The Information Explosion (and its implications to the Future of Publishing)

by Thad McIlroy, The Future of Publishing, August 15, 2010

Summary

> "The Internet easily defeats advanced filters, delivering millions of words per second to brains that can process only 10 words per second."
>
> — Peter J. Denning, Director of the Cebrowski Institute for Information and Innovation and Superiority at the Naval Postgraduate School in Monterey, CA

> "It's a vicious circle," Sabrina said. "People are going deaf because music is played louder and louder. But because they are going deaf, it has to be played louder still."
>
> — From The Unbearable Lightness of Being
> © 1984 by Milan Kundera

> "Facts are the enemy of truth."
>
> — Miguel de Cervantes, Man of La Mancha

The ongoing information explosion has created a situation where it's now impossible for any one person to stay up-to-date with the changes in any topic area (unless that topic is perhaps so minute in its focus that only a few dozen are following it and contributing to the existing body of knowledge). Not surprisingly, I cannot find statistics on what those topics would be. I do recall years ago a learned colleague of mine mentioning a friend whom he said was the world's leading expert on the "semiotics of the circus." Indeed I see today that if I Google (with quotation marks) "semiotics of the circus" I find only 28 entries, three of which reference an earlier version of this article. No doubt there are additional references in academic texts and journal papers (*Google Scholar* lists 6) but I do believe that should this subject grab your interest, you may indeed find that you could reach a point where you were comfortable that you were aware of all that is known on the topic.

Now try and become all-knowledgeable about the future of publishing. I spend a great many hours in this Sisyphean quest, but feel like I lose ground with each moment spent.

The very interesting author Gabriel Zaid, discussed below, highlights the dilemma:

> "Books are published at such a rapid rate that they make us exponentially more ignorant. If a person read a book a day, he would be neglecting four thousand others, published the same day. In other words, the books he didn't read would pile up four thousand times faster than the books he did read, and his ignorance would grow four thousand times faster than his knowledge."

Data, Information, Knowledge, Understanding and Wisdom

I'm not alone in making a distinction between information, knowledge and wisdom. More recently I see that some pundits add "data" to the front of the list, and something called "understanding" before "wisdom."

The message is the same whether you live by just three categories or by four or five. At the core of the information explosion is just numbers and provable facts — raw and bare — still subject to interpretation and augmentation. When you are able to find context, then knowledge may result. If you have a brain that can interpret this knowledge, and extend the knowledge beyond its obvious implications, you may become wise. Few do. But then wisdom is not a commodity, as information is.

While we can access the thoughts and writings of the wise from the Internet they don't come with certificates of wisdom issued in our names. Just because people read a message, doesn't mean that they understand it. There is an inherent risk in all forms of communication.

It's clearly our good fortune that there is a great deal of data/information available on the Internet. And of course, as with this Web site, it is possible to take advantage of much of this free information. Can we find "knowledge" on the Internet? Perhaps. Wisdom? Less frequently. Nonetheless, there has never been a richer and deeper accessible source of data and information than on the Internet.

But, sadly, there is now far too much information at our disposal. And so we need to approach our quest for data, information, knowledge, understanding and wisdom with careful planning.

The plan must address two challenges. The first is simply trying to improve our methodology for seeking data and information, and transforming that into knowledge, understanding and wisdom. The second is more mechanical and matter-of-fact. We need to learn how to filter the data that only distracts so that we can get on with the task of understanding.

Why the Information Explosion Matters to the Future of Publishing

1. The last several decades have seen an essential shift in the raw material of knowledge, of education – indeed of the creation and expression of culture. Moving from a position of never knowing if you had the answer, or whether the answer was available, or whether you were just searching the wrong sources, we now find ourselves in the reverse position. We find thousands of links on Google, but still must ask at each link: Is this information I have uncovered an *authoritative* or *definitive* word on the subject, or should I keep looking? Is the source to be trusted (c.f. Wikipedia)? How do I assess this possibly tangential or inaccurate information?

2. The authority of the author and publisher are thrown into question. If we can question our own methods of seeking information and knowledge, how are we to approach any other person's purported statement of fact, of truth, of knowledge?

3. How will people successfully restructure their daily lives to physically and emotionally cope with information overload? Many studies indicate that this is becoming an increasing source of stress and even illness, and certainly of declining productivity for most workers.

4. Most importantly, where are the tools that can provide the filters to information we can trust? We are still awaiting "intelligent avatars." Mark Anderson, famous for his Strategic News Service, has proposed a relatively radically alternative (for more see my blog entry on the subject). The concept behind his new service, SNS iNews, is joining a trusted community of personal and mutual interest, and then looks to the news solely from that community. We would ordinarily join multiple communities to cover our varied information requirements – a *trusted community* is the key element (something Google certainly does not provide).

The Information Explosion

The information explosion attacks us from every angle. For example the Internet Movie Database (IMDb) states that it lists 1,472,014 individual film/TV productions (games and more), and some 3,128,262 names of people who have worked on these productions.

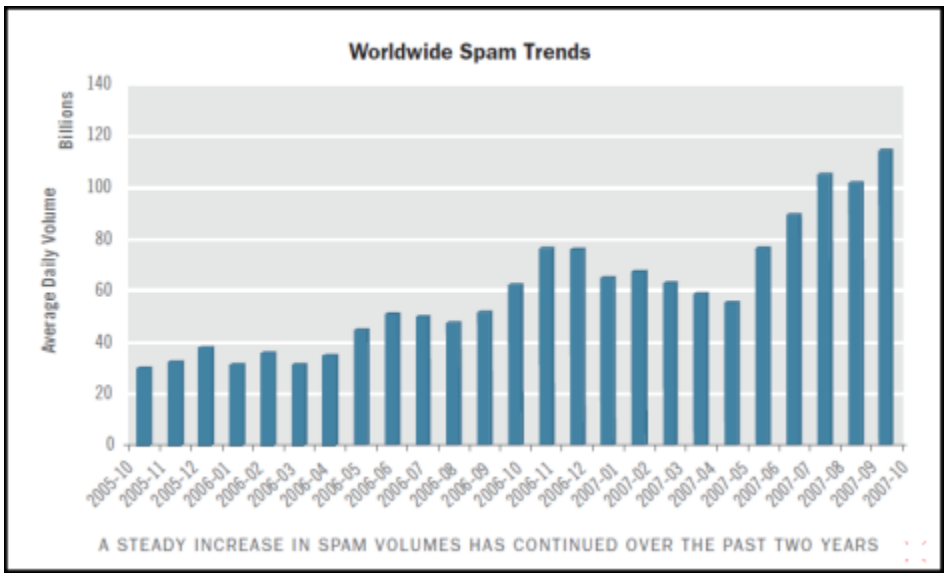
Ah, perhaps you'd have no trouble absorbing this soupçon of film and TV credits. Perhaps you'd like to get a handle on what people are blogging about. Technorati, an Internet search engine for blogs, indicates that it has indexed 133 million blog entries since 2002.

And what about email? An April, 2008 *USA Today* article cites ComScore Media Metrix figures for February, 2008:

Microsoft webmail properties: 256.2 million users
Yahoo: 254.6 million users
Google: 91.6 million users
AOL webmail properties: 48.9 million users

According to a June 2008 *New York Times* article, "A typical information worker who sits at a computer all day turns to his e-mail program more than 50 times and uses instant messaging 77 times, according to one measure by RescueTime, a company that analyzes computer use habits. The company, which draws its data from 40,000 people who have tracking software on their computers, found that on average the worker also stops at 40 Web sites over the course of the day."

At the same time, spam continues to remain a distraction for many people with inadequate filters, and even the best filters are constantly faced with combating new techniques from spammers to bypass the filters' defenses. The 2008 Internet Security Report, published by IronPort and Cisco reported that "by the end of 2006 many companies were seeing spam messages making up as much as 90 percent of their inbound mail flow."



Source: 2008 Internet Security Report, © 2008 by IronPort and Cisco

Books and Information Overload

As mentioned in my section on book publishing, according to Bowker, publishers in the United States, United Kingdom, Canada, Australia and New Zealand released 375,000 new titles and editions in English in 2004. Further Bowker reports that in 2008, as a result of on-demand print technology, "total output (in the U.S.) rose 38%, to 560,626 titles." Bowker's Books in Print offers a database of "over 5 million book, audio book, and video titles." For the ambitious reader, that's over 1,000 books per day just to keep up. My prayers go out to those who speak several languages!

There was a time when mankind's entire knowledge held in book form was less than a few hundred volumes. The notion of absorbing all of this knowledge within your lifetime was not in the least fanciful.

"So Many Books" is a delightful short tome by Gabriel Zaid, published in Spanish in 2003, and in an English translation in 2004 (it is still in print). As the dust jacket explains the book is far-ranging, full of history, philosophy, and thoughts about books and about reading. It also contains some excellent data very relevant to the consideration of information overload.

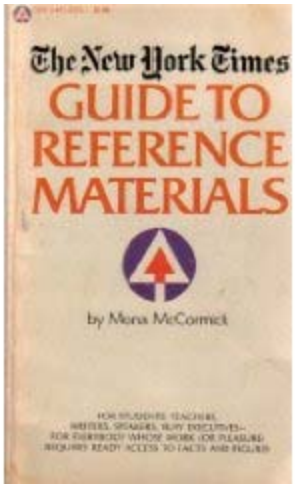
In 1450, after the invention of Gutenberg's press, 100 titles were published per year, or 0.2 per million inhabitants of the planet. By 1950, when color television was approved by the FCC, there were 250,000 titles published per year (in all languages), or 100 per million inhabitants. By the year 2,000, one million titles were published; 167 per million inhabitants.

Just as fascinating, Zaid lays out the cumulative bibliography (total number of books published) over time. He estimates that by 1550 it had reached 35,000 titles; by 1850 it was 3.3 million, and by 2000, an astounding 52 million.

As he writes, “In the first century of printing (1450-1550) 35,000 titles were published; in the last half century (1950-2000) there were a thousand times more – 36 million.”

The Extent of Information Overload

I have a used book, published originally in 1971 (though now out of print), written by Mona McCormick, called *The New York Times Guide to Reference Materials.* It cheerfully proclaims that with all of the search technology available, and the “relatively” modest number of volumes, it is still possible to “know” all that has been written and discovered. She specifically states: “So, after all, the information explosion is still a challenge we can meet.” Though now obviously a false statement, it clearly represents the pre-Internet sense of how we could cope with the information at our disposal.



The Internet has laid waste to that claim.

The information explosion cannot be “blamed” on the Internet; it existed long before that. But the Internet and the web have greatly accelerated the explosion of information (and other content) available online. This is indeed one of the key characteristics of the web: there is no limit (in theoretical terms) to the amount of information that can be published in this medium. (Although several sources are now warning that there will be insufficient storage available for the next few years to actually contain all of the data created.)

So why is this important to our exploration of the future of publishing? Simply stated, the information equation has radically changed. In order to get a handle on a subject, it’s no longer significant to aim for the assimilation of large amounts of information. It’s far more important to find the appropriate and authoritative tools and methods that also *restrict* the amount of information you’re exposed to. Google is widely-considered to be the best search engine for delivering relevant information. Its mission is “to organize the world’s information and make it universally accessible and useful.” But it may represent a primitive breakthrough in this quest. Even Google’s greatest fans admit to its many limitations. In 2004, Google claimed its site index increased to 4.28 billion web pages; in a July 2008 blog on Google appears the claim, “Recently, even our search engineers stopped in awe about just how big the web is these days -- when our systems that process

links on the web to find new content hit a milestone: 1 trillion (as in 1,000,000,000,000) unique URLs on the web at once." The statement is misleading – a little later in the blog we read, "We don't index every one of those trillion pages – many of them are similar to each other, or represent auto-generated content…that isn't very useful to searchers. But we're proud to have the most comprehensive index of any search engine, and our goal always has been to index all the world's data."

We can no longer determine with any certainty how many pages Google misses. Some estimates claim alarming large numbers. But this may be competitive nitpicking. Google doesn't miss much, and it catches new pages very quickly. What it does miss is the value of those pages it uncovers. You have to be an expert Google searcher to find the best that Google has to offer. Many search terms, even for the expert, will often fail to harvest the most relevant material. (Optimizing the process of web searching is becoming an important subject in librarianship, as well as for analysts.)

So, as you contemplate the truly vast information resources on our planet, pause twice: once to remember your own human limitations in absorbing that which you can access, and second in lamenting the many information sources that are somewhere between difficult and impossible to locate.

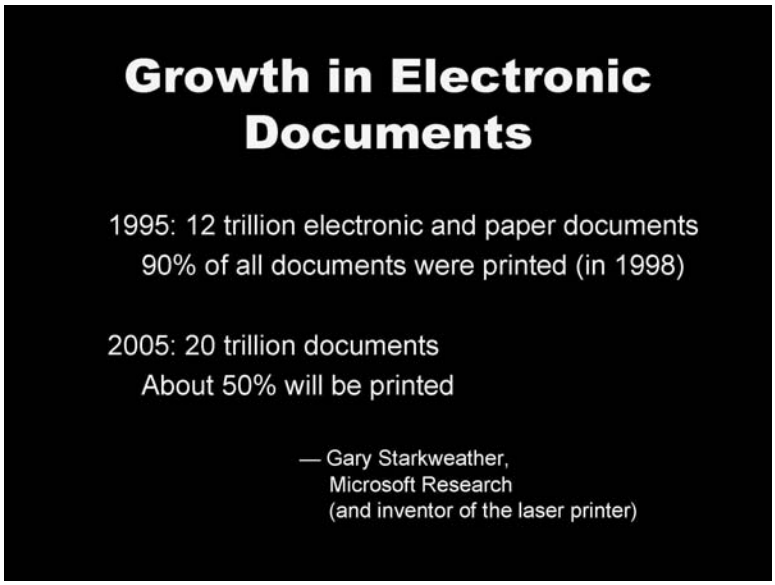
I come back to the central tenet of how to cope with information in the Internet era. The challenge is to find the information you need, to make sure that it is timely, and at least reasonably accurate (measured against the "*New York Times*" scale "All the News That's Fit to Print" — whatever that might mean today).

The downside is that you will be inundated with tough-to-sift-through garbage.

The upside is that you will find facts and analysis previously available only to graduate students and their professors.

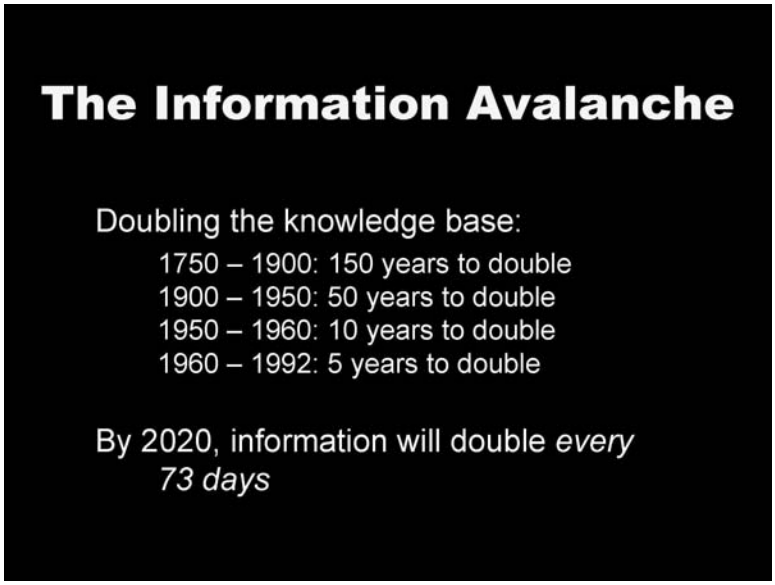
A further downside is that certain copyright holders are still holding onto their data for dear life, and not making it available on the open web.

My colleague Gary Starkweather, best known as the inventor of the laser printer, spoke in a session I moderated at an industry conference in mid-2000 and made two startling claims, as illustrated by two slides from that presentation:



This first claim is profound and unsettling. Though the statistics are dated, they point to the very clear idea that there is much more "information" available today than any human can possibly peruse. But despite the laser printer, much less is printed now than it was a decade ago. I believe this to be the case, although Hewlett-Packard may not. We become accustomed both to reading on the screen, and to parsing the information offered.

More significantly, and perhaps more subtle way, "The Information Avalanche" tells another story.

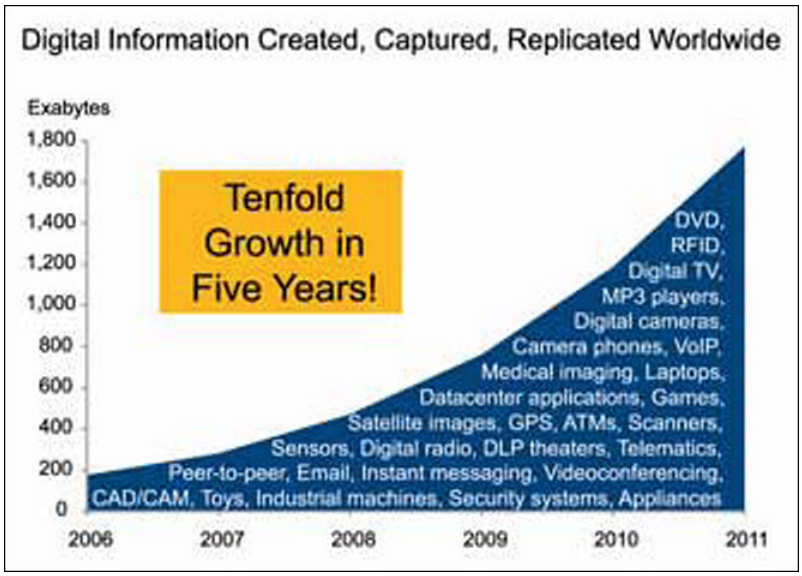


Source: *1992 Conference Teach America*, quoted by Gary Starkweather.

Although this data comes from a 1992 conference, the point is very cogent. Whether it will be 2020 or 2030, and whether it will be every 73 days, or every 100 days, the trend is backed up by a lot of industry data. See, for example, in the References section the first item: Hal Varian's team's landmark work in this area, including their "estimate that new

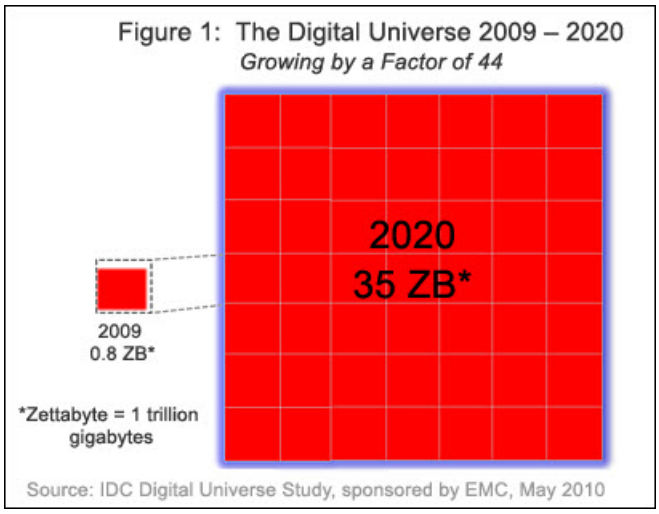
stored information grew about 30% a year between 1999 and 2002," and also "almost 800 MB of recorded information is produced per person each year. It would take about 30 feet of books to store the equivalent of 800 MB of information on paper."

A more recent report, published in March 2008, is entitled *The Diverse and Exploding Digital Universe – An Updated Forecast of Worldwide Information Growth Through 2011*. Sponsored by storage vendor EMC, and conducted by IDC, it provides some useful additional data: "The IDC research shows that the digital universe — information that is either created, captured, or replicated in digital form — was 281 exabytes in 2007. In 2011, the amount of digital information produced in the year should equal nearly 1,800 exabytes, or 10 times that produced in 2006 (see Figure 1). The compound annual growth rate between now and 2011 is expected to be almost 60%." (For definitions of these terms check the References section.)



The awkwardly-named May 2010 site update to the 2008 PDF report, *The Digital Universe Decade – Are You Ready?*, offers additional insights into issues related to information explosion, albeit with a specific focus on data storage concerns.

The latest report states that "between now and 2020 the amount of digital information created and replicated will grow to an almost inconceivable 35 trillion gigabytes as all major forms of media – voice, TV, radio, print – complete the journey from analog to digital." This report estimates that by 2020 the digital universe will be 44 times the size it was in 2009!



The Information Explosion and Serendipity

Many observers of Internet trends lament the ability of users to filter data to their own interests; thereby filtering out so much else that might serve to broaden their perspective on the world as it unfolds. Most recently Cass R. Sunstein in his revised Republic.com 2.0 argues that "…people should be exposed to materials that they would not have chosen in advance. Unplanned, unanticipated encounters are central to democracy itself. Such encounters often involve topics and points of view that people have not sought out and perhaps find quite irritating." He repeats an often-stated argument "You might, for example, read the city newspaper and in the process find a range of stories that you would not have selected if you had the power to do so. Your eyes might come across a story about ethnic tensions in Germany, or crime in Los Angeles, or innovative business practices in Tokyo, or a terrorist attack in India, or a hurricane in New Orleans, and you might read those stories although you would hardly have placed them in your Daily Me."

The simple proposition behind this argument, that serendipitous encounters with the unexpected broaden one's perspective and possibly make us more questioning and involved citizens, is not worth disputing. I do however dispute the foundation on which this argument is based. It assumes first of all that traditional media actually *did* expose readers and viewers and listeners to ideas and information and points of view that might jolt them out of their myopic worldview. It also assumes that the Internet in some way inhibits serendipity.

I question deeply whether unexpectedly encountering a wire service story in a print periodical of a recent terrorist attack in India offers anything of value to the reader. The story will be brief and superficial. Context will likely be non-existent. Commentary will be provided generally by conservative commentators: American politicians and/or the current ruling powers in India. Does this make the reader a stronger participant in democracy?

It's true that Web users can easily filter out any and all stories about India with the tools now available. But, far more significantly, should they encounter information about

what's happening today in India, they have at their disposal the richest resources in history to dig quickly and deeply into the story, and to discover a range of viewpoints *never* available from the North America media.

The information explosion can be blight or a blessing. Keep faith in the citizenry: those who care are becoming stronger, better informed and more empowered than ever before. Those who wish to doze were dozing also in the era of one-newspaper towns and three major television networks.

The Implications of the Information Explosion for Publishers

In researching this topic I stumbled upon a 1992 research paper called "The Information Explosion: Fact or Myth?" by William J. Clark of the Department of Computer Information Systems at Colorado State University in Fort Collins, CO. In the abstract for the paper comes the statement "An examination is made of data collected to measure patterns of information growth during the last century in the US. Results show that it is *information distribution* and not *information production* that has experienced explosive growth during the time frame." (Emphasis mine.)

As this article demonstrates and as you'll find illustrated from the research available throughout the article and in the References section below this statement is incorrect. Information production has seen ever-increasing growth, exploding particularly since roughly 1950. Information distribution has grown rapidly since the explosion of newspapers in the 19th century, through magazines, paperback books, telephones, radios, TV, and now the Internet. Both information production and distribution must be considered side-by-side. Information that is produced but not distributed does not contribute significantly to the information explosion.

While improvements in literacy in developing countries are creating new audiences for all forms of content, the business beneficiaries are largely (although not exclusively) local firms. Multinational conglomerates aside, most North American publishing companies have to depend for the most part on their own country first, and, to a lesser extent, some export sales.

Publishers are in the midst of an unprecedented era. More content is being created than ever before, and more channels are available to distribute that content, while at the same time population growth remains modest and business growth is not providing sufficient advertising revenue to support all of the new channels' distribution efforts.

The implication is that the heyday of mass-media has peaked and passed. New content will continue to accumulate at an increasingly rapid pace. The distribution challenge becomes how to still make a business success in reaching one customer at a time and addressing their individual requirements.

Conclusion

The impact of the information explosion is enormous and far-reaching. We currently find ourselves as its victims. Yet I firmly believe that the web itself will one day provide the

solution to the problem. The semantic web, as proposed by the World Wide Web's founder, Tim Berners-Lee, is to my knowledge the key tool with a sufficient following to begin to address the simple problem: if we cannot know everything on a topic, then let us at least know that which is most credible and relevant. More solutions will emerge over time. In the meantime we'll be gasping for breath whenever we try to become authoritative commentators on a subject of wide public interest.

References

1. How Much Information? 2003

A famous study by faculty and students at the School of Information and Management Systems at the University of Berkeley.

Professor Hal Varian of the University of Berkeley and his colleagues set out to answer the enormous question "How much new information is created each year." This team of researchers estimates that the world's total yearly production of information in the four physical media of print, film, magnetic and optical content would require roughly 1.5 billion GB of storage, the equivalent of 250 MB per human. Where possible, they have also compared their findings to their similar study done in 2000, to find out how much the amount of information had increased. Read on…and try to get your mind around this one!

http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/

http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/printable_report.pdf

2. The Global Information Industry Center (GIIC)

Located here, this group "seeks to identify and describe through its research programs the underlying issues and consequences of technology enabled change in information and communications practices in government and industry, and those affecting individuals. The Center functions as a collaborative research and learning environment for faculty, industry professionals and students to engage in projects, discussion forums, and events focused on the major program areas of the Center."

The key program underway from the Center is intended as the follow-up to Dr. Varian's work mentioned above. Called, of course, How Much Information? (or HMI? for short), the long-term research project addresses: "What is the rate of new information growth each year? Who produces the greatest amounts of information annually? Individuals? Enterprises? How does information growth in North America compare with growth in other geographies, markets, and people globally?

"To answer these questions and others, an updated and expanded How Much Information? (HMI) research program is underway."

Some of the first research briefs are now available for download.

3. How Much Information Is There In the World?

By Michael Lesk (http://www.lesk.com/mlesk/ksg97/ksg.html)

A shorter and more modest paper on Varian's topic, probably written about 1997.

4. HotTopics: Information Industry Outlook 2010: A New Dawn, New Day, New Decade
From Outsell Inc.

This yearly report from Outsell presents "information industry" predictions and expectations for the coming year. It's an good resource for getting a handle on the information industry. (It used to be free and 40 pages, but the 14-page 2009 report is $495!)

5. Slow Down, Brave Multitasker, and Don't Read This in Traffic,
by Steve Lohr, *The New York Times*, March 25, 2007

A very good overview of current theory and research into the failure of multitasking, whose ease of mastery is underestimated by too many "knowledge workers." "Several research reports, both recently published and not yet published, provide evidence of the limits of multitasking. The findings, according to neuroscientists, psychologists and management professors, suggest that many people would be wise to curb their multitasking behavior when working in an office, studying or driving a car."

6. Was I Right About The Dangers of The Internet in 1997?
by David Shenk, *Slate*, July 25, 2007

In 1997 Shenk published a book called *Data Smog*. He explains the genesis of the book: "…while doing research in Washington into public political knowledge, I started to realize that our postindustrial society was in the midst of a true phase shift — from information scarcity to information glut. Even for a culture with a basic faith in human progress and technology, such a transformation clearly presented serious personal and political challenges." This article revisits his study and its conclusions a decade later: he was mostly right!

7. Defining huge amounts of storage

It's always handy to have a simple reference nearby when trying to get your mind around the amount of information out there, expressed in computer storage terms.

kilobyte	KB	1000 bytes
megabyte	MB	1000 kilobytes
gigabyte	GB	1000 megabytes
terabyte	TB	1000 gigabytes
petabyte	PB	1000 terabytes
exabyte	EB	1000 petabytes
zettabyte	ZB	1000 exabytes
yottabyte	YB	1000 zettabytes

A piece of byte-sized trivia: According to Ian Ayres in his interesting book *Super Crunchers: Why Thinking-by-Numbers is the New Way to Be Smart* (New York: Bantam Books, 2007) the prefix "tera" derives from the ancient Greek word for monster, and as he comments: "A terabyte is truly a monstrously large quantity."

8. "There are three kinds of lies: lies, damned lies, and statistics."

This quotation is attributed often to Mark Twain, who in turn attributed it to the British politician Benjamin Disraeli. The phrase has made its way into the title of several books on the abuse of statistics, including *Lies, Damn Lies and Statistics: The Manipulation of Public Opinion in America* by Michael Wheeler (published in 1976 and now quite dated), *Damned Lies and Statistics: Untangling Numbers from the Media, Politicians, and Activists*, by Joel Best (followed by *More Damned Lies and Statistics: How Numbers Confuse Public Issues*), and the 1954 classic *How to Lie with Statistics* by Darrell Huff. I'm very aware of this issue in my work on this website, and try as hard as I can to find data from multiple sources before putting forward an argument.

9. The Information Overload Research Group

The Information Overload Research Group seeks to "work together to build awareness of the world's greatest challenge to productivity, conduct research, help define best practices, contribute to the creation of solutions, share information and resources, offer guidance and facilitation, and help make the business case for fighting information overload." It includes a link to the:

10. Information Sanity blog

"Named for the challenges we all face managing the information that comes at us everyday, from every possible direction, making human attention a scarce commodity, this blog serves as a quick reference point for coping with information overload. See what we are up to – cool events, breaking news, interesting trends and industry commentary – in the spirit of providing relevant information to help you work smarter, faster and better." Not frequently updated…I'd like to find something better.