Panizzi, Lubetzky, and Google: How the Modern Web Environment is Reinventing the Theory of Cataloguing

Panizzi, Lubetzky et Google : comment l'environnement du Web moderne réinvente la théorie du catalogage

D. Grant Campbell
Faculty of Information and Media Studies
University of Western Ontario
gcampbel@uwo.ca

Karl V. Fast
Faculty of Information and Media Studies
University of Western Ontario
kfast@uwo.ca

Abstract: This paper uses cataloguing theory to interpret the partial results of an exploratory study of university students using Web search engines and Web-based OPACs. The participants expressed frustration with the OPAC; while they sensed that it was "organized," they were unable to exploit that organization and attributed their failure to the inadequacy of their own skills. In the Google searches, on the other hand, students were getting the support traditionally advocated in catalogue design. Google gave them starting points: resources that broadly addressed their requirements, enabling them to get a greater sense of the knowledge structure that would help them to increase their precision in subsequent searches. While current OPACs apparently fail to provide these starting points, the effectiveness of Google is consistent with the aims of cataloguing as expressed in the theories of Anthony Panizzi and Seymour Lubetzky.

Résumé : Cette étude utilise la théorie du catalogage pour interpréter les résultats partiels d'une étude exploratoire d'étudiants universitaires utilisant les moteurs de recherche Web et les catalogues publics en ligne. Les participants expriment leur frustration envers les catalogues publics en ligne. Bien qu'ils perçoivent que les catalogues sont « organisés », ils sont incapables d'utiliser cette organisation et attribuent leur échec au manque d'adaptation de leurs propres capacités. Lors des recherches avec Google, d'autre part, les étudiants ont reçu le soutien traditionnellement proposé dans la conception d'un catalogue. Google leur a donné des points de départ : ressources qui répondent largement à leurs besoins, leur permettant ainsi

© The Canadian Journal of Information and Library Science / La Revue canadienne des sciences de l'information et de bibliothéconomie 28, no. 3 2004

d'obtenir une meilleure compréhension de la structure des connaissances qui pourraient les aider à augmenter leur précision lors de recherches subséquentes. Alors que l'incapacité des catalogues publics en ligne à fournir ces points de départ est évidente, l'efficacité de Google, quant à elle, correspond parfaitement aux objectifs de catalogage exprimés par les théories de Anthony Panizzi et Seymour Lubetzky.

Introduction

As the World Wide Web moves into its second decade, the on-line public access catalogue has lost its old pre-eminence. We are now witnessing the rise of a new generation of information seekers: users who are coming to OPACs with significant experience with other electronic information tools: Web directories, Weblogs, discussion lists, resource guides, and search engines. This new generation of users is giving fresh urgency to a familiar question: What is the role of the on-line library catalogue in an information environment in which users are accustomed to using search engines? In particular, what new expectations are these young users bringing to the library catalogue and to what extent can and should these expectations be met and satisfied? Lurking beneath these questions is another, still more urgent question: Are we witnessing an evolution in information design and delivery, one that is different from, but continuous with, the information systems we have used in the past? Or are we witnessing a major disruption, a large-scale redefinition of information design and delivery so radically different from the traditional library environment that it renders irrelevant all our experience in bibliographic control?

The answers to these questions have far-reaching implications for libraries and their catalogues, particularly in academic settings, where scholarly research continues to rely on a relatively seamless access to various media and where there is continuing pressure to design portals and interfaces that enable users to search multiple information sources simultaneously. This paper confronts these questions within the context of a user study comparing student use of OPACs and Web search engines. While we found ample evidence of a disruption between the world of OPACs and the world of search engines, we also found surprising strands of continuity. Even more surprising, we found that these strands originate in the most traditional of cataloguing theories: those of Sir Anthony Panizzi and of Seymour Lubetzky.

Background

For most of their history, OPACs have been the most prevalent information retrieval systems available to the general public. As recently as 1996, it was stated that OPACs are "the most widely-available automated retrieval systems and the first that many people encounter" (Borgman 1996, 501). However, with the increased popularity of the Internet, we can no longer assume that OPACs are either the first or the most important retrieval systems that people use. And while some would argue that the problems of the Internet are problems that the library world quietly solved years ago (Bates 2002), the sheer volume of traffic on the popular search engines is staggering.[1] Web searching is shaping user expectations of what an information retrieval system looks like, how it behaves, and how to interact with it. An analysis of transaction logs for the Excite search engine found that "web search users seem to differ significantly from users of traditional IR systems" (Jansen, Spink, and Saracevic 2000, 226). The authors concluded that the design of Web-based information retrieval systems, search engines, and Web sites themselves should be approached "in a significantly different way than the design of IR systems as practised to date" (226).

In order to gain insights into these differences, we designed a study that attentively observed how university students perceive and interact with both OPACs and search engines and how they understand the differences between the two. Our findings were disturbing for libraries and their catalogue systems: Students expressed a distinct preference for search engines over library catalogues, finding the catalogue baffling and difficult to use effectively.[2]

Such results are hardly surprising: Adapting traditional systems for the Web environment is easier said than done, and library catalogues, encoded in MARC format, have already gone through the transition from physical cards to local on-line systems, usually available through Telnet. Furthermore, OPAC design is rooted in traditional information retrieval theory and practice, assuming the presence both of highly structured data and of expert searchers "who have a rich conceptual framework for information retrieval" (Borgman 1996, 501). Nonetheless, the fact remains: Descriptive cataloguing theory and practice both have an extensive set of concepts and procedures in place to help users find their way through complex information spaces. If these concepts and procedures are no longer working, we need to know why.

Cataloguing theory and practice

Library catalogues are based on a set of descriptive standards, established by international consensus through the International Standard Bibliographic Descriptions and manifested in cataloguing codes such as the *Anglo-American Cataloguing Rules*. These descriptions are in turn rendered machine-readable through the use of MARC, which enables bibliographic records to be transferred electronically from place to place and loaded into catalogues for searching and display. Equally important, these descriptive standards are based on a body of theoretical thinking and writing about library catalogues and their purposes, a body that effectively began with Sir Anthony Panizzi in the nineteenth century and has continued to the present day with the IFLA Working Group on the Functional Requirements of Bibliographic Records.

Much of this cataloguing theory makes an assumption that information searching often takes place in two stages: an initial stage, in which the user, in a state of relative ignorance, poses some initial, rudimentary queries of the system to establish a general understanding of the library's holdings in a particular area; and a second stage, in which the user draws on this new understanding to pose precise queries that attempt to satisfy the information demand as thoroughly as possible. Panizzi, defending his elaborate cataloguing rules in 1850, makes an important distinction between a work and an edition: "A reader may know the work he requires; he cannot be expected to know all the peculiarities of different editions; and this information he has a right to expect from the catalog" (1850, 348). The reader who wants a copy of *Crime and Punishment*, for instance, will first wish to ascertain whether the library contains the work and will then use the catalogue to establish more precise criteria for the selection of an edition: the translator, the presence of notes, and other such details.

Seymour Lubetzky (1969) uses this distinction between works and editions to clarify the purpose of the catalogue as a bibliographic tool, as opposed to either an inventory list or a full-blown reference tool. A library catalogue, particularly one that scrupulously follows the rules of main entry, functions as far more than an inventory list: It guides the user into a coherent bibliographic structure, populated by works and editions, all of which are linked by intricate relationships:

> If, as Butler maintained and as has been increasingly recognized, the function of the library is to provide for its users not only the materials needed by them but also the "bibliographical" guidance they require to help them make optimum use of the

> materials, then the catalog will have to be made to tell an inquirer in search of a book not only whether the library has that book but also what other editions and translations of the work the library has …
>
> … But there is yet another "bibliographical" relation of both direct and indirect interest to many catalog users: it is the interrelation between the works of an author. To show what works the library has of a particular author is of direct interest to many users concerned, not with any particular book or work, but rather with a particular author who may be represented by his works in the library. (Lubetzky 1969, 271)

As any cataloguer knows through experience, the creation of a bibliographic tool such as Lubetzky describes takes time, experience, training, and money. And as any user of Google knows through experience, Web search engines, with their spiders, indexes, and search and ranking algorithms, do not come from this theoretical background. While Web retrieval systems frequently draw on concepts and practices from the library community, in the form of metadata standards and controlled vocabularies, they are not designed to be bibliographic tools in the traditional sense.

We therefore went back to the data we gathered for our study, to ask a fresh set of questions:

1. Do the theoretical principles and practices of traditional bibliographic description have any relevance to the needs and behaviour of university students who know how to use search engines?
2. Does conventional OPAC design provide an adequate vehicle for those principles and practices?

Methods

A qualitative study was developed to observe and compare students searching the Web and a Web-based OPAC (Fast and Campbell forthcoming). The study was based on observation, retrospective verbal reports (think afters), and interviews. Data was collected from a questionnaire, video captures of the search sessions, and audiotapes of the think afters and interviews. Students performed the searches, then watched a recording of their searches for the think after, and finally participated in a follow-up interview. Sixteen university students from the University of Western Ontario (London, Ontario) participated in the study. Eight of the students were undergraduates taking a first-year course on information retrieval in the Media, Information, and Technoculture program (MIT): a course that draws students from a variety of undergraduate programs.

None of them was studying to become an information professional. The other eight participants were Masters students in the Library and Information Science program (MLIS). In this report, the undergraduate students will be referred as the MIT participants, and the graduate students as the MLIS participants. The research was conducted in three phases, between January and April 2002.

Participants were asked to imagine they were taking an introductory university course on economics and were starting research for a paper about "globalization and how it has affected the Canadian economy." Half the participants searched the OPAC first and then Google, the other half searched Google and then the OPAC. The University of Western Ontario OPAC is a Web-based OPAC using INNOPAC software.[3] It supports common catalogue search features: title, author, call number, subject heading (both LCSH and Medical), and keyword. The search sessions were recorded using Camtasia v3.01.[4]

After both searches were completed, the video recording was played back to participants. They were asked to describe "not what they were doing, but what they were thinking." Their comments were recorded and transcribed for later analysis. An interview was conducted immediately after the think after concluded. The interview typically lasted between 10 and 20 minutes. As with the think-after protocol, the interviews were recorded, transcribed, and analysed.

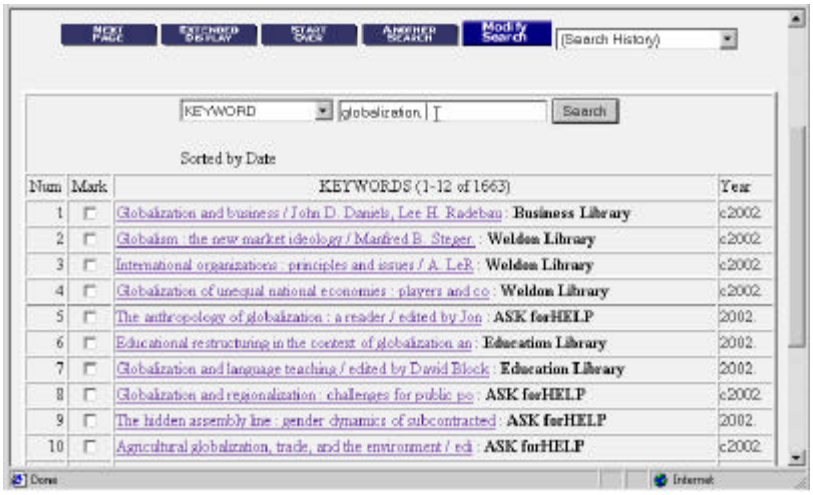
Findings

Three sets of findings are of particular relevance to this inquiry. First, the screen recordings showed that when searching the library catalogue, the participants tended to hover over the hit list. Whether the participants began with a keyword or a controlled vocabulary search, they tended to pause anxiously at the list of titles or vocabulary terms that the terms produced (See Figure 1). By contrast, when confronted with the Google hit list, the participants scrutinized the textual excerpts and were far bolder about trying items on the list.

Second, while the participants generally preferred Google to the library catalogue, they expressed respect and admiration for the very qualities of organization that the catalogue has historically represented and condemned the Web for its extraneous information. To one student, the OPAC had materials that were "clean"; it tended not to have a lot of "gar-

bage" or "things that didn't belong there." Another student said the catalogue "doesn't have any of the extraneous … articles, and things that you don't really need." The Web is cluttered; the catalogue is organized. However, this organization was not always helpful: it was admired, but not understood. One student began by admiring the organization of the OPAC and then immediately described it as confusing: "What do I like? It's very organized, I think.… The thing I don't like about it is that it's very confusing because of all these numbers and all these, yeah, it's just a little bit harder to [interpret]."

Finally, while the participants were generally happy with their understanding of search engines, they frequently expressed a low opinion of their ability to search the catalogue. Such an ability, they predicted, would emerge with greater experience: An undergraduate expressed the need to "spend more time on it and do more," while an MLIS student sheepishly suggested, "I'm experienced at searching [OPACs] but I may not be really good at it."



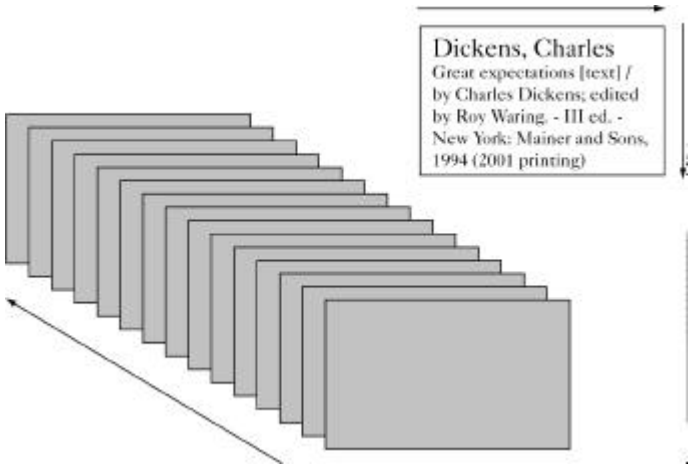
Discussion

Does the OPAC do justice to cataloguing theory and practice?

The findings from this small exploratory study would suggest that they do not. Like many other OPACs, the INNOPAC system presents a hit list as a list of hyperlinks, and this list is sufficiently similar to the Google hit list

to confuse young users, without any tangible benefit. The think-after data suggested that the users were trying to use the OPAC hit list as they would the Google hit list but were unable to derive enough information from the lists of titles or headings to gain any extra understanding.

What is more, the two-dimensional display on the OPAC robs the catalogue of the third dimension available in the card catalogue. As Figure 2 illustrates, the order of headings that characterizes the gathering effect of a bibliographic tool emerges along the length of the card row, as distinct from the two-dimensional space of the cards themselves. This has the advantage of making a bibliographic record available within the visible context of an ordering principle that works along different visual lines.



Does cataloguing theory have relevance to search engine design?

On the face of it, the answer is a resounding "No." Catalogues are founded on complex principles of control over details: procedures that prevent potential ambiguities that could dilute recall and precision. The detailed records, with their intricate MARC coding and their use of extensive and elaborate authority files, reflect a deliberate, human imposition of structure that is simply impossible on a wide scale and that cannot be automated with any degree of effectiveness. The World Wide Web, on the other hand, was founded upon a rejection of such centralized control: Its primary inventor, Tim Berners-Lee, was committed to a decentralized system of information generation that allowed patterns to emerge through wide access and few restrictions: "I would have to create a system with

common rules that would be acceptable to everyone. This meant as close as possible to no rules at all" (Berners-Lee and Fischetti 2000, 15). This absence of rules, while perhaps causing chaos on a small scale, would cause coherence to emerge on a large scale:

> The system had to have one other fundamental property: it had to be completely decentralized. That would be the only way a new person somewhere could start to use it without asking for access from anyone else. And that would be the only way the system could scale, so that as more people used it, it wouldn't get bogged down.… I wanted the act of adding a new link to be trivial: if it was, then a web of links could spread evenly across the globe. (Berners-Lee and Fischetti 2000, 16)

Search engines, particularly Google, with its popularity-based rankings, are founded on principles of emerging order rather than imposed order: Given a sufficiently large information space, with a sufficiently large and diverse number of users, patterns can emerge through ranking and hyperlinking systems that can be of major use for users looking for information. In this vision of information, user feedback causes the information space to evolve: Connections through hyperlinks encourage the formation of other connections and the ordering and ranking of items shifts constantly.

Given this divide between the underlying principles of the two tools, the admiration expressed by the users for the OPAC must be taken with a grain of salt. Their expressions of respect for the catalogue's "order" and "cleanliness" may well have been a sentimental indulgence towards a relic that has outlived its usefulness but not its power to charm. Or, they may have been expressing what they thought the researchers in Library and Information Science wanted to hear.

This gloomy outlook, however, proceeds from an assumption that both Google and the library catalogue are "information systems" in the classical sense: what Buckland describes as "retrieval-based information services" in which "collected and stored information-as-thing is sought and retrieved by the user" (1991, 30). It is worth remembering, though, that Buckland also identified other types of information system; in particular, "communication, in which information is conveyed, intentionally and more or less directly, to the receiver, as in a conversation, a letter, or a lecture" (30).

Communication as an information act has gained greater prominence with the rise of new Web-based forms of human interaction, particularly developments in social software. Weinberger (2002) goes so far as to suggest

that the Web puts meat on the bones of our "anorexic" traditional views of information by enhancing "authoritative" information with communication that provides rhetoric, emotion, speculation, and, above all, human connectedness: "The Web is a new social, public space. But because the Web has no geography, no surface, no container of space that preexists its habitation, we can't make the old mistake about what constitutes our sociality. The Web is a shared place that we choose to build, extend and inhabit. We form groups there because our interests aren't unique" (119).

In a similar spirit, John Seeley Brown and Paul Duguid (2000) use Jerome Bruner's distinction between "learning about" and "learning to be" to suggest that the most profound stage of knowledge acquisition lies at the point where the human being stops saying, "I'm learning to do something," and starts saying, "I'm learning to be something": "Learning a practice … involves becoming a member of a 'community of practice' and thereby understanding its work and its talk from the inside. Learning, from this point of view, is not simply a matter of acquiring information; it requires developing the disposition, demeanor, and outlook of its practitioners" (126).

If we treat both the catalogue and the search engine as vehicles and expressions of human communities, the data collected in this study starts to take a different shape. The participants were remarkably aware that, in entering the library catalogue, they had entered a place with rules, and rather tyrannical rules at that. There were things they had to remember, and things they had to know. Many participants appeared to sense the presence of a community of supposedly expert catalogue users, a community that they were as yet unable, or unwilling, to join. The MLIS students tended to use Google as a "starting point" before moving on to tools of the trade, such as the OPAC and on-line databases. The undergraduate students were more varied in their responses. Some saw the skills of catalogue searching as a representation of a community they wished to join:

> I thought that I would probably find more. But at the same time, because I hadn't had as much experience using the library catalogue I figured that I'd probably need to spend more time doing that. But at the same time I think the library catalogue is good because since I'm here and if I'm doing research at school it's probably, you know, it's going to be. I don't mind going to the library. I work well in the library.

Still others expressed ambivalence or outright hostility, as if their confusion about the catalogue's details were creating a permanent sense of having broken rules:

> I'm getting used to this computer and, I know how to use it, I know how to get information from the Net. If I go right now to the library I don't know how to get information from the library. Like I'm not going to be, it's not going to be the same as before because I, before I know, I knew how to do that. I knew what words should I do, what parts I should be in, and all that stuff. Right now, I don't think I can, I can get the best information from the library.

Google, by contrast, presented few barriers, and the participants felt little shyness about experimenting with it; indeed, some started from "Advanced Search" simply because they felt like trying it.

What, then, does the library catalogue have to do with community? For the undergraduate participants, it represents a community of university researchers that they are trying to join, often with great difficulty. For MLIS students, it can represent a set of skills they will have mastered by the end of their program. But if we analyse Lubetzky's principles, we find that community is at the heart of conventional cataloguing in two different but equally important ways.

First, Lubetzky's impassioned defence of main entry rests upon a vision of a catalogue in which entries are grouped and ordered around authors as the defining entities. The catalogue, for Lubetzky, provides representations of authors, just as it provides representations of documents, editions, and works. These "authors," as represented by personal—and corporate—author authority records, form an essential part of the catalogue as a bibliographic tool..

Second, the emphasis that Lubetzky places on the edition is first and foremost a device for fostering community. In modern library catalogues, the edition is the fundamental entity that triggers the creation of new bibliographic records: It is that collection of all copies of an item that come from a single master copy and can therefore be considered identical. Elizabeth Eisenstein (1983) suggests that the printing press revolutionized learning in Europe primarily by the creation of multiple identical copies of texts, and these identical copies enabled fruitful collaboration and knowledge building across time and distance that had hitherto been almost impossible: "[Early printed editions] were sufficiently uniform for scholars in different regions to correspond with each other about the same citation" (51). The user of a library catalogue, therefore, is using a tool that has been designed for intellectual collaboration: By identifying identical texts and differentiating different texts, it makes meaningful interchange possible.

Classical cataloguing theory, then, assumes that the bibliographic universe is a highly populated one, which is structured around the idea of mutual intellectual nourishment. In coming to the catalogue, the user is entering a community of authors, both personal and corporate, whose works appear in the collection. And these works are grouped into editions of identical copies, thereby facilitating communities of readers, learners, and sharers.

The Google ranking algorithm, by incorporating user feedback in the form of linking popularity, fosters a sense of community in a very different way. By ranking highly those pages that are heavily linked to other pages and the relative popularity of those linking pages, Google guides the user towards those resources that are highly travelled and implicitly highly regarded. Users hungering for a "ball-park notion" of the major concerns in a topic can therefore use Google to gain access to the most widely-used treatment of specific topics and themes, thereby identifying what others consider to be the most important concepts and issues. The anxiety surrounding the initial stages of the search process is mitigated by joining a community: using Google to go where there are others.

Conclusion

The evidence from this study, therefore, indicates that while OPACs may be unpopular among a generation raised on the Web, their unpopularity may not be insurmountable. Further and more innovative development in OPAC interface design is badly needed. This innovation could tap into the emerging capabilities of the World Wide Web, to provide better and more flexible descriptions for a variety of uses (Campbell and Fast forthcoming). But it is equally important to use new visualization techniques to restore dimensions to traditional catalogues, particularly in card form, that have been lost in current OPAC design, so that the richness of the catalogue structure can be fully exploited.

Most important of all, the evidence of this study suggests that catalogues and search engines, for all their differences, have a common objective that is all too often obscured: the creation of communities of authors, readers, searchers, and learners through a variety of grouping and ranking mechanisms. This common objective places search engines and catalogues on a continuum, rather than in opposition across a technological divide. And this complementary relationship has enormous, if currently unrealized, potential for new and exciting innovations in the information landscape.

Acknowledgement

The abstract of this paper is printed in Julien, Heidi, and Sharon Thompson, eds. *Access to Information: Technology, Skills, and Socio-Political Context*. Proceedings of the 32nd Annual Conference of the Canadian Association for Information Science, Winnipeg, MB, June 3–5, 2004 (http://www.cais-acsi.ca/2004proceedings.htm).

Notes

1 In February of 2002, the search services of Yahoo!, MSN, and Google drew a combined audience of 95 million visitors (Nielsen/NetRatings 2002).

2 A full account of the study can be found in Fast and Campbell forthcoming.

3 http://www.lib.uwo.ca/.

4 Camtasia is produced by Techsmith, http://www.techsmith.com/.

References

Bates, M. 2002. After the dot-bomb: Getting Web information retrieval right this time. *First Monday* 7(7): 1. http://www.firstmonday.dk/issues/issue7_7.

Berners-Lee, T., with M. Fischetti. 1999. *Weaving the Web: The original design and ultimate destiny of the World Wide Web*. New York: HarperBusiness.

Borgman, C.L. 1996. Why are online catalogs still hard to use? *Journal of the American Society for Information Science* 47(7): 493–503.

Brown, J.S., and P. Duguid. 2000. *The social life of information.* Boston, MA: Harvard Business School Press.

Buckland, M. 1991. *Information and information systems.* Westport, CT: Praeger.

Campbell, D.G., and K.V. Fast. Forthcoming. Academic libraries and the semantic Web: What the future may hold for research-supporting library catalogues. *Journal of academic librarianship*.

Eisenstein, E. 1983. *The printing revolution in early modern Europe*. Cambridge: Cambridge University Press.

Fast, K.V. and D.G. Campbell. Forthcoming. "I still like Google": University student perceptions of searching OPACs and the Web. *Proceedings of the 2004 Conference of the American Society for Information Science and Technology, November 2004*. Medford, NJ: Information Today.

Jansen, B.J., A. Spink, and T. Saracevic. 2000. Real life, real users, and real needs: A study and analysis of user queries on the Web. *Information Processing and Management* 36(2): 207–27.

Lubetzky, S. 2001. *Principles of cataloging: Final report—Phase I: Descriptive cataloging*, in *Writings on the classical art of cataloging*, comp. and ed. Elaine Svenonius and Dorothy McGarry, 255–341. Englewood, CO: Libraries Unlimited.

Nielsen/NetRatings. 2002. Yahoo!, MSN, and Google lead in online search, according to Nielsen/Netratings. http://www.nielsen-netratings.com/pr/pr_020404.pdf.

Panizzi, Sir A. 1850. Response to question 9814. *Report of the Commission Appointed to Inquire into the Constitution and Government of the British Museum, with minutes of evidence*. London. Quoted in S. Lubetzky, *Seymour Lubetzky: Writings on the classical art of*

cataloging, comp. and ed. Elaine Svenonius and Dorothy McGarry (Englewood: Libraries Unlimited, 2001) 348.

Weinberger, D. 2002. *Small pieces loosely joined: A unified theory of the Web.* Cambridge, MA: Perseus.