10-15-2003

Librarians and Link Rot: A Comparative Analysis with Some Methodological Considerations

David Tyler
University of Nebraska - Lincoln, dtyler2@unl.edu

Beth McNeil
University of Nebraska-Lincoln, mcneil@iastate.edu

Librarians and Link Rot: A Comparative Analysis with Some Methodological Considerations

David C. Tyler and Beth McNeil

abstract: The longevity of printed guides to resources on the web is a topic of some concern to all librarians. This paper attempts to determine whether guides created by specialist librarians perform better than randomly assembled lists of resources (assembled solely for the purpose of web studies), commercially created guides ('Best of the web'-type publications), and guides prepared by specialists in library science and other fields. The paper also attempts to determine whether the characteristics of included web resources have an impact on guides' longevity. Lastly, the paper addresses methodological issues of concern to this and similar studies.

Almost since its arrival as a publicly available resource, web users and bibliographers have found themselves frustrated by the instability of the World Wide Web and with the intermittent availability and in many instances the outright disappearance of sites, pages, and other web objects. Perhaps as a result of this near-universal frustration, a number of researchers have made attempts to determine just how unstable a resource the World Wide Web is and just how useful, in terms of the longevity of their accuracy, are the guides, finding aids, and bibliographies that have sprung up around it. In reviewing this literature, the authors have had a nagging sense both that the studies that examine web bibliographies may not have taken advantage of some of the discoveries made in studies of the web as an unstable entity and may have been, to some degree, inadvertently biased by either the size or character of their lists/samples or by their methodologies. To strengthen our understanding of the usefulness of web bibliographies over time and to discover and discuss some of the limitations of such studies, we examined a large and various collection of web bibliographies published serially over a eight-year period, the *College & Research Libraries News* "Internet Resources" columns published from 1994 to 2001 (referred to henceforth as the "*C&RL News* web bibliographies").

portal: Libraries and the Academy, Vol. 3, No. 4 (2003), pp. 615–632.

Review of Literature

As mentioned above, many of the studies of website, page, and/or object availability in the literature may be divided into two camps: those primarily interested in the behavior of the entities that make up the World Wide Web and those primarily interested in how the usefulness of the prepared finding aids and bibliographies that have grown up around the web are affected by that behavior. The former studies, let us call them "web-directed," tend to examine randomly assembled groups of items so that their conclusions can be extrapolated to apply to the web as a whole. These studies tend to be more "diachronic" in their approach: they usually involve the checking of websites, -pages, and/or -objects regularly and continuously over a period to determine whether and how they occur and/or change over time. The latter sort of studies, let us call them "bibliography-directed", tend to use lists that were consciously assembled, either directly by the researcher or at one remove, and to be more "synchronic" in their approach: they usually involve checking once or twice during brief discrete intervals that the listed websites, -pages, and/or -objects are still available.

As one might expect, the web-directed studies show more variety in approach and focus than do the bibliography-directed studies. For example, the 1997 study by Fred Douglis et al., "Rate Change and Other Metrics: a Live Study of the World Wide Web," which is primarily concerned with web caching, attempted to quantify the rate and extent of changes to web resources by collecting two traces at the Internet connections of two large corporate networks over 17 and 2 days, respectively.[1] The larger trace comprised "95,000 records from 465 clients accessing 20,400 distinct servers and referencing 474,000 distinct URLs."[2] In their recurring *State of the Web* surveys, Terry Sullivan and the site "All Things Web" (ATW), have used ATW's harvester to randomly gather pages from the web—44 pages in 1997, 213 in 1998, and 200 in 1999, the latest sample year available—and have examined, among other things, changes in average total page size and the "incidence and prevalence of broken hyperlinks."[3] For his study into web-site and -page mortality rates and into the rates and types of change they experience, Wallace Koehler randomly selected 361 sites and 361 pages in late 1996.[4] He then checked them regularly over a 53-week period, the sites being checked during the period at three separate intervals and the pages being checked weekly.[5] Koehler also published in the same year a piece on the incorporation of web documents into library collections based on his findings.[6] He later published a follow-up article that re-examined the behavior of the 361 web pages of the original study so as to provide some insight into the life cycles and change rates of an aging set of web-pages.[7]

The studies focusing on bibliographies, guides, and/or finding aids are fairly similar in focus and approach. One of the earlier studies, S. Mary P. Benbow's "File Not Found" focused on two of Benbow's articles published in 1995 and 1997 in *Internet Research: Networking Applications and Policy* that contained 74 and 69 URLs, respectively, and were checked for accuracy in late 1997 or early 1998.[8] Two later articles, Joel Kitchens and Pixey Anne Mosley's "Error 404: Or, What Is The Shelf-Life Of Printed Internet Guides?" and Mark Taylor and Diane Hudson's "'Linkrot' And The Usefulness Of Web Site Bibliographies," examined the accuracy of URLs published in several various bibliographies.[9] Kitchens and Mosley reviewed samples from several "Best of the web"

books (sample size: 3,941 URLs).[10] Taylor and Hudson reviewed the URLs in the *C&RL News* web bibliographies published between October of 1997 and October of 1998 immediately after the publication of the last article (sample size: 482 out of 510 URLs), with a follow-up review of the active links performed six months later.[11] Thomas O'Daniel and Chew Kok Wai studied 3,236 sites submitted by their students for a course in electronic commerce.[12] They checked not only link failure at 2 intervals separated by six months but also analyzed for correlation between domain names and the regional allocation of IP addresses.[13] Most recently, two researchers at the University of Nebraska, John Markwell and David W. Brooks investigated on a monthly basis, from August 2000 to May 2002, the incidence of "link rot" in the 515 hyper-links contained in the on-line materials of three graduate-level biochemistry courses created in August of 2000.[14] They also have reported their results on-line, and have been the subject of an article by Vincent Kiernan that appeared in *The Chronicle of Higher Education's* on-line edition.[15]

In any attempt at categorization, there are always exceptions to the rule. Although they do not quite fit our model and are concerned with particular collections rather than with web bibliographies or with the web proper, we would also like to draw the readers' attention to Michael Nelson and Danette Allen's "Object Persistence and Availability in Digital Libraries," and Steve Lawrence et al.'s "Persistence of web References in Scientific Research."[16] The former measured, with thrice-weekly checks over slightly more than a year's time, the persistence and availability of 1,000 objects selected randomly from twenty digital libraries. The latter investigated the continued accuracy of 67,577 URLs cited in research papers using NEC Research Institute's scientific digital library ResearchIndex (formerly CiteSeer).

The results of the studies mentioned above will be integrated later into this paper through analysis and comparison with our own findings.

Methods and Definitions of Terms

With the assistance of a student worker, we manually attempted to access the 2,729 URLs of the http-based resources (gopher and ftp sites, listservs, and e-mail addresses were ignored) listed in the *C&RL News* web bibliographies published over the eight year period from 1994–2001, using Microsoft's Internet Explorer 6.0 with the recommended security settings. The URLs were first checked as a whole in mid-June 2002 so that, for the purposes of our later discussion, each year's bibliographies would be as a group an average of an even calendar year old. We also performed two follow-up examinations involving two separate portions of the URL lists, six weeks later which will be detailed at the end of this section. In compiling and numerating the lists, URLs with obvious typographical errors were corrected when caught, and duplicate URLs were removed when caught. Also, a few URLs were missed when the authors were first numbering the addresses for inclusion, and while most of these missed URLs were later added to the lists some may still have been missed. If there are discrepancies in the number of URLs recorded in this and in Taylor and Hudson's study of portions of the 1997 and 1998 *C&RL News* web bibliographies, they may largely be attributed to these causes.

In examining the URL lists, our first intent was to review them for their usefulness—in terms of their incidences of success and failure, their apparent "half-lives,"

and their inferable "rates of decay" and/or "decay curves"[17]—from the perspective of the casual user. In our first examination we tried to determine if the URLs listed were valid without inquiring into whether the listed sites and pages were still extant elsewhere and could be located with some effort. Our first step was to determine which addresses were "live," which provided a "re-route," and which were "dead." For our purposes, a "live" URL is one that returns the intended site or page as annotated in the *C&RL News* web bibliographies, a page that redirects the user to a subscription-free registration or login page that subsequently automatically delivers the desired page, or a page that directs the user to a subscription page if the annotation indicates that such a redirection is to be expected. A "re-route" URL is one that either results in one's being taken to the intended site's or page's new and currently correct address automatically or that calls up a page that provides said new and currently correct address. A "dead" URL is one that returns a "404 Not Found," "403 Forbidden," or other such error message, that returns a *persistent* domain name server (DNS) error message, or that fails to meet the criteria for either *live* or *re-route* URLs above (e.g., a URL that returns a site or page that is active but that is not the one described, a URL that requires a subscription when such is not indicated by its annotation, which, we recognize, may be an annotator's error, and so forth). In the case of DNS errors, URLs that returned such errors were rechecked twice, once during the following morning when University web traffic was low and then again three days later. If the error persisted or a wrong page was eventually returned, the site or page was recorded as being *dead*; if the correct site or page or an accurate re-routing page was eventually returned, the URL's status was recorded as *live* or *re-route*, respectively.[18]

Our second purpose was to discover whether the lists prepared by specialist librarians for *C&RL News* were superior in their staying power to those assembled randomly and to those prepared by others. To this end, the incidences of failure from several comparable lists will be presented throughout our discussion.

Our third aim was to attempt to investigate whether some of the characteristics of the URLs in question had any bearing on their status as *live*, *re-route*, or *dead*. To accomplish this we disaggregated our lists by three criteria. First, we checked the URLs' top level domain types (e.g., "com," "edu," etc.) and grouped them into one of five types ("TLD type"); second, we assigned the bibliographies one of four broad topical headings and grouped them accordingly ("Topic"); and, third, we recorded the URLs' server-level domain addresses and grouped them into one of three types ("Server Domain Level" or "SDL").

The assigning of URLs to TLD types was largely straightforward for four of the five types: URLs were noted as ending in ".gov," ".com," ".edu," and ".org," as appropriate. For the fifth grouping, all other types—".mil," ".net," numerical addresses, addresses ending in country-specific TLDs, and so forth—were placed together under the heading "Other," as the occurrences of each were dwarfed by the occurrences of the other four types. We gave some consideration to reclassifying country-specific two-letter tags that were identifiable as belonging to one of the other types, such as "co.uk" and "net.de," and to reclassifying those tags whose membership in another type could be inferred, such as McGill University's homepage's address (http://www.mcgill.ca/), but decided against doing so to avoid inconsistencies.[19]

The assigning of topical headings to the bibliographies was also rather straightforward. Each bibliography, upon the basis of its subject, was classified under one of four headings: "Sciences," "Social Sciences," "Arts and Humanities," and "Library and Information Science." The last topic was used to cover not only library and information science web bibliographies but several general, ready-reference web bibliographies that the authors could not justifiably assign to one of the other three groups, as well.

Lastly, we grouped the URLs into three types by the third variable whose impact we decided to investigate, server domain level. SDL "Type A" comprises "zero-level" addresses, those with no directory structure (http://aaa.bbb.ccc/) and those addresses with sub-files that returned exactly the same page as their "zero-level" addresses (these addresses usually had just one or two levels and ended in " . . . /index.html," " . . . /default.htm," and so forth). Addresses assigned to "Type C" were those that specified a port number (e.g., http://aaa.bbb.ccc:8080), including the standard port. We considered reclassifying URLs that specified the standard port, but we encountered a few sites that somehow were *dead* when the port number was included and were *live* without it and so decided against reclassification. SDL "Type B" comprises all URLs with sub-files and/or file names that did not meet the criteria for *Type A* and *Type C* (e.g., http:aaa.bbb.ccc/xxx/ or http://aaa.bbb.ccc/xxx/page.html).

As mentioned above, our final step was to perform two follow-up checks on subsets of the original list six weeks after the initial data gathering. For the first follow-up, we re-checked all of the links we had tagged as *dead* to determine if website and -page intermittency had skewed our initial data collection and to some extent invalidated our findings. Taylor and Hudson, after an interval of six months, had performed a similar follow up with just the URLs that they had recorded as being *live* to establish a rate of failure over time; Benbow and Kitchens and Mosley performed no follow-ups; O'Daniel and Kok Wai checked all of their sites twice, with the checks being separated by a six-month interval; and Markwell and Brooks, much like Koehler, checked the whole of their lists at regular intervals.[20] In our follow-up, URLs that returned a once-*dead* page or that provided an accurate *re-route* were classified as "undead;" URLs that did not were deemed to be still *dead*.

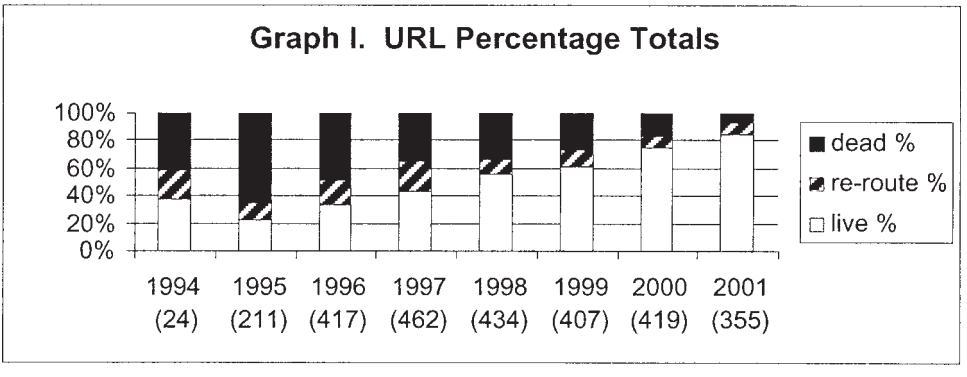
For the second follow-up, we used MetaCrawler, Dogpile, Google, AltaVista, and HotBot to search for sites and pages from the 1998 bibliographies that were still *dead* after the performance of the first follow-up.[21] The 1998 bibliographies were chosen as they were, at more than three years old, beyond the pale for most of the consulted studies. Searches were performed using just the information provided in the *C&RL News* web bibliographies' annotations. Sites and pages were searched for with several engines, and if a site or page as described did not appear in the first 50 results returned by the various engines, it was deemed to be "Lost." Sites and pages that were returned were classified as "Found." While additional searching at the higher domain level of the original websites might have yielded additional "found" websites, we chose to search multiple search engines to widen results instead and to stay with the original plan to search solely using *C&RL News* information.

Findings and Results

For the purposes of our discussion, the data and results for the 1994 bibliographies, though included and presented with the later years', will be largely ignored. The 1994 bibliographies, which contained a scant total of 24 URLs, provide less than 10 percent of the average number of URLs provided by the later years' bibliographies and are far less varied in their character. Rather than allow an odd year to skew the results, we have chosen to simply present it alongside the others largely without comment.

I. Persistence and Failure of URLS, By Year

The initial analysis focused on the validity of the *C&RL News* web bibliographies. As Graph I shows the *C&RL News* web bibliographies hold up rather well.



Graph I. Values within parenthesis ( ) indicate the URL total for the year.

Our analysis indicated that the strict half-life for these guides, which takes only *live* URLs into account, is somewhere between 4 and 5 years. Were one to include the semi-valid *re-route* links as well, the soft half-life of the web bibliographies would appear to be just over 6 years. Koehler, whose sample was randomly selected between December 1996 and January 1998, reported a half-life of 2.9 years for his sites (and he had lost 66.6 percent of his sample after four years),[22] as compared to 34 percent *dead* at the four-year mark in this study. Among the researchers working with consciously assembled lists, Benbow also reported a half-life of 3 years for the list of URLs that she selected in mid-1995.[23] Markwell and Brooks, however, who selected their URLs in August of 2000, five years after Benbow collected hers, posited a half-life of 55-months,[24] which is comparable to our initial findings.

Our incidences of failure for the one- and two-year-old bibliographies also compare favorably to most of the other studies' findings (even to those of Taylor and Hudson, who also examined *C&RL News* web bibliographies). We found that 7 percent of URLs (24) were *dead* at an average of one year after publication and that 17 percent (73) were *dead* at two. Markwell and Brooks found 16.5 percent of their links to be non-viable after 13 months.[25] Taylor and Hudson found 22.2 percent to be outdated at an average

of one year after publication for their group.[26] Both Benbow and Kitchens and Mosley found nearly 30 percent of their URLs to be *dead* after two years.[27] Interestingly, O'Daniel and Kok Wai found just 2.7 percent of their links to be *dead* 6 months after collection, and Markwell and Brooks later reported 18.6 percent of their links to be *dead* after 19 months, a change of just 2.1 percent over the 6 months following a rate of decay of 16.5 percent over 13 months.[28] Both of the latter studies used URLs collected near the year 2000, as were the last two years' worth of URLs in our study, and these lower incidences of failure could suggest that the web, though still very unstable, may be becoming a more stable environment, at least insofar as "selected" URLs are concerned. These results also suggest, when one compares the incidences of failure in the randomly selected and consciously assembled topical lists over time, that URL selection is very much a value-adding service, one that is perhaps even more so when the lists are assembled by specialist librarians. Obviously, as a result, findings from examinations of consciously assembled topical lists of URLs should not be used to discuss the behavior of the World Wide Web in general.

Obviously, as a result, findings from examinations of consciously assembled topical lists of URLs should not be used to discuss the behavior of the World Wide Web in general.

Surprisingly, Lawrence et al. and Nelson and Allen, whose studies were concerned with URLs and objects from digital libraries, found similar incidences of invalid URLs in their samples. Lawrence et al. found that 23 percent of URLs in their study were invalid after one year and that 54 percent were invalid after six; and Nelson and Allen found that 3.1 percent of objects were no longer available at the end of their 161 day testing period,[29] a rate similar to O'Daniel and Kok Wai's. Their results are a bit surprising, as one would assume that objects' "being placed in a digital library is indicative of someone's desire to increase the persistence and availability of an object."[30] It would seem, however, in terms of the percentage of *dead* URLs, that a managed collection of web-available resources may not perform much better than a collection of URLs that has been carefully assembled by a specialist librarian or other knowledgeable professional.

II. By Top Level Domain Type, By Year

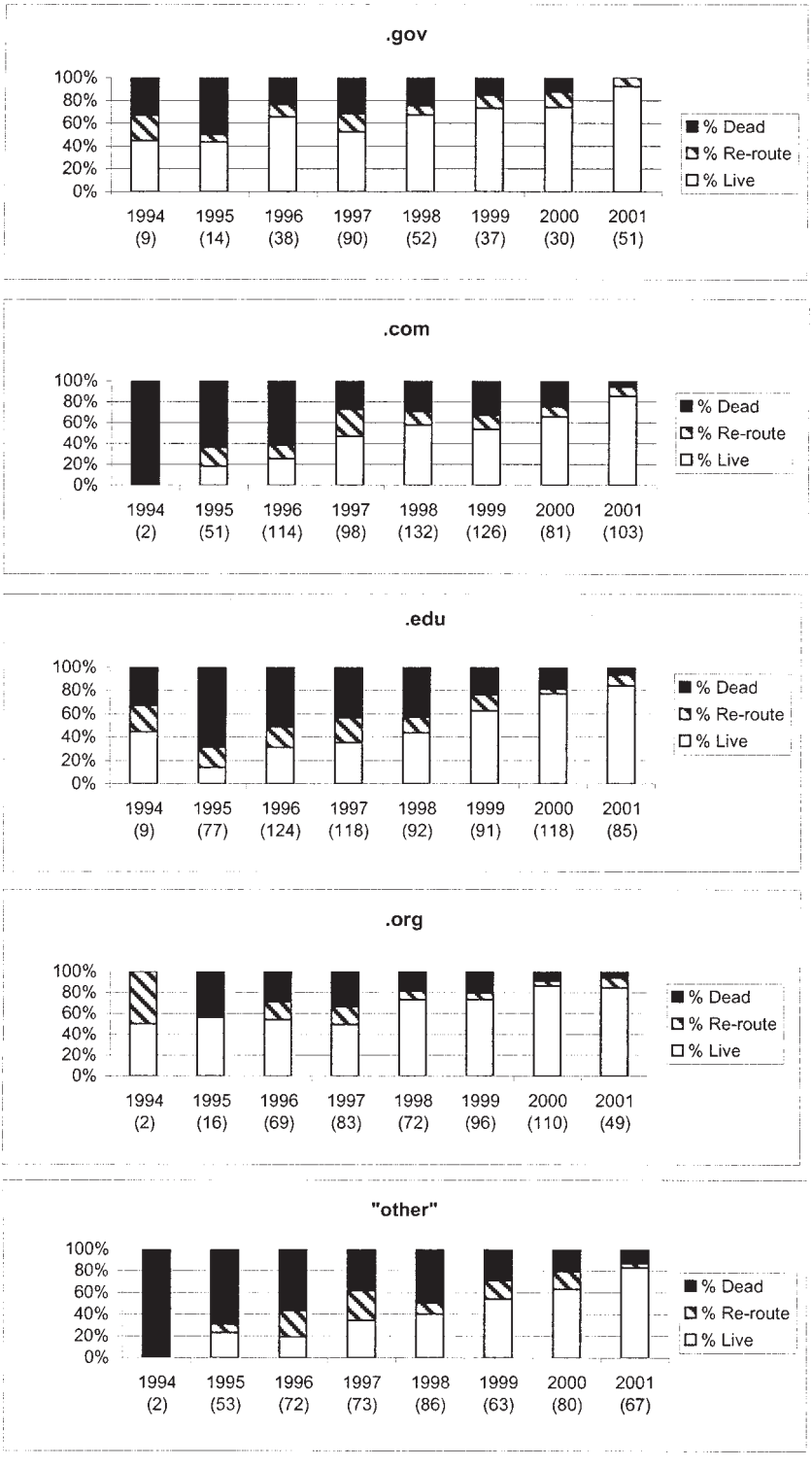
Our second step was to examine our results by top-level domain type to determine if the sites' and pages' types had any bearing on the viability of the listed URLs over time. The *C&RL News* web bibliographies tend to include URLs from TLD types *.com* (702) and *.edu* (715) considerably more often than they include URLs from the TLD types *.org* (497) and the types in the *Other* (496) category. They also include *.com* and *.edu* more than twice as often as they include *.gov* URLs (322), which is ironic in that, as graph II reveals, with the exception of the URLs in the *Other* category, TLD types *.com* and *.edu* are fairly consistently the worst performers in terms of percentage of *dead* URLs while *.gov* URLs are usually the best. Interestingly, in addition to having a high percentage of *live* URLs, *.gov* URLs also had comparably high *re-route* percentages for the 1999–2001

bibliographies, a reflection, perhaps, of government's tendency to initially publish items at one address and later archive them at another.

Our finding that *.com* and *.edu* TLD types are more likely to be *dead* and that *.gov* URLs are more likely to persist is consistent with findings of Markwell and Brooks, Koehler, and Taylor and Hudson.[31] However, the results of Koehler's web-directed study and those of the several bibliography-directed studies differ greatly over *.org* addresses. Koehler found that *.org* sites were available 100 percent of the time over 3 samples and that only 8.3 percent of *.org* pages were "comatose" (i.e., consistently unavailable over a six-week period) over his initial 1999 search period.[32] Koehler's failure rates for randomly selected *.org* URLs seem awfully low, especially when compared to the rates from the studies of consciously selected URLs. For example, Markwell and Brooks' 11.6 percent failure rate for *.org* sites and pages is comparable to our incidences of *dead* URLs for the 2001 and 2000 bibliographies (6 percent and 9 percent, respectively),[33] and even Taylor and Hudson's failure rate for *.org* URLs of 20.7 percent seems more in keeping with the other studies' results, relative to the other domains' percentages in their study (although, in absolute terms, in our sample *.org* URLs do not reach 21 percent *dead* until the bibliographies are an average of 3 years old, whereas Taylor and Hudson's reaches 20.7 percent after the bibliographies are an average of only 1 year old).[34] Our initial results suggest that this relationship between the randomly and consciously selected lists should be reversed, as it is for almost all of the other domain name types. Our first inclination here is to suggest that there may be something unusual about Koehler's *.org* sites and pages or to suggest that the other studies may contain more pages and fewer sites than Koehler's study.

Interestingly, *.net* addresses, which make up a large portion of the *Other* category in our study, do very poorly in two of the three studies just mentioned (Taylor and Hudson, and Koehler), but country tags do a bit better in Koehler's initial study.[35] As part of our *Other* category, both did rather poorly in this study (.net=101 of 496; country tags =348 of 496). We believe that some of the discrepancies between our results and Koehler's and Taylor and Hudson's in this area could be attributed to their reclassifying geographic and inferable URLs, and to our not doing so.

In terms of list half-life, in our study *.gov* URLs maintained over 50 percent *live* links all the way back to 1996, while *.com* dipped below 50 percent *live* after 1998 and *.edu* after 1999. The saving grace for our *.com* and *.edu* URLs, perhaps, was their tendency to provide a fairly large number of *re-route* links for the year of their half-lives and for the year after. Surprisingly, *.org*, once it reached its half-life, after a sharp drop-off between 1998 and 1997, tended to hover around 50 percent *live* URLs for several years. Therefore, we would concur with Koehler that *.gov* and *.org* URLs tend to remain valid longer and/or to a greater extent, and that the converse appears to be true for *.com* and *.edu* addresses.[36] However, we would again like to direct attention to the discrepancy between our incidence of *dead* URLs for 1-year-old bibliographies and Taylor and Hudson's failure rate for 1-year-old bibliographies, which is nearly equal to our failure rate for a 3-year-old bibliography.[37] As mentioned before, this discrepancy may point to a new stability in the "useful" sites and pages selected by web bibliographers. It may also, as will be discussed later, indicate a gross methodological failure.



Graph II. Values within parenthesis ( ) indicate the URL total for the year.

III. By Topic, By Year

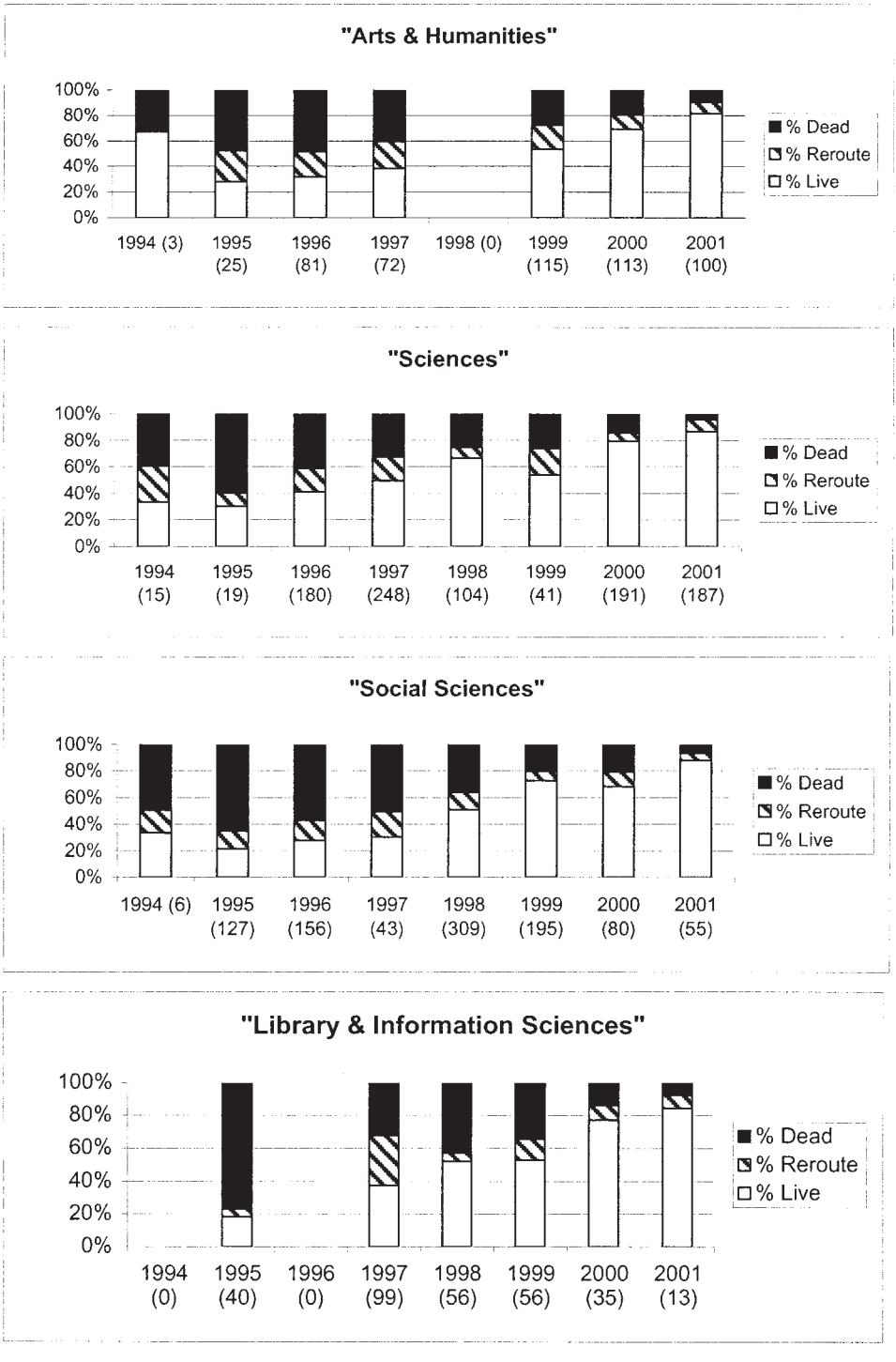
Our third point of investigation was the impact of general *Topic* over time, and to this end we arranged our results by the four aforementioned topical categories. The *C&RL News* web bibliographies provided nearly the same number of URLs for both *Sciences* (985) and *Social Sciences* (971), but the *Arts and Humanities* (509) bibliographies provided just bit more than half as few as either. *Library and Information Science* (264), which includes "ready-reference" bibliographies of a general nature, provided just over one-quarter as few. We cannot say whether this seeming bias is the result of *C&RL News'* preferring to accept web bibliographies in the sciences and social sciences or, as seems more likely, of librarians' producing more *Sciences* and *Social Sciences* bibliographies, but a clear disparity exists.

Disappointingly, the variable *Topic* for the three major topics does not seem to have a clear effect upon URLs' persistence over time, as graph III shows. If one discounts the partial data from 1994, the *Sciences* seem to maintain a large percentage of *live* URLs over the long run, but over the short run the other major fields' percentages for *live* URLs appear to be comparable. Their percentages for *dead* URLs also appear comparable to *Sciences'*, though the *Social Sciences* appear to fare a bit worse, perhaps because the *Sciences* and certainly the *Arts and Humanities* proved to be better about providing *re-route* links as the bibliographies get older.

Library and Information Science started out very handsomely and then finished worse than the other three topics. However, it is difficult to say anything definitive about the topical grouping *Library and Information Science* because it is a considerably smaller group than the other three and, more importantly, because its year-to-year appearance is much more erratic.

If one were to assign strict and soft half-lives to the topical groupings, graph III shows that the *Sciences* group would have reached its strict half-life in 1997 and its soft half-life somewhere between 1996 and 1995. The *Social Sciences* fared a bit worse, with 1998 and 1997, respectively. The *Arts and Humanities* lists are a bit more difficult to pin down as, astoundingly, there were no *Arts and Humanities* web bibliographies in 1998. It would appear, however, that *Arts and Humanities* reached their strict half-life for *live* URLs just after 1999, but, because of their superior use of *re-route* URLs, the *Arts and Humanities* soft half-life for combined *live* and *re-route* URLs may be extended to 1996 or 1995. Also of note: the number of *dead* links for *Arts and Humanities* never climbed above 50 percent.

In looking at the three major topical groups, one would be inclined to suggest that *Sciences* web bibliographies perform the best over time, with *Social Sciences* web bibliographies performing nearly as well and with *Arts and Humanities* web bibliographies performing the worst. However, when one takes into account our earlier findings regarding the effect of TLD types on URL longevity, it seems likely that the great differences between the *Sciences* and *Arts and Humanities* web bibliographies could in large part be accounted for by the relative preponderance of *.gov* URLs in the *Sciences* web bibliographies and of *.edu* and *Other* URLs in *Arts and Humanities* web bibliographies. As graph II indicated earlier, *.gov* URLs persist longer than other types, and *.edu* sites tend to provide more *re-route* URLs, especially as the web bibliographies age. Graph III



Graph III. Values within parenthesis ( ) indicate the URL total for the year.

indicates that the *Sciences* and *Arts and Humanities* web bibliographies behave in much the same manner as their dominant TLD types. As a result, we are unwilling to assign too much importance to the half-life differences between the topical groupings, as such, and are more inclined to ascribe the larger part of the topical web bibliographies' differences to disparities between the behavior of the domain name types that predominate in each *Topic's* web bibliographies. We feel it unlikely that the differences between the topics noted here necessarily point to any great differences in individual URL longevity between the URLs of *Sciences*, *Social Sciences*, and *Arts and Humanities* web-sites so much as they point to the differences in the make-up of the topical lists. It is therefore unlikely that one could conclude from our results that an *.edu* URL from a *Sciences* web-bibliography would be more or less likely to be valid over time than an *.edu* URL from a *Arts and Humanities* web-bibliography published in the same year.

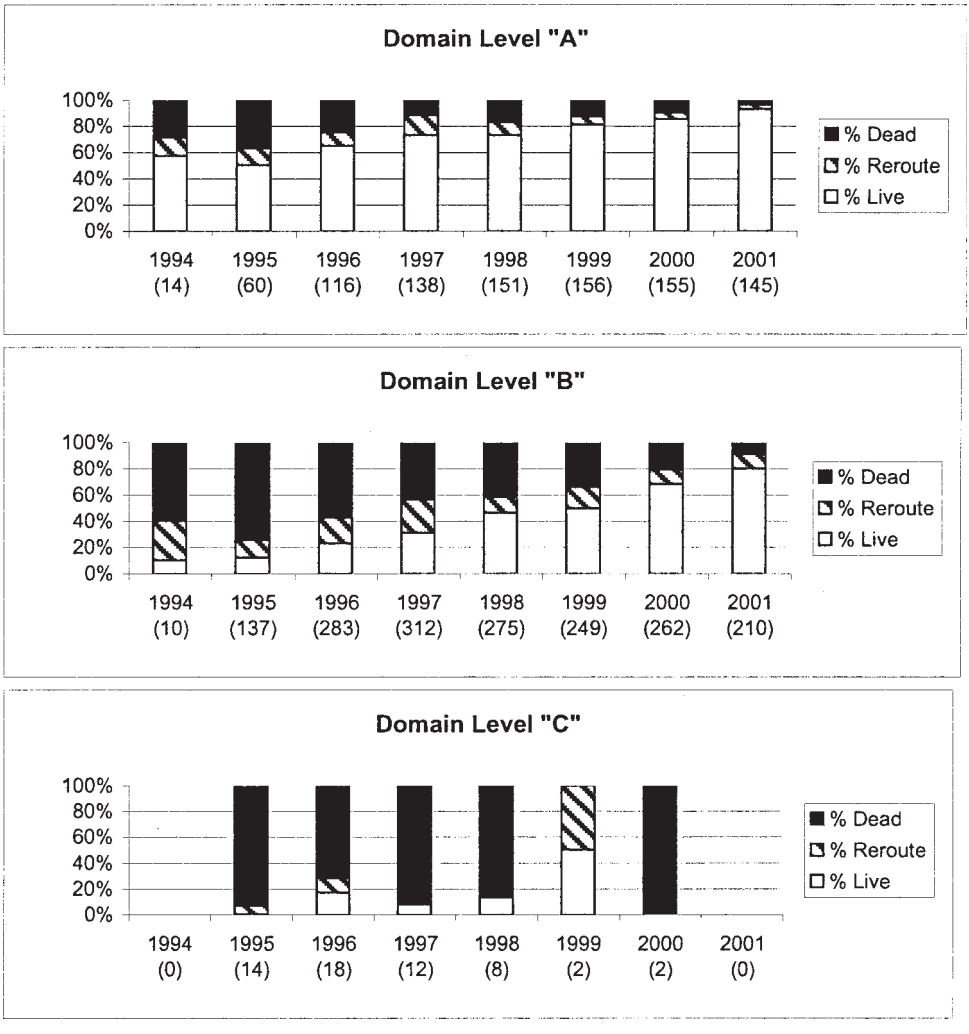
IV. By Server Domain Level, By Year

The final URL characteristic that we examined for its possible effect was placement. It was our expectation that a site's or page's place in its server's directory structure could have an effect upon URL validity over time. In examining our sample, we found that *Type A* (935), or top-level and inferred top level, domains made up considerably less of the sample than *Type B* domains (1738), by nearly one-half, and that *Type C* domains (56), those specifying a port number, made up an insignificant 2 percent of the sample.

It was our expectation that a site's or page's place in its server's directory structure could have an effect upon URL validity over time.

However, as graph IV indicates, *Type A* and *Type B* positions were nearly inverted where *live* and *dead* links are concerned (*Type C*, of course, presented by far the worst performance, but as with *Library and Information Science* above *Type C* provided too small and erratic a sample to comment meaningfully upon). In terms of the types' strict half-lives, *Type A* may have reached its in 1995, but *Type B* URLs certainly reached theirs by 1999 or 1998. If one were to combine the *live* and *re-route* URLs, *Type B* would fare much better, with a soft half-life between 1997 and 1996. It should be obvious that a preponderance of *Type B* URLs, those with lengthier directory structures and those that specify a particular file name, would be the greatest factor that contributes to most web bibliographies rapid obsolescence (excluding of course the unlikely possibility of the inclusion of a large number of *Type C* URLs).

These results confirm the findings of Koehler, though his results were less extreme in their disparities and his analysis was considerably more sophisticated, and those of Kitchens and Mosley, though they suggest that URLs that specify particular files are more detrimental to a web finding-aid's longevity than are URLs with subdirectory references.[38] Regardless, it would seem clear that the server domain level of its URLs should have a large if not determining impact on a web bibliography's longevity.



Graph IV. Values within parenthesis ( ) indicate the URL total for the year.

V. The Incidence of Undead URLs

As mentioned previously, Taylor and Hudson followed-up their initial examination with a second check of their *live* URLs after an interval of six months to determine how many additional URLs had become invalid over the span. Koehler's rather more extensive work with the intermittency of web-pages, however, suggested to the authors that Taylor and Hudson's may have been the wrong tack and that we ought, rather, to re-check how many of our *dead* links had remained so. As table I shows, after the 6-week interval 8 percent (70) of the 868 *dead* links from our initial examination had returned to activity, or were *undead*, with the highest percentage returns occurring among the most-recently published bibliographies and the lowest among the oldest. What may be most surprising about our results is that the web bibliographies that are between 3 and 5 years old show *undead* rates of 8–9 percent.

Our results, although they initially seemed a bit better, are roughly in keeping with Koehler's findings: he initially posited a rate of intermittency of 5.2 percent for sites and 5 percent for pages over a six-week interval, but in his later study he remarked that the intermittency rate for his sample had "varied from between less than 5 percent in a given week to more than 20 percent."[39] This characteristic of site and page intermit-

Table 1

Undead/Dead Follow-up

	Dead	percent Dead	Undead	percent Undead	Total
1994	9	90 percent	1	10 percent	10
1995	136	99 percent	1	1 percent	137
1996	201	98 percent	4	2 percent	205
1997	149	91 percent	15	9 percent	164
1998	137	92 percent	12	8 percent	149
1999	96	91 percent	10	9 percent	106
2000	54	74 percent	19	26 percent	73
2001	16	67 percent	8	33 percent	24
Total	798	92 percent	70	8 percent	868

tency, or the fact that many *dead* URLs can become *undead* over a short interval and can continue to behave thusly over several years, suggests to us that there are methodological problems for most of the bibliography-directed studies. Koehler's and our study's findings suggest that synchronic examinations of a body of URLs or of sites and pages are unlikely to produce an entirely accurate expiration rate or half-life measure. Koehler's and Markwell and Brooks' more diachronic methods, which are more likely to catch sites at several points on the wave of availability and unavailability, should be more correct for the subject matter.

VI. Lost and Found URLs

For our final analysis, we repeated Taylor and Hudson's search of the web to determine if the information in annotations in web bibliographies helped them to remain useful after many of their URLs were no longer accurate. Taylor and Hudson had been able to find all but 4.4 percent of their "outdated" URLs using the MetaCrawler engine.[40] At the time of their searches, Taylor and Hudson's oldest bibliographies were no more than 1 year old. For this study, largely the same body of web bibliographies, those from 1998, was searched using MetaCrawler, Google, Dogpile, AltaVista, and HotBot. For these four-year-old bibliographies, we were able to locate sites and/or pages for 40

percent of the *dead* URLs (54 of 137) using just the titles and other information provided in the bibliographers' annotations.

With respect to the several recommendations that Taylor and Hudson provide for making lost sites and pages more easy to find, we would have to concur, based upon our experiences, that listing the name of the resource given in the title bar and including the name of a contact person are immeasurable helps in both finding and verifying resources, and we would add that including a site author or host institution, especially in the case of ".edu" sites and pages, would be very helpful as well. Regardless of how detailed the annotations may be, our finding 40 percent of *dead* sites and pages for bibliographies from 1998 suggests that web-finding aids retain value for years after publication, perhaps even well beyond the two-plus years that Kitchens and Mosley suggested in their study.[41]

Conclusion

Upon review of the data, we are inclined to concur with authors Taylor and Hudson that well-prepared paper-based bibliographies of websites and pages are likely to remain useful over a fairly long period of time despite the ephemeral accuracy of URLs: A large percentage of the URLs in our study remained valid or semi-valid over several years, and a large percentage of the listed sites with *dead* or inaccurate URLs remain locatable as well, well after the initial publication of web bibliographies. The viability of the bibliographies would appear to some degree to be a function of their contents, with those comprised of a large number of *Type A* URLs and *.gov* addresses remaining accurate longer than those containing mostly *.com*, *.edu*, *.net*, and geographic TLD names, and those comprised largely of URLs that contain subdirectories or that specify particular page addresses.[42]

We also congratulate the profession, for it would seem that bibliographies prepared by librarians for *C&RL News* performed as well as or better than both the randomly assembled lists and the web bibliographies prepared either commercially or by practitioners and students in other disciplines. In fact, contrary to Markwell and Brooks' suggestion that the archiving of sites and pages could improve the longevity of link lists,[43] our findings may signify that well-assembled lists perform as well as or better than managed collections in some instances. However, our study's web bibliography results were considerably better than Taylor and Hudson's even though we were both, in part, studying approximately the same web bibliographies.

We also congratulate the profession, for it would seem that bibliographies prepared by librarians for *C&RL News* performed as well as or better than both the randomly assembled lists and the web bibliographies prepared either commercially or by practitioners and students in other disciplines.

This last finding leaves the authors concerned as to the effectiveness of the methodology we, like several other researchers, have chosen to use in examining the accu-

racy and longevity of printed web bibliographies, a method we had earlier indicated was "synchronic" in its approach. Although we were heartened to discover that our and Taylor and Hudson's findings produced fairly long half-lives and inferable decay curves and/or rates that were gentle and gradual (see Graph I) and similar to those of Markwell and Brooks' more diachronic study of educational links in biochemistry,[44] Koehler's findings regarding the intermittency of sites and pages and our own discovery that a significant portion of the URLs that we had labeled *dead* were active again after a six-week interval (e.g., 33 percent of *dead* URLs from 2001's and 26 percent of *dead* URLs from 2000's web bibliographies were *undead* after the interval) suggest that the common method of performing a synchronic examination or two of web bibliographies is unlikely to produce an entirely accurate picture of their behaviors or accuracy over time. Because of this, we are unlikely to have an accurate picture of the long-term utility of web bibliographies until either a large number of comparable synchronic examinations are performed or until easily employable and reliable diachronic methods are developed.[45]

David C. Tyler is a Reference Librarian and Beth McNeil is Assistant Dean of Libraries at University of Nebraska-Lincoln; the authors may be contacted via e-mail to Ms. McNeil at: mmcneil1@unl.edu.

Notes

1. Fred Douglis, Anja Feldmann, Balachander Krishnamurthy and Jeffrey Mogul, "Rate of Change and Other Metrics: A Live Study of the World Wide Web" (December 1997): 1–13. Available: <http://www.douglis.org/fred/work/papers/roc.pdf> or <http://www.douglis.org/fred/work/papers/roc/> [September 24, 2003]. Fred Douglis et al., "Rate of Change and Other Metrics: A Live Study of the World Wide Web," in *Proceedings of the USENIX Symposium on Internet Technologies and Systems; December 8–11, 1997, Monterey, California* (Berkeley, CA: USENIX Association, 1997), 147–158.
2. Ibid., 3.
3. Terry Sullivan, "All Things web: How Much Is Too Much?" *All Things web* (May 28, 1999). Available: <http://www.pantos.org/atw/35654.html> [September 24, 2003].
4. Wallace C. Koehler, "An Analysis of Web Page and Web Site Constancy and Permanence," *Journal of the American Society for Information Science 50*, 2 (1999): 162–180.
5. Ibid., 166–168.
6. Wallace C. Koehler, "Digital Libraries and World Wide Web Sites and Page Persistence," *Information Research* 4, 4 (July 1999). Available: <http://InformationR.net/ir/4-4/paper60.html> [September 24, 2003].
7. Wallace C. Koehler, " web Page Change and Persistence: A Four-Year Longitudinal Study," *Journal of the American Society for Information Science and Technology* 53, 2 (2002): 162–171.
8. S. Mary P. Benbow, "File Not Found: The Problems of Changing URLs for the World Wide Web," *Internet Research: Electronic Networking Application and Policy* 8, 3 (1998): 247–250.
9. Joel D. Kitchens and Pixey Anne Mosley, "Error 404: Or, What Is the Shelf-life of Printed Internet Guides?" *Library Collections, Acquisitions, & Technical Services* 24 (2000): 467–478; Mark K. Taylor and Diane Hudson, "'Linkrot' and the Usefulness of Web Site Bibliographies," *Reference & User Services Quarterly* 39, no. 3 (2000): 273–277.
10. Kitchens and Mosley, 469–470.
11. Taylor and Hudson, 273–274.

12. Thomas O'Daniel and Chew Kok Wai, "Domain Name and Site Hosting Preferences: Empirical Evidence," *Internet Research: Electronic Networking Applications and Policy* 10, 4 (2000): 308–316.
13. Ibid., 311–312.
14. John Markwell and David W. Brooks, "Broken Links: The Ephemeral Nature of Educational WWW Hyperlinks," *Journal of Science Education and Technology* 11, 2 (2002): 105–108.
15. John Markwell and David W. Brooks, "Broken Links: Just How Rapidly Do Science Education Hyperlinks Go Extinct?" (September 15, 2003). Available: <http://www-class.unl.edu/biochem/url/broken_links.html> [September 24, 2003]; Vincent Kiernan, "Nebraska Researchers Measure the Extent of 'Link Rot' in Distance Education," *The Chronicle of Higher Education* (April 10, 2002). Available: <http://chronicle.com/free/2002/04/2002041001u.htm> [September 24, 2003].
16. Michael L. Nelson and B. Danette Allen, "Object Persistence and Availability in Digital Libraries," *D-Lib Magazine* 8, 1 (January 2002). Available: <http://www.dlib.org/dlib/january02/nelson/01nelson.html> [September 24, 2003]; Steve Lawrence et al., "Persistence of web References in Scientific Research" (2001). Available: <http://www.neci.nec.com/~lawrence/papers/persistence-computer01/persistence-computer01.pdf> [September 24, 2003].
17. "Half-life," "decay," "rate of decay" or "decay rate," and "decay curve" are, of course, terms that properly belong to the physical sciences, and, as so often occurs in the social sciences, we will be somewhat improperly borrowing them and adopting them for conditions and processes that are merely analogous to those found in the sciences.

    In the physical sciences, one of several uses for the term "half-life" is to refer to the time required for a quantity of a radioactive material to decay to a lesser state or be eliminated. We will be employing the term in this sense, and, for our purposes, the "strict" half-life of the several grouped *C&RL News* web bibliographies will be that point where half of URLs in the grouping are no longer valid (i.e., *live*). Unfortunately, for clarity's sake, our discussion of the URLs involves three states: valid or *live* URLs, invalid or *dead* URLs, and semi-valid or *re-route* URLs. Therefore, in the discussion to come, we will often be providing both "strict" half-lives for the *C&RL News* web bibliographies that countenance only the number of *live* URLs and "soft" half-lives that combine the counts for the *live* and semi-valid *re-route* URLs.

    The term "decay," obviously, refers to the process whereby a substance declines into an inferior or weakened condition (i.e., the *C&RL* web bibliographies' URLs decay from a *live* state to *re-route* and *dead* states). The "rate of decay" or "decay rate" would be the changing rate of disintegration or change in a quantity of a material in which the number of disintegrating or changing parts in a unit of time is proportional to the total number of remaining parts of the quantity of the material then present. The "decay curve" would be a graph that would track the activity of a decaying quantity of material over time, showing how much changed and unchanged material existed at several given times.

    In this study, we will not be tracking the decay rate of a single quantity of our "material" (i.e., web bibliographies) and then plotting its decay curve. Rather, we will be discovering incidences of invalidity, semi-validity, and validity at particular averaged intervals for the several grouped *C&RL News* web bibliographies and plotting those incidences under the assumption that what holds true for any one of the grouped web bibliographies at a particular point in their publication lives should hold true for all of them as a group at that point (except, of course, for the 1994 grouping, which is considerably smaller and less varied in its make-up than the later groupings). Therefore, the decay curves and rates of decay that we may present or refer to for the *C&RL News* web bibliographies should be understood to be *inferred* rather than *actual* curves and rates.
18. These definitions are somewhat at variance with our colleagues' definitions. For example, Taylor and Hudson included the returning of a site map or site search engine as a *re-route*,

Kitchens and Mosley appear to have counted *re-route* URLs as *live*, and Benbow disregarded DNS errors.

19. Koehler supplied his data both "as is," as we did, and also with re-classed geographic and inferred domains (Koehler, "An Analysis," 166 and "Web Page Change"). To some degree, Taylor and Hudson mixed their URLs, as well ("'Linkrot,'" 274–275).
20. Taylor and Hudson, "'Linkrot,'" 274; O'Daniel and Kok Wai, "Domain Name," 311–312; Markwell and Brooks, "Broken Links: The Ephemeral," 106–107; Markwell and Brooks, "Broken Links: Just How Rapidly."
21. Taylor and Hudson performed a similar search using just the MetaCrawler engine ("'Linkrot,'" 274–275), and Lawrence et al. also performed a variety of searches ("Persistence of web References," 27–28). The other papers were largely uninterested in the locating of *lost* -sites and -pages.
22. Koehler, "An Analysis," 172; Koehler, "Web Page Change."
23. Benbow, "File Not Found," 248.
24. Markwell and Brooks, "Broken Links: The Ephemeral,"107; Markwell and Brooks, "Broken Links: Just How Rapidly;" Kiernan, "Nebraska Researchers."
25. Markwell and Brooks, "Broken Links: The Ephemeral," 106.
26. Taylor and Hudson, "'Linkrot,'" 274.
27. Benbow, "File Not Found," 248; Kitchens and Mosley, "Error 404," 470–471.
28. O'Daniel and Kok Wai, "Domain Name," 312; Markwell and Brooks, "Broken Links: Just How Rapidly."
29. Lawrence et al., "Persistence of web References," 26–27. Lawrence et al. were later able (with results considerably better than those in our study's 'Lost and Found' follow-up test) to reduce after two searches the number of *lost* items in a random sample of 300 invalid URLs to 3 percent by using search engines, by guessing or browsing, and by settling for highly related information or a formal citation (27–28); Nelson and Allen, "Object Persistence."
30. Nelson and Allen, "Object Persistence."
31. Markwell and Brooks, "Broken Links: The Ephemeral," 106; Markwell and Brooks, "Broken Links: Just How Rapidly;" Koehler, "An Analysis," 173–174; Koehler, "Web Page Change;" Taylor and Hudson, "'Linkrot,'" 275.
32. Koehler, "An Analysis," 173–174.
33. Markwell and Brooks, "Broken Links: The Ephemeral," 106.
34. Taylor and Hudson, "'Linkrot,'"275.
35. Ibid., 275; Koehler, "An Analysis," 173–174.
36. Koehler, "An Analysis," 173–174; Koehler, "Web Page Change."
37. Taylor and Hudson, "'Linkrot,'" 274–275.
38. Koehler, "web Page Change;" Kitchens and Mosley, "Error 404," 471, 473.
39. Koehler, "An Analysis," 172, 173; Koehler, "Web Page Change."
40. Taylor and Hudson, "'Linkrot,'"275.
41. Kitchens and Mosley, "Error 404," 477.
42. Recall that Kitchens and Mosley have suggested that URLs specifying particular filenames were the greater culprit, so one might suggest that, in future, librarians give preference to navigational sites and pages over specific content pages to enhance the longevity of their bibliographies.
43. Markwell and Brooks, "Broken Links: The Ephemeral," 107.
44. Ibid., 106; Markwell and Brooks, "Broken Links: Just How Rapidly."
45. Koehler briefly discusses the difficulties that the menace of 'phantom' web-pages and -sites has caused his automated site/page checking program in "Web Page Change," but readers desiring a more complete discussion should also see: Wallace C. Koehler, "The Management of Web Page Dynamics in Web Catalogs: The Phantom Page Problem," in vol. 1 of *Libraries and Associations in the Transient World Proceedings* (Sudak, Ukraine: International Conference "Crimea 2000", 2000): 211–214. Available: <http://www.gpntb.ru/win/inter-events/crimea2000/doc/tom1/444/Doc3.HTML> [September 24, 2003].