2
Methods and Metrics

What's a liblog? I define it as a blog written by one or more library people (librarians, library staff and fringe types like me) *or* a blog written about library issues—excluding official blogs of all stripes. In a few cases, I've excluded blogs by library people that are clearly not library-related.

You may have seen the term "biblioblogosphere." I used that term for the first in this series of studies. I don't use it anymore because it's too broad and too narrow—the term should include library blogs (blogs that are official library projects) and blogs related to books and publishing wholly outside the library arena, and the term should exclude liblogs that have nothing to do with books as such.

Requirements and Exclusions

To be included in this study, a blog had to meet five requirements and avoid one exclusion:

➢ **Viewable**: I had to be able to reach the blog on the open web, using Firefox, without passwords, at some time between June 1 and September 14, 2010.

➢ **Blog**: It must have at least minimal characteristics of a blog—that is, a stream of individual posts presented in reverse chronological order.

➢ **English**: Most posts (based on inspection of the first viewed page) must be in English.

➢ **Liblog**: Either the blogger(s) must identify themselves as library people (not necessarily librarians) or the blog

must relate to libraries or librarianship. I'm interpreting "library" to include archives and museums.

➢ **As of May 31, 2010**: There must be at least one post dated on or before May 31, 2010.

➢ **Exclusion**: **Official blogs**: Blogs that appear to be official as opposed to personal, including library blogs, corporate blogs and others, are excluded. This is frequently a judgment call—e.g., a blog with an association's name that is clearly written by members without association approval of each post would be included.

For the three somewhat ambiguous requirements (English, liblog, not an official blog), my usual rule is "when in doubt, leave it in." There are certainly some blogs included in this study that might be considered official blogs and a few that are not exclusively English.

Building the Universe

I started with the spreadsheet from the previous study—including liblogs that had been included in 2007-2008 but didn't meet narrower criteria for 2007-2009. Current filters (English, not official, still visible) yielded 563 liblogs and 62 exclusions. My own set of blog subscriptions (in Bloglines) added 43 additional liblogs and nine more exclusions.

Checking and rechecking half a dozen liblog directories, including Salem Press's lists (which yielded a *lot* of liblogs I'd never encountered) added 415 additional liblogs—and 426 more exclusions. At this point, there were just over one thousand liblogs and just under 500 exclusions. Two more directories, totaling 681 entries, yielded 275 possibilities that became 49 more liblogs and 226 more exclusions.

Then I started gathering and updating metrics for those liblogs—and copying blogrolls when the blogrolls appeared to be library-related and weren't so long as to be unmanageable. That process yielded considerably more than 2,000 possible candidates, which boiled down to some 900 blogs I

hadn't already checked—which, after checking, yielded enough new liblogs and exclusions to bring the totals to 1,277 liblogs and 1,308 exclusions.

Finally, I asked for new names on my blog and a couple of lists and accidentally encountered one or two new liblogs that weren't in any of the other sources. One or two liblogs disappeared as I was checking metrics. The final total: 1,304 liblogs and 1,327 exclusions.

Is this the *complete* universe? Almost certainly not. I didn't check every blogroll (that would have involved checking literally tens of thousands of blogs, almost all of them outside the library field). Some liblogs don't show up anywhere: Nobody links to them, they're not in any directory, they're not in any blogroll. I'd *guess* this is at least 90% of the English-language liblog universe as it existed on May 31, 2010, but I could be wrong.

The universe keeps changing. If I rechecked all 1,304 liblogs on, say, October 10, 2010, I can be sure some previously-visible blogs would have disappeared (turning into parking pages, spam blogs or protected blogs, or yielding a "Blog removed" page from WordPress, Blogger and similar hosts)—most likely at least half a dozen and fewer than 50.

What about the Exclusions?

I categorized each exclusion as I checked it. Here's a tabular version of why blogs are excluded, followed by notes on each category—noting that a blog could be excluded for more than one reason, but I only chose one.

Reason	Count
Broken	2
Empty	21
Malware	1
Not a blog	67
Not English	160
New	12

Not library-related	506
Not visible	306
Official	132
Podcast	1
Renamed	118

Table 2.1: Exclusions

Notes on the categories:

➢ **Broken** blogs were there but didn't work, for one reason or another.

➢ **Empty** blogs had pages and possibly banners, but no posts whatsoever.

➢ **Malware** blog—fortunately singular—attempted to infect my computer with some form of malware. I ran a full scan and used multiple disinfection tools, removing three different pieces of malware. (The blog was "Information Knot"; it may since have been fixed or deleted.)

➢ **Not a blog:** These didn't seem to be blogs but had been included in blogrolls.

➢ **Not English:** Most or all posts on the first page were in languages other than English.

➢ **New:** Liblogs where the very first post appeared on June 1, 2010 or later. If another study is done next year, they'll be included. It's heartening to see *some* additional liblogs being created!

➢ **Not library-related**: Neither written by "library people" (or archivists or museum people) nor with any apparent focus on libraries, archives, museums or librarianship. This is the largest category by far, which makes sense: Many libloggers include all their interests in a single big blogroll, and those interests will and should go beyond other liblogs.

➢ **Not visible**: The URL either yields a parking page, a "blog deleted" page, a 404 error or the like, or a request

for a password. While some of these 306 blogs may never have been liblogs, I'd guess most of them are defunct liblogs.

➤ **Official**: Blogs from a library, group or other agency that appear to reflect official viewpoints, rather than blogs hosted by a group that have posts by group members.

➤ **Podcast**: Purely a list of podcasts with no other text.

➤ **Renamed:** Most of these are liblogs that have since changed names (and appear under the new names); some of them are names that appear in blogrolls but were never actual liblog names.

If you add "Not visible" to the set of blogs I examined, you come up with just over 1,600. Is it fair to say there have been 1,600 unofficial liblogs? Probably not—chances are there are others that existed so briefly that they never made it to a directory or blogroll.

Metrics and Completeness

The primary spreadsheet for this year's study has 24 columns, each representing some *direct* piece of information or metric for the blog in each row. There are quite a few more *derived* metrics—e.g., posts per month, comments per post.

With few exceptions, metrics are incomplete: I was either unable to collect information on some blogs or, in some cases, unwilling to spend the time to do so.

Here are the metrics and the number and percentage of blogs for which I was able to record each metric. Metrics appear in the order in which they are analyzed and discussed in this study.

Name, URL, Group and Category

These four metrics appear for all 1,304 blogs—in two cases because it wouldn't be possible to include the blog otherwise, in the other cases because they're assigned metrics.

➢ **Name**: The name of the blog, normally as it appears in the banner or page title. There's one case in which two blogs have exactly the same name, "@ the library"; I've added "[2]" to the younger of the two (the one at librarianwoes.wordpress.com).

➢ **URL**: The last known location for the blog. URLs generally don't appear in this study but will be in the portion of the spreadsheet that may eventually appear as an online table.

➢ **Group**: A number from 1 to 4, assigned based on a blog's activity, visibility and currency. Described further in the next section.

➢ **Type**: Not a detailed breakdown similar to that done in 2008 (e.g., law, science, museums, academic), but a three-way breakdown: "Books" (blogs primarily consisting of book or media reviews), "Technology" (blogs with technological slants) and "Other" (everybody else).

Where, How, Visibility and When

➢ **Country**: Country in which the blogger(s) resided when the blog was checked, when that was clear either from author information or from posts themselves (or when my request for information yielded results). Present for 1,216 of 1,304 blogs: 93%.

➢ **Software**: Blog software, only recorded for seven possibilities. Present for 1,251 of 1,304 blogs: 96%, a higher rate than I would have expected.

➢ **Google Page Rank:** A crude but easy indicator of blog visibility—the extent to which other blogs and websites refer to a given blog. Taken at the first time I checked a blog between June 1 and September 14, 2010. Technically, there's a number for every blog, but "0" is the same as "no information." This report does not *ever* associate a given

blog with its GPR. Greater than 0 for 1,054 of 1,304 blogs: 81%.

➢ **Year and Month**: The year and month of the first post that I could locate or that is claimed in archive lists. Present for all blogs.

➢ **Longevity**: The number of months between the start of a blog and the last post on or before May 31, 2010. Also present for all blogs. (This isn't a derivative figure—you can't just subtract the month and year from June 2010—because many blogs aren't still active.)

➢ **Currency**: How current the most recent post was as of May 31, 2010, grouped into a small number of buckets. Present for all but seven blogs.

Activity

The remaining 13 metrics—and many derivative metrics—deal with blog activity: Number of posts, length of posts, number of comments. "Present" really means "I was able to calculate this without too much effort." What's too much? I wouldn't spend an hour on a metric for one blog; I *might* spend half an hour.

➢ **Total Posts**: Total number of posts from the start of the blog through May 31, 2010. Present for 1,186 of the 1,304 blogs: 91%.

➢ **Count2007**: Number of posts in March-May 2007. In this case, there are three categories: Blogs with countable posts, blogs with no posts, and blogs that didn't exist yet. (A few of the "no posts" blogs might be cases where I couldn't figure out how many posts there were.) 683 blogs have positive numbers (52%), 158 blogs started before June 2007 but have no recorded posts during this period (12%), and 463 blogs started after May 2007 (36%).

➢ **Length2007**: Total length of posts in March-May 2007. Of the 683 blogs with posts in 2007, 624 (91%) have recorded lengths; the others were difficult or impossible to measure, for various reasons.

➢ **Comments2007**: Number of comments in March-May 2007. 550 of the 683 blogs with countable posts had at least one comment (81%). 104 more blogs with countable posts had no comments, either because they weren't allowed or because nobody commented. The rest are cases where the comments couldn't be counted easily (29) or the blog began after June 2007.

➢ **Count2008**: How do you track blogs that have formally ceased or have no new posts? For now, it makes sense to count them as 0 posts, length and comments. For March-May 2008, 810 blogs have positive numbers (62%), 257 blogs existed but had no countable posts (20%), and 237 blogs started after May 2008 (18%).

➢ **Length2008**: 746 of the blogs with countable posts have recorded lengths (92%).

➢ **Comments2008**: 642 of the blogs with countable posts have at least one comment in March-May 2008 (79%). Another 129 with posts had no comments.

➢ **Count2009**: Number of posts March-May 2009. For this period, 876 blogs (67%) have positive numbers, 340 (26%) existed but had no countable posts (including quite a few that were already moribund or dead), and 88 (7%) started after May 2009.

➢ **Length2009**: 790 of the blogs with countable posts have recorded lengths (90%).

➢ **Comments2009**: 664 of the blogs with countable posts have at least one comment in March-May 2009 (76%). Another 148 with posts had no comments.

➢ **Count2010**: Number of posts March-May 2010. For this period, 842 blogs (65%) have positive numbers. The re-

mainder, 462 blogs, either skipped those months or were already moribund or dead.

➤ **Length2010**: 736 of the blogs with countable posts have recorded lengths (87%).

➤ **Comments2010**: 603 of the blogs with countable posts have at least one comment in March-May 2010 (72%). Another 180 with posts had no comments.

Metric	Present	Percentage
Country of blogger	1,216	93%
Blogging software	1,251	96%
Google Page Rank	1,054	81%
Total Posts	1,186	91%
Count2007	683	52%
Length2007	624	91%*
Comments2007	550	81%*
Count2008	810	62%
Length2008	746	92%*
Comments2008	642	79%*
Count2009	876	67%
Length2009	790	90%*
Comments2009	664	76%*
Count2010	842	65%
Length2010	736	87%*
Comments2010	603	72%*

*As percentage of blogs with counted posts for that year, noting that Comment count is for blogs with at least one comment.

Table 2.2: Metrics and Completeness

Metrics will be discussed in more detail in later chapters, as will the many derived metrics that appear in other chapters.

This Year and Last Year

How much broader is this study than last year's selective study? It varies—anywhere from half again as broad to more than twice as broad. The following table may help—and, for comparability, I've included blogs where comments were countable but there weren't any.

Metric	2010	2009	Percentage
Total blogs	1,304	521	250%
Count2007	683	449	152%
Length2007	624	412	151%
Comments2007	654	441	148%
Count2008	810	486	167%
Length2008	746	452	165%
Comments2008	771	476	162%
Count2009	876	434	202%
Length2009	790	394	201%
Comments2009	812	415	196%

Table 2.3: Blog counts for metrics in 2010 study compared to 2009 study

What: Types and Groups

In *The Liblog Landscape 2007-2008*, I attempted to break down blogs by affiliation of bloggers—e.g., academic library, vendor, law library or librarian. I dropped that attempt in 2009; in most cases it didn't seem to add useful information.

As I was adding new blogs this year, I encountered quite a few that seemed to fall into a category or type, a type that wasn't well represented in previous versions: blogs devoted primarily to book reviews and news (or other media reviews and news). It was clear that these have different characteristics than most liblogs. I wondered whether "technology-oriented" blogs—ones that specifically note "2.0" or "tech" in

the name or banner, or are otherwise fairly obviously technology-oriented—might differ from the run of blogs as well.

Type b blogs (books and other reviews) make up 9% of the universe, 115 in all—but that's 5.2 times as many as the 22 type b blogs in the 2009 study.

Type t blogs (technology) make up 31% of the universe, 405 in all, and that's 2.5 times as many as the 181 type t blogs in the 2009 study, almost exactly the norm.

Everything else—"type o" blogs (other)—make up 60% of the universe, 784 in all—2.3 times as many as last year.

You'll see breakdowns for type b and type t blogs in some metrics where there seems to be a significant difference. One clear difference for type b blogs shows up in tables 2.4 and 2.5: Book review blogs tend to be strong survivors.

Groups

I developed a rough grouping of blogs based on activity, currency and visibility before I started gathering data—in part because I originally planned to do a superficial "broad look" at the whole universe and restrict the deeper look to a subset of blogs somewhat comparable to (but even more narrowly defined than) the 2009 study.

As time went by, I decided to apply the broad look to the whole universe (although, for obvious reasons, some metrics don't apply to one of the four groups) and to increase the number of groups from three to four.

Here's how the four groups are defined—based entirely on observation in June-September 2010:

➢ **Group 1**: These blogs must have Google Page Rank of 4 or higher and must have at least three posts between March 1, 2010 and May 31, 2010. 443 blogs (34% of the total) fall into this group. I'd call these the **core blogs**.

➢ **Group 2**: These blogs *either* have a Google Page Rank of 3 and at least one post between March 1, 2010 and May 31,

2010, *or* have a Google Page Rank 4 or higher, but have only one or two posts during the quarter. 207 blogs (16% of the total) fall into this group. I'd call these **less active visible blogs**.

➢ **Group 3**: These blogs don't fall into Groups 1 or 2 but had at least one blog within the year ending May 31, 2010—that is, the most recent post is on or after June 1, 2009. This is a hodgepodge, as it combines some very active blogs with no GPR (which can happen for several reasons—e.g., few of the *Library Journal/School Library Journal* blogs have GPRs) with a bunch of blogs that either have very little activity or have been abandoned within the past year. 364 blogs (28% of the total) fall into this group. Call these **also alive**.

➢ **Group 4**: These blogs had at least one post visible on the web—but no posts between June 1, 2009 and May 31, 2010. (A few of them have come back to life since May 31.) I'd guess 80% to 90% of these blogs are defunct, but haven't actually been removed from the web. Call these **mostly defunct**.

Table 2.4 breaks blogs down into groups (columns) and types (rows):

Group	1	2	3	4	Total
Type					
b	57	14	35	9	115
o	235	139	218	192	784
t	151	54	111	89	405
Total	443	207	364	290	1304

Table 2.4: Blogs by group and type

The combination of Groups 1 and 2, 650 blogs, is roughly comparable to the 2009 study group of 521 blogs (except that the 2009 study excluded GPR 3 and included quite a few blogs now regarded as official).

It's interesting but coincidental that Groups 1 and 2 and Groups 3 and 4 are each almost precisely half of the universe. Group definitions were made without regard to actual numbers before the universe of blogs was understood.

Table 2.5 is the same breakdown as Table 2.4—but by percentages. It's important to note that each percentage in the first four columns is a percentage of the total *for that row*—that is, 50% of type b blogs are in group 1.

Percentage	1	2	3	4	Total
b	50%	12%	30%	8%	9%
o	30%	18%	28%	24%	60%
t	37%	13%	27%	22%	31%
Total	34%	16%	28%	22%	100%

Table 2.5: Blog percentages by group and type

As with types, you'll see some group breakdowns in metrics where it appears useful, most often to differentiate Group 1 from the rest of the universe.

Averages and Medians

Averages are mostly meaningless when discussing liblogs: The universe is far too heterogeneous. I do mention averages whenever it seems appropriate or when there's a need to point out just how heterogeneous the liblog landscape actually is. When used, "average" carries its standard meaning: The sum of all the values for a given metric divided by the number of items carrying that metric. So the average of 1, 89, and 900 is 330—and that's about as meaningful an average as most liblog metrics would yield.

Medians can sometimes be more meaningful. The median of a group is the value at which as many members of the group have a higher value as have a lower value—it's the center as you're counting.

Quintiles and Other Notes

I use *quintiles* extensively in looking at metrics. Quintiles break a population down into five groups based on a particular metric. So, for example, when looking at number of posts in March-May 2010, the first quintile is the (roughly) 20% of blogs with the most posts, the second quintile the 20% of blogs with more posts than average (60%-80%), the third quintile (40%-60%) blogs with roughly average number of posts, fourth quintile (20%-40%) blogs with fewer posts than average, and fifth quintile (0%-20%) blogs with the fewest posts. I say "roughly" because it makes no sense to split two blogs with the same number of posts between two quintiles, so quintiles can vary in size.

Why quintiles? Because they provide a compact picture of a universe that's not *too* compact—one that allows for a meaningful "average" range. When you split a population into four parts, the average is split badly. No matter how little an individual differs from the median, it's forced into the second or third quartile. With quintiles, there's a broad "average group"—the third quintile.

Most quintile tables provide the median for each quintile. That allows you to look at, for example, the top 10% in a category: The median for Quintile 1 marks the breakpoint between the top 10% and the second 10%.

Triplets

Quintiles are great for looking at overall distribution and for pointing up how an individual blog fits into the universe of blogs. They're not so great for certain other purposes, including change patterns involving more than one metric or more than one change period.

When looking at some of those patterns, I'll use triplets rather than quintiles, where a triplet divides the universe into three parts: Significantly increased, Roughly unchanged,

and Significantly decreased. I use 20% as the cutoff point for significance. So, for example, for post counts from 2009 to 2010, the triplets would be blogs that had at least 20.1% more posts in 2010 than in 2009, those that had 80% to 120% as many, and those with 79.9% or less.

Why March-May?

When measuring blog activity, you need to strike a balance between a long enough sample period to be meaningful and a short enough period so it's feasible to do the metrics. Three months seems to be a good compromise.

You want a three-month period when most bloggers are reasonably likely to be active, regardless of the type of library or activity. That argues against summer months (June-August) and heavy vacation months (November and December).

January is problematic, both because it's partly a vacation month and because of ALA Midwinter, one of two megaconferences that can skew blogging activity. That leaves two possible three-month slots: February through April or March through May. The latter period is a little longer, and in the U.S. each of those three-month periods has one three-day weekend, so I chose the longer quarter. (Yes, there are significant library conferences in the March-May period, but that's true for every quarter.)

(If you read *But Still They Blog*, you've read this section already—it's copied with almost no revision.)

The Rest of the Study

Is this liblog landscape the totality of English-language liblogs? Absolutely not: Some have disappeared and some just don't show up either on blogrolls or in directories and lists.

I do believe thus study includes most of them, probably at least 90% of those that are still visible. If so, it's reasona-

ble to suggest that there are around 800 to 900 active liblogs at any time in recent years.

The rest of the study looks at the landscape as it appears now and, in some cases, as it's changed over time.

First we'll look at the software used for liblogs, where liblogs came from, when they began and how long they've lasted. Then we'll look at *overall* posts—a new metric—in two ways: The totality of posts and the average posts per month over the life of blogs.

After that, three chapters look at one category of metric as it appears during the four March-May quarters studied to date and as it's changed. First there are posts. Then come blog length—and, more significantly, the average length of each post in a blog. Finally, there are conversations: The total comments for a blog and the average number of comments per post.

Additional chapters look at patterns of change involving the three key metrics (posts, length per post, comments per post), standouts and standards—blogs that stand out for one reason or another, and correlations and miscellaneous notes.

If you're reading this, you're reading Chapter 2 of *The Liblog Landscape 2007-2010* in late draft form as it appears in the December 2010 *Cites & Insights*. To see Chapter 1, to see the final version of this chapter, for an index to blogs named in this book and to help support future research, buy *The Liblog Landscape 2007-2010* in book form (trade paperback or PDF download) when it's announced in late 2010 or early 2011.