraries* 67(2):106-122.

Ojokoh, B. A., and M. F. Asaolu. 2005. Studies on Internet Access and Usage by Students of the Federal University of Technology, Akure, Nigeria. *African Journal of Library, Archives & Information Science* 15(2): 149-153.

Olufemi Omotayo, Bukky. 2006. A Survey of Internet Access and Usage among Undergraduates in an African University. *International Information and Library Review* 38(4): 215-224.

Organ, Michael. 2006. Download Statistics: What Do they Tell Us? The Example of Research Online, the Open Access Institutional Repository at the University of Wollongong, Australia. *D-Lib Magazine* 12(10). Available at http://www.dlib.org/dlib/november06/organ/11organ.html.

Palmer, Carole L. 2005. Scholarly Work and the Shaping of Digital Access. *Journal of the American Society for Information Science and Technology.* 56(11): 1140-1153.

Papin-Ramcharan, Jennifer, and Richard A. Dawe. 2006. The Other Side of the Coin for Open Access Publishing—A Developing Country View. *Libri* 56: 16-27.

Rieger, Oya Y. 2008. *Preservation in the Age of Large-Scale Digitization*. Washington, D.C.: Council on Library and Information Resources.

Ribes, D., and Finholt, T. A. 2007. Tensions across the Scales: Planning Infrastructure for the Long-Term. Proceedings of the 2007 International ACM Conference on Supporting Group Work. Sanibel Island, Fla., November 4-7. New York, N.Y.: Association for Computing Machinery. Available at http://doi.acm.org/10.1145/1316624.1316659.

Saad, Mohd Sharif Mohd, and A. N. Zainab. 2004. Undergraduates in Computer Science and Information Technology Using the Internet as a Resource. *Malaysian Journal of Library & Information Science* 9(2): 1-16.

Sale, Arthur. 2006. The Acquisition of Open Access Research Articles. *First Monday* 11(10). Available at http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1409/1327.

Sale, Arthur. 2007. The Patchwork Mandate. *D-Lib Magazine* 13(1/2). Available at http://www.dlib.org/dlib/january07/sale/01sale.html.

Sandlin, Jennifer A. 2007. Netnography as a Consumer Education Research Tool. *International Journal of Consumer Studies* 31(3): 288-294.

Spink, Amanda, Dietmar Wolfram, B.J. Jansen, Tefko Saracevic. 2001. Searching the Web: The Public and Their Queries. *Journal of the American Society for Information Science and Technology* 52(3): 226-234.Tague-Sutcliffe, Jean. 1992. An Introduction to Informetrics. *Information Processing & Management* 28(1): 1-3.

Tennant, Roy. 2005. The Open Content Alliance. *Library Journal* 130(20): 38.

Thelwall, Mike, Liwen Vaughan, and Lennart Björneborn. 2005. Webometrics. *Annual Review of Information Science and Technology*. 39(1): 81-135.

Thelwall, Mike. 2008. Extracting Accurate and Complete Results from Search Engines: Case Study Windows Live. *Journal of the American Society for Information Science and Technology* 59(1): 38-50.

Thomas, Chuck, and Robert H. McDonald. 2007. Measuring and Comparing Participation Patterns in Digital Repositories: Repositories by the numbers, part 1. *D-Lib Magazine* 13(9/10). Available at http://www.dlib.org/dlib/september07/mcdonald/09mcdonald.html.

Unsworth, John. 2006. *Our Cultural Commonweath: The Report of the American Council of Learned Societies' Commission on Cyberinfrastructure for the Humanities and Social Sciences*. New York, N.Y.: American Council of Learned Societies. Available at http://www.acls.org/programs/Default.aspx?id=644&linkidentifier=id&itemid=644.

Urquhart, Christine, Ann Light, Rhian Thomas, Anne Barker, Alison Yeoman, Jan Cooper, Chris Armstrong, Roger Fenton, Ray Lonsdale, and Siân Spink. 2003. Critical Incident Technique and Explicitation Interviewing in Studies of Information Behavior. *Library and Information Science Research* 25 (1): 63-88.

van Westrienen, Gerard and Clifford A. Lynch. 2005. Academic Institutional Repositories: Deployment Status in 13 Nations as of Mid-2005. *D-Lib Magazine* 11(9). Available at http://www.dlib.org/dlib/september05/westrienen/09westrienen.html.

Waters, Donald J. 2006. Managing Digital Assets in Higher Education: An Overview of Strategic Issues. *ARL Bimonthly Report* (244): 1-10.

Witten, Ian H., Michel Loots, Maria F. Trujillo, and David Bainbridge. 2002. The Promise of Digital Libraries in Developing Countries. *The Electronic Library* 20(1): 7-13.

Wolfram, Dietmar. 2000. Applications of Informetrics to Information Retrieval Research. *Informing Science* 3(2): 77-82.

Wormell, Irene. 2000. Informetrics–A New Area of Quantitative Studies. *Education for Information* 18 (2/3): 131-138.

Zuccala, Alesia, Mike Thelwall, Charles Oppenheim, and Rajveen Dhiensa. 2007. Web Intelligence Analyses of Digital Libraries: A Case Study of the National Electronic Library for Health (NeLH). *Journal of Documentation* 63(4): 558-589.

Zuccala, Alesia, Mike Thelwall, Charles Oppenheim, and Rajveen Dhiensa. 2006. *Digital Repository Management Practices, User Needs and Potential Users: An Integrated Analysis*. Project Report. Joint Information Systems Committee. Available: cybermetrics.wlv.ac.uk/DigitalRepositories/FinalReport.pdf.