Virtual Dave Lankes

Pre-Prints

http://www.DavidLankes.org

TITLE: Collecting Conversations in a Massive Scale World
AUTHOR(s): R. David Lankes
PUBLICATION TYPE: Journal
DATE: 2007
FINAL CITATION: "Collecting Conversations in a Massive Scale World" Lankes, R. David (forthcoming). Library Resources & Technical Services
KEYWORDS: Participatory Librarianship, Massive Scale, Credibility, Conversation Theory

Please Cite as:

Lankes, R. David (forthcoming). "Collecting Conversations in a Massive Scale World" Library Resources & Technical Services

Collecting Conversations in a Massive Scale World

R. David Lankes

Director, Information Institute of Syracuse &

Associate Professor

School of Information Studies

Syracuse University

Abstract

This article highlights the growing importance, challenges and opportunities of massive scale computing as they relate to libraries. Massive scale computing is defined as the predictable wide scale availability of computing power, storage and network speeds at an immense levels. The article, based on a presentation to the Association for Library Collections & Technical Services 50th anniversary conference, argues that libraries must help shape the emerging world of nearly unlimited computing capacity, and further more outlines an approach to library service in such an environment: participatory librarianship.

Introductory Note

The following paper was derived from a speech given at the Association for Library Collections & Technical Services 50th anniversary conference. As with any talk, there is some level of informality, and some large ideas given short treatment. Where possible, citations have been provided so that the reader may delve more deeply into the topics mentioned.

It is presented in a time of near universal mission seeking by libraries. Where once libraries were arguably the home to the largest information stores in the world, today the floodgates of digital data have been opened and libraries are now seen as a much more selective center of documents. This focus on books and formal documents, while they have served the library very well in the past, begin to inhibit the libraries evolution. The tools developed over the past two hundred years focused on items and artifacts (books, albums, etc), have begun to show both their age and their rigid assumption in a world of real-time information production and distribution.

It is nothing unusual that our field, or any field, must engage in a series of self-reflections and justifications of its purpose and tools. It is the sign of an active and important pursuit that questions arise. Libraries have had such mission discussions as they moved from ivory towers to the public, as video challenged books as central modes of information disseminations, and now it is happening again as the field struggles with digital items that do not neatly fit the definition of "document" or "item." I argue that such changes and challenges need to be embraced, and embraced by returning to libraries core mission: the facilitation of knowledge acquisition in our

communities. And this is what I told to ALCTS, starting with an example from the transportation industry.

<1>Gigabyte per Mile

In the process of a Transportation Research Board study on information management in the transportation industry), several panel members observed that soon every mile of road will generate a gigabyte of data a day.[1] This data will come from road sensors embedded into asphalt to detect temperature for winter salting, real-time traffic data from roadway cameras, weather information, toll data from RFID (Radio-Frequency Identification) expressway systems, car black boxes, and a myriad of other data sources. It is assumed that this will become a gigabyte an hour as more and more technology finds its way into our vehicles and management systems (GPS data, real time environment monitoring, etc.). As there are 3.5 million miles of highways in the U.S. that would be 3.3 petabytes of data per hour, or 28 exabytes per year.

Some readers may not be familiar with an exabyte. It is the name for a very large volume of storage like megabytes, gigabytes (1024 megabytes) and terabytes (1024 gigabytes); technically 2^60 bytes. Table 1 will give the reader some sense of the scale involved.

Table 1
Data Powers of Ten

Byte	1 byte: a single character;
Kilobyte	2 Kilobytes: A typewritten page;
Megabyte	2 Megabytes: A high resolution photograph;
Gigabyte	2 Gigabytes: 20 meters of shelved books
Terabyte	2 Terabytes: An academic research library
Petabyte	2 Petabytes: All US academic research libraries;

Exabyte	5 Exabytes: All words ever spoken by humans.
Zettabyte	
Yottabyte	

Derived from University of Berkeley, "How Much Information?" (2000), http://www2.sims.berkeley.edu/research/projects/how-much-info/datapowers.html (accessed Sept. m16, 2007).

What the reader needs to realize is that each succeeding row in the table, from megabyte to gigabyte to terabyte and so forth, is an exponential increase. By and large people do not think in exponential terms. Gladwell uses the analogy of folding paper to demonstrate just how big the shifts involved in exponential change are. Imagine you have a huge piece of paper.[2] While the paper is large in terms of its width and height, it is only .01" thick. You fold it in half. You then fold it in half again fifty times. How tall would it be? Many people might say as thick as a phone book, or get really brave and predict as high as a refrigerator. The actual answer is approximately the distance between the earth and the sun.

How can this be? Certainly if I stack fifty pieces of paper on top of each other ,the stack would not be that large. However, stacking separate sheets is a linear progression, that is not what you accomplished by folding the paper. With every fold, you doubled the thickness of the paper. So one fold, the paper is twice as thick as when you started. With the second fold the paper is four times as thick—the next fold is eight times as thick and so on. In first few folds you do not see a major increase, but at about fold forty you are doubling a mile. We are not used to thinking in terms of exponential growth because most things we deal with grow linearly. However, technology is not.

<1>Predictable Change

In 1965 computer pioneer Gordon E. Moore predicted that the number of transistors that could be fit on a chip (roughly equivalent to the speed at which the chip could process information) would double every eighteen months.[3] The prediction has become so reliable it is referred to as Moore's Law. The law is an exponential change just like the paper folding. Computers have not just gotten faster over the past decade, they have gotten exponentially faster. What is more, currently makers of storage technologies—hard drives, solid state flash memory and the like—are exceeding Moore's Law. The emergence of massive scale computing in our every day lives is a predictable change unlike the Web.

The Web and associated wide spread Internet penetration was a discontinuous event. No one could truly predict a world where URLs come with every can of soda, or where an online search company would become one of the biggest corporations on the planet. Libraries can be excused for taking some time to adjust their service models to such an unpredictable and disruptive force. Yet libraries, by and large, have adapted to the new reality. Be it providers of access, guiding online research, supporting distance education, providing virtual reference, or developing metadata schema, libraries have adapted to this change and continue to do so.

The question now lies before the library community: will massive scale computing be another disruptive force, or, as it is a predictable change, will libraries proactively engage in the massive scale computing world? This question is not theoretical, nor is it a question that can long be delayed. Consider that following quote from *Wired Magazine*:

> Ask.com operations VP Dayne Sampson estimates that the five leading search companies together have some 2 million servers, each shedding 300 watts of heat annually, a total of 600 megawatts. These are linked to

> hard drives that dissipate perhaps another gigawatt. Fifty percent again as much power is required to cool this searing heat, for a total of 2.4 gigawatts. With a third of the incoming power already lost to the grid's inefficiencies, and half of what's left lost to power supplies, transformers, and converters, the total of electricity consumed by major search engines in 2006 approaches 5 gigawatts . . . almost enough to power the Las Vegas metropolitan area—with all its hotels, casinos, restaurants, and convention centers—on the hottest day of the year.[4]

Consider also that many universities, companies, and even primary and secondary schools have run out of power to add new computing equipment. Either their own electrical infrastructure cannot handle the load of computing, or their municipalities literally have no more power to send.

<1>Options?

So how can the library community respond to the emerging reality of massive data stores, unimaginable processing power, and super fast networks? In particular how will libraries respond when the limitations of storing the world's information indefinitely disappears, and the production of new data and information grows exponentially from today? Let us explore some options.

<2>Option 1: Ignore It

No one said the library has to take on every challenge presented it. In fact, many criticize libraries for taking on too much. Perhaps the problem of massive scale computing and storage is not a library problem. Certainly for those who argue that libraries are in the business of literacy and cataloging, there is plenty to do with published documents.[5] After all, libraries are plenty busy with published documents and digitizing historical documents. Why add the problem of real-time information stores and digital items that don't remotely look like documents? Furthermore, there are already plenty of other disciplines lining up to tackle this issue. From e-commerce to computer science to individual industry sectors like transportation and medicine many have begun to acknowledge the problem of massive scale computing. The National Science Foundation and National Endowment for the Humanities alike have begun "cyberinfrastructure" initiatives. In addition the computing industry has certainly taken care of these problems to date. With faster processors, smarter software, and bigger hard drives, no doubt Apple, Microsoft, or the other industry players can solve these issues.

The answer to "why not ignore it," I argue, comes down to a simple ethical consideration. If libraries do not address these issues with their foundation of praxis and principles, the consequences for society and the field of libraries itself, could be grave. Look at the largest portal and search engine companies. When they partner with libraries, such as in large-scale digitization efforts, these commercial organizations gain credibility, and have negotiated safeguards of the material they are digitization (scans being re-deposited with libraries for example). However, look at the data these organization store on users. How comfortable is the library profession with these data stores when search engine providers cooperate with

governments (domestic and abroad)? Will principles closely aligned with civil liberties and privacy be preserved? Will data stores of unique resources beyond the current library collections be made widely accessible? The answer is obvious—only as long as the business model is served.

The ultimate result may well be the commercialization of data stewardship in the massive scale world. We have already seen how well that works with scholarly output and journals. To be sure, I am not arguing that libraries must do it all, but they must be a vital part of the massive scale landscape. If we truly value our principles of privacy, access, and so on, we must see them as active, not simply passive. We cannot, in essence, commit the sin of omission by not engaging the massive scale world, and allowing access and privacy to be discarded or distorted. We should be working to instill the patron's bill of rights throughout the information world, not simply when they enter our buildings or Web sites.

<2>Option 2: Limit the Library

A closely related strategy to ignoring the issue is to acknowledge the issue and redefine our mission around it. In essence, libraries are in the knowledge business, and that is now going to be defined as document-like objects, with some sort of elite provenance, and well synthesized. In fact arguments have been made that sound very close to this approach. The distinction is sometimes subtle, as in this quote from Crawford and Gorman:

> Libraries and librarians serve their users and preserve the culture by acquiring, listing, making available, and conserving the records of humankind in all media and by providing services to the users of those records.[6]

Here, while the mission sounds expansive, the key comes in defining what a "record of humankind" is. Do large-scale datasets fit into this category? What about blog entries or reference inquiries? Certainly they appear not to in Gorman's later essay "Web 2.0: The Sleep of Reason."[7]. Here Gorman bemoans "an increase in credulity and an associated flight from expertise."[8] The problem, of course, has always been in defining and agreeing on an expert. Such notions are almost always situational (for a much more detailed discussion on this issue see Lankes, (forthcoming)).[9]

However, there is a much deeper problem in this line of logic. Namely. it pits two long-standing practices and ideals in librarianship: selection and intellectual freedom. Selection and weeding is common library practice. It grew out of resource limitations. Shelf space, book budgets, availability, use of jobbers, and the like are all about existing in a world of scarcity. All of these resources in the physical world constrain the size and scope of the collection. Not since the days of monks and illuminated manuscripts have libraries been convincingly able to collect it all. Today the concept of "comprehensive" is often limited to a serial run or manuscript series.

Yet in a truly digital world the growing prospect of cheap storage makes digital artifacts *very* different. While licensing and cost may still restrict access to some items, collecting massive, effectively limitless, digital items makes the selection due to scarcity argument all but moot. Imagine an academic or school library collecting every paper (including every draft and note) ever written by all of its students. Imagine every public library collecting video and

minutes and audio from every public meeting held. The old arguments of not enough room to accomplish such tasks are clearly disappearing. Certainly by having a library collect and disseminate such information we are providing free and open access to information. Whether we should, whether it is worth doing so is no longer a selection from scarcity debate. It becomes a selection by choice debate. Can libraries choose what to collect and still say they are providing free and unencumbered intellectual access to these materials? In a massive scale world, libraries will have to choose between these ideals.

<3>Option 3: Catalog it All

Some have argued that cataloging lies at the heart of librarianship.[10] I and many others take issue with the argument equating "human intervention for the organization of information" solely to cataloging, it is hard to refute the more general concept that information organization lies at the heart of the profession. Why not then extend the current praxis of the field, i.e., metadata generation, to the growing mass of digital information?

It is a pretty commonsensical argument that the library field (or indeed any given field) is unable to provide the raw person power behind indexing the world of networked digital information. However, we also have some pretty good empirical reasons to show this is not an acceptable means of proceeding. The first is that as a field we have already tried this. From early OCLC experiments with CORC (Cooperative Online Resource Catalog**)** to the Librarians Index to the Internet (claiming over 20,000 sites indexed), librarians have tried to selectively catalog the net. They all cite problems of timeliness and a rapidly changing Internet environment (catalog it today, the page will move tomorrow) in trying to catalog the world.

Ignore the problems of shifting pages and dynamic content, and suppose for a minute that every page on the Internet was not only static, but never changed its location. In 2005 Yahoo! estimated it indexed 20 billion pages.[11] If we had our 65,000 American Library Association (ALA) members spend one minute per record indexing these pages, the good news is that the entire Internet could be indexed in a little over seven months. The bad news is that those ALA members would have to work the seven months straight without eating, sleeping, or attending a committee meeting. At the same time Google was claiming its index was three times as large.

The fact is that the Internet is, however, very dynamic. Blogs, gateway pages, news outlets, and other dynamic content represent a growing portion of the Web. If all of those ALA'ers did decide to spend seven months cataloging the Web, they would have to start in the eighth month doing it all over again. Of course, they might also want to spend sometime on the four billion new pages created each year also (using a conservative estimate from OCLC's growth data.[12]

All of this debate, however, ignores the most interesting aspect of massive scale computing—the invention of whole new records that defy traditional cataloging. Take, for example, gigapixel images. According to the Gigapxl Project:

> It would take a video wall of 10,000 television screens or 600 prints from a professional digital SLR camera to capture as much information as that contained in a single Gigapixil exposure.

Imagine an historian creating a directory of gargoyles on the façade of the Notre Dame cathedral. Instead of taking a series of images of each sculpture, the historian simply takes four

gigapixel images (one for each face of the building). Any user of the directory can zoom in from the entire front of the cathedral to any individual gargoyle at high resolution from a single image. How does one catalog that image? As Notre Dame? A Cathedral? A collection of gargoyles? What about a later scholar who uses the same image to explore the stained glass, or construction, or weathering of the façade or any number of other details that can be explored in the image. At such high resolutions what is foreground, what is background, what is predominant, or what is detail becomes messy at best.

<2>Option 4: Embrace It

I obviously favor the option of engagement. In fact, I would further argue that it is the ethical responsibility of library and information science education to prepare librarians for the world of massive scale computing. By not preparing future information professionals to deal with terabytes of data per second, we are limiting their ability to live up to the ideals of the profession and the needs of the future (and many current) patrons.

In order to embrace massive scale computing in libraries we must:

- *Expand and Enhance Current Library Practice.* As previously discussed, librarians must become conversant in not only processing elite documents, but real time information as well.
- *Go Beyond a Focus on Artifacts and Items*. As will be discussed, books, videos, even Web pages themselves are simply artifacts of a knowledge creation process. To concentrate on containers and documents is to be overwhelmed. By focusing instead on

knowledge creation itself and directly incorporating patron knowledge, librarians should be better able to manage and add value to the tsunami of digital data being created.

- *See Richness and Structure Beyond Metadata*. To move from processing containers to capturing and organizing knowledge means going beyond traditional methods of classification and cataloging. Too often librarians enter a discourse community, and drive it to taxonomy creation when the vocabulary, the very concepts, of the discourse community are still formative. Instead, librarians need to look to other structures in knowledge products and the creation process such as provenance, linking (citations), and social networks to provide a useful method of information discovery and enrichment.
- *Change at the Core of the Library*. All of this needs to be done at the core of library service, not as some new service, or by adding new systems and functions to an already labyrinthine array of databases, catalogs, and software.

There is now an effort to evolve our understanding of librarianship to accommodate these shifts in approach, and this should help the field engage in the world of massive scale computing.

<1>Participatory Librarianship

Simply put participatory librarianship recasts library and library practice from the fundamental concept that knowledge is created through conversation. Since libraries are in the knowledge business, they are, therefore, in the conversation business. Participatory librarians approach their work as facilitators of conversation. Be it in practice, policies, programs, or tools

(or all of these), participatory librarians seek to enrich, capture, store,and disseminate the conversations of their communities.

The other implication of this approach is that books, videos, documents are by-products of conversations. That is not to say they are unimportant, rather acknowledges they are only a pale reflection of the knowledge creation process. By the time you read this article, for example, it has already been re-written and edited numerous times. By the time the ideas are encoded into words, they have been debated and discussed by a wide spectrum of people. The citations at the end give only an idea of the resources used to develop these arguments (the ones written down and easily addressed). The article will also no doubt lead to a few discussions and disagreements after it is published. Yet, it is this written document that will be indexed in the databases. The rich conversational space around it is lost.

The idea of conversation in librarianship or a "conversational space" around articles is not all that new. Bechtel talked about how scholarly communications should be taught as an ongoing conversation in information literacy programs.[14] Conversational organizational approaches can also be seen in: citations and scholarly communication, law and precedents; bibliometrics, Web of Science, reference, and special collections, and it plays a large role in collection development. In may ways, libraries have been in the conversation business, they have simply developed technologies centered on items—so much so we are now struggling to re-capture the conversation in initiatives such as federated searching and *FRBR* (*Functional Requirements for Bibliographic Records*).

Now turn this problem around for a moment. Let us say that we could capture this conversational space. We would have audio files of class conversations, video of presentations, the full text of the articles cited (including the citations used in those articles hot linked), drafts,

editor's notes—the whole work. Approached as items, each would need a catalog record, and all might be available in the catalog. Yet what holds all of them together as a conversation? In fact, the conversational aspects of this collection of artifacts exists between the catalog records themselves. It is the relationship of items, not the items. This is the kind of information we capture in an annotated bibliography.

If in addition to capturing the items, we captured the relationships, how might that work? Imagine now finding this article online. Once there, you should be able to instantly find the rest of the items. Click—you see a previous draft. Click—there is a citation. Click—here is another article by that author. You are now surfing the conversation itself. It also allows you to rapidly find lots of heterogeneous data. Click on this article and see the text, find a graph and click on it. Up pops access to a large dataset. Run some new analysis on the data and post it. Now someone finding your article can find both the original dataset and the original article that was published. It is in the relationships between items we gain navigation, not in the items themselves.

As a field, we must think in threads. The way to handle a terabyte of data per second is not to try and catalog items in less than a second, it is to know what thread the new terabyte extends. "Oh, this is more weather data from NOAA, I'll attach it to my NOAA thread." Once available scientists, students, and the general public can use that new dataset as a starting point for yet a new thread.

Take our gigapixel image of Notre Dame. The image is simply one item in a thread about gargoyles as created by the author. The same item, however, can also become the starting point for threads by the architect, historian, theologian, etc. Furthermore, by finding any point in the conversation about Notre Dame, be it architectural, historical or spiritual, you can find any other conversation.

If this begins to sound like the Web itself, you are right. However, imagine imbuing the web with the ideals and tools of librarianship. These threads we create can incorporate fundamental concepts such as authority files. The search tools and "thread" (annotation) tools can both preserve privacy, and provide new structures for the library community to capture and add value to.

<1>Conversations: It Takes Two

So by organizing materials into threads and capturing and adding value to the relationships between items, the library can begin to approach massive scales of information. However, just as with trying to catalog the world of digital information, creating and capturing threads can quickly overwhelm the resources of professional librarians. More to the point, with networked technology we want to capture these threads at the point of knowledge creation, with the authors of ideas. In order to o this, we must expand our systems and services to truly incorporate our patrons into them.

In the library science field we have seen an evolution in thinking about the relationship between systems and users. Early computer systems were designed by programmers, and more reflected the system designer than those the system was intended for. This so called system view was challenged, and eventually supplanted (at least rhetorically) by a user-based design paradigm.[15] In the user-based approach, the user's needs and habits needed to be well understood and then reflected in the systems we created. However, today we see a further evolution to truly user systems. In today's spate of social Internet tools, the systems only provide a sparse framework of functionality for users to populate and direct. Wikis, blogs, video sharing

sites, and the like have shown that when users construct the system around themselves they gain greater ownership and utility. We call these participatory systems.

Participatory systems and participatory librarians, do not seek to construct a system of functions and information and then bring the users to them, but rather seek to support users as they construct their own systems and information spaces. Once again, rhetorically, this fits well with the rhetoric of librarianship. After all, from reference to collection development to cataloging (in the concept of literary warrant), we claim the users direct our services. Yet, look at the systems we use to instantiate these ideals. The catalogs we provide only accept queries from users, not actual documents. In reference, we have a conversation between librarian and patron, not patron and patron. It is time to take our ideals and make systems that reflect that the library is an agent of the community, not simply a service to it.

In other venues, these ideas are much more fully developed, and I recommend to the interested reader seeking out the more fleshed out discussions of participatory librarianship.[16] For now, let me simply state that to be of the community means that you have trust in your patrons and they have voice. To be a service to a community implies a paternalistic relationship and a separation.

<1>Recommendations and Conclusion

Libraries must be active participants in participatory networking. This must be done at the core of the library, not on the periphery. Anything less simply adds stress and stretches scarce resources even further. The reason we should be looking at technologies such as blogs and Wiki's, is to: get closer to the community and knowledge generation; and to make all of our

library systems more inclusive of community. By thinking in threads, and using the social intelligence of our service community the library profession is actually well poised to take on the world of massive scale computing.

However, the library field will only thrive in the massive scale world is to engage the ideas and current massive scale stakeholders. To ignore the implications of massive scale computing is dangerous. It abdicates serious decisions and consequences to others who do not have our experience and firm principles. Participatory librarianship is an opportunity not only enhance the mission of the library, but proactively to position librarians at the forefront of the information field . . . where they belong!

References

1. Transportation Research Board, Committee for a Future Strategy for Transportation Information Management, *Transportation Knowledge Networks: A Management Strategy for the 21st Century,* TRB Special Report 284 (Washington D.C.: 2006), www.trb.org/news/blurb_detail.asp?id=5789 (accessed Sept. 15, 2007).

2. Malcolm Gladwell, *The Tipping Point: How Little Things Can Make a Big Difference* (San Francisco: Back Bay Books, 2002).

3. Gordon E. Moore, "Cramming More Components onto Integrated Circuits," *Electronics* 38, no. 8 (April 1965), http://download.intel.com/museum/Moores_Law/Articles-Press_Releases/Gordon_Moore_1965_Article.pdf (accessed Sept. 16, 2007).

4. George Gilder, (October, 2007) The Information Factories. www.wired.com/wired/archive/14.10/cloudware.html?pg=1&topic=cloudware&topic_set (accessed online Sept. 14, 2007).

5. Martha M.Yee, with a great deal of help from Michael Gorman, "Will the Response of the Library Profession to the Internet be Self-immolation?" online posting, July 29, 2007, JESSE, http://listserv.utk.edu/cgi-bin/wa?A2=ind0707&L=JESSE&P=R14797&I=-3.

6. Walt Crawford and Michael Gorman, *Future Libraries: Dreams, Madness and Reality*. American Library Association; (Chicago: American Library Assn., 1995), 120.

7. Michael Gorman, "Web 2.0: The Sleep of Reason. Part I ," online posting, July 11, 2007, Britannica Blog, http://blogs.britannica.com/blog/main/2007/06/web-20-the-sleep-of-reason-part-i/ (accessed Sept. 15, 2007).

8. Ibid.

9. R. David Lankes, Joanne Silverstein, and Scott Nicholson, Participatory Networks: The Library as Conversation *Information Technology and Libraries* forthcoming.

10.Yee, "Will the Response of the Library Profession to the Internet be Self-immolation?"

11. John Battelle, (September 26, 2005). Google Announces New Index Size, Shifts Focus from Counting," online posting, Septe. 26, 2005), John Battelle's Searchblog, http://battellemedia.com/archives/001889.php (accessed Sept. 15, 2007).

12. OCLC, Size and Growth Statistics, www.oclc.org/research/projects/archive/wcp/stats/size.htm (accessed Sept. 15, 2007).

13. Gigapxl, The Gigapxl Project (August 2007), http://www.gigapxl.org (accessed Sept. 15, 2007).

14. Joan M. Bechtel, "Conversation: A New Paradigm for Librarianship? *College & Research Libraries* 47, no. 3 (May 1986): 219-224.

15. Brenda Dervin and Michael Sanford Nilan, " Information Needs and Uses," *Annual Review of Information Science and Technology* 21 (1986): 3-31.

16. R. David Lankes, Joanne Silverstein, and Scott Nicholson, Participatory Networks: The Library as Conversation (commissioned technology brief for the American library Association's Office of Information Technology Policy, 2007), http://ptbed.org (accessed Sept. 16, 2007).