MYSTERIES OF ONLINE

EXTRACTS FROM THE BEYOND SEARCH WEB LOG

HTTP://WWW.ARNOLDIT.COM/WORDPRESS

Table of Contents

Introduction

I wrote these brief essays in early 2009. Each of the sections is based on talks I had given prior to 2007. I did some updating. The graphics have been deleted. I made each of the sections of this collection available to the readers of my Web log, "Beyond Search." A person at the MarkLogic user conference in May 2009 asked me about the essays and wondered if I would make them available as a single publication.

I am not too keen on doing any more "regular books". The economy has taken the wind from the sails of the publishers with whom I work. I am, therefore, offering this PDF without charge on an "as is" basis.

The information may be used for personal, library, and academic purposes without asking me. If you want to use the information for commercial purposes, please, send me an email explaining what you want to do.

This material conforms to the editorial guidelines set forth for the Web log. You can read these at http://www.arnoldit.com/about.

If you want to provide feedback on these essays, please, use the comments section of my Web log at http://www.arnoldit.com/wordpress. Just run a query for "mysteries of online" and add your comment to the relevant posting.

If you want more detailed information about the topics in these brief essays, let me know. I can be reached by sending an email to seaky2000 at yahoo dot com.

Stephen Arnold
Harrod's Creek, Kentucky
July 13, 2009

1 Cannibalization

Big news. Visits to newspapers' Web sites rose in the last few months. You can read the Associated Press story here. Sorry. I won't quote from the write up because I don't need the AP's quite traditional, legal eagles occluding the sun in Harrod's Creek. The guts of the AP story is that some dead tree — aka traditional — publishers with Web sites are getting more traffic. More traffic is good, right? More traffic equals more revenue, right? The AP seems blissfully unaware that at the same time click fraud, which is consistently under estimated, is increasing, said TechCrunch here. But there's a bigger problem with Web centric online services offered by some traditional newspaper publishers.

The good news, bad news dynamic underscores the paradoxical nature of online information. For every upbeat innovation, there may be one or more downbeats. Paraphrasing Jacques Ellul, it takes technology to tame technology. The consequences are more technology and unexpected consequences. One consequence is trouble controlling costs and another is a net shortfall in revenue. The online revenue was supposed to be additive. In fact, the online revenue is lower than forecast and traditional revenues sources continue to slump, often faster with the online service than without it.

That's the core of the cannibalization issue.

Definition

Cannibalism as I use the term means eating another snail. I am using my first hand knowledge of South Africa's cannibal snail as my inspiration in this Web log post. The cannibal snail works up an appetite. Looks around for food that is close. If there is nothing tasty, the cannibal snail eats any other cannibal snails even if the other snail is a relatives. Really close relatives.

As applied to online, most people who use this term mean that one product may suck customers, money, and attention from some other product. The cannibal part refers to eating money from one member of the product family to fatten another member.

Here's how this works. A company publishes a big fat encyclopedia with inclusions (meaty ad inserts). The publisher gets the great idea to break up the big directory into smaller chunks; for example, split out the research firms from the drug distribution companies or maybe the descriptions of the Microsoft products. A price is slapped on these spin outs which can be offered online if the big book is a print only product. I am going to use examples that are about 25 years old because the examples are useful and relevant even in 2009. The cannibalization angle is that most spin out or derived products don't earn as much money as the daddy product.

Inelasticity: Just Charge a Lot and Forget Spin Outs

Even more surprising, based on my tests with the Pharmaceutical News Index, between 1981 and 1986 is that for certain products, a high price or a low price has little impact on the number of information products

sold. I learned you can pump up the price of an information product and some customers will pay for the product regardless of the cost. The surprise came when I left the price unchanged and created child products. Revenue declined. The people who wanted a segment were happy to pay less. The market for the information was not growing. Therefore, I got bitten by one of cannibalism's fangs. Spinning out child products directly hurt sales of the big daddy product. Cannibalization, in short, left me hungry for revenues.

Aggregation Versus Local News Web Site

A newspaper has a tough time competing with aggregated news outside of a newspaper's core market. Ric Manning, I, and a number of other Courier Journal professionals created Business Dateline. In the database's first year of existence, the product generated hefty online revenues and turned a profit, tough to do in the Stone Age of online. The Courier Journal shared that revenue with the 50 or so participating business news organizations. The database did not cannibalize the revenues of any one news partner. We selected only certain stories and packaged them with other stories that had a local angle but addressed a broader business issue; for example, labor management, site selection, marketing tactics, etc. The local angle provided an anchor, but the business message was broader. Local news Web sites don't blend big picture and microscopic view.

The reason is that aggregated information does not meet the same types of information needs as a local or sharply focused news source. When an innovative outfit launches Craigslist.org, the local Web sites and the local newspapers take a hit. Craigslist.org is more hip, cheaper to use, and more timely. Why wait for a dead tree edition of the Courier Journal? Go online when you want.

When a local newspaper creates an online Web site, that Web site cannibalizes the local hard copy newspapers. The bite may be modest but when every penny counts, the reality is that ad revenue from single source Web sites costs money to get and then is not as lucrative as selling the used car dealers full page ads on Sunday. The other gotcha is that the most sophisticated local news consumer will use the Web site for free. The better job the local newspaper Web site does, the quicker the sophisticated Web user is to drop the subscription. The Web is free, doesn't ask for money at the holidays, and does not leave the paper in the mud.

Let's assume that a local newspaper signs a deal with an aggregator and offers a Web site. Makes no difference. The revenue from the aggregator combined with the revenue from the local Web site won't make up for the costs of the dead tree operation as well as the costs of the Web site. Advertisers won't pay as much for online. As print display rates rose, advertisers looked elsewhere. (Think Google and other online networks.) But the revenue expectations of the dead tree crowd cannot be met.

Dead tree publishers have 400 or 500 years of business practices behind their business models. Jumping online throws some of those cherished operational procedures out the window. I get calls from PR people who think I am a journalist. I am not a journalist. I am a researcher, and when I don't get answers quickly, I post my opinions and ask for input. Whatever the journalism schools teach, I don't know. I'm 65 and happy to work in my small corner of the vineyard.

Publishers are like me when they have to deal with online information. There's a technology hurdle. There's a velocity of action issue. There's a nerd management issue. There's the cannibalization issue. What I have observed is that dead tree outfits talk a good game and then set up online services that are almost certain to accelerate three bob sled runs for their financial department:

- Loss of subscribers which means less dough comes in

- Declining ad revenue at the same time the cost of selling those ads goes up
- Inability to control non linear costs that are common to online services.

When dead tree wizards talk to me about economies of scale in online or a method to avoid cannibalization, I head for the exit.

What dead tree publishers want from their online services. Google and Facebook have one of these vehicles. Anyone can buy one, right? Money talks.

What publishers get is more like this. A traditionally engineered approach based on market research, committees, and strategic planning. Works really well. Problem was that most people did not want an Edsel. Whatever sales the Edsel generated came at the expense of Ford Motor's other products. In the 1960s, the Edsel had zero impact on GM and Chrysler.

A company like Yahoo is supposed to be pretty good at online. The company has emulated many successful online services; for example, email, news aggregation, and online photo sharing. The company is a mess.

A newspaper lacks Yahoo's technical focus and business methods. Furthermore, if Yahoo can't make its high tech business hum like a Veyron in online, how can a dead tree publisher. Microsoft is in the same sub division, rubbing shoulders with Yahoo.

The outfits that seem to be generating revenue have little in common with dead tree publications and seem to have more craft in dealing with the vagaries of online information. Amazon is doing okay. Facebook seems to be on track. There are quite a few services that are so so. But the majority of the savvy tech outfits are struggling.

Cannibalization is more common among the dead tree crowd. Savvy online companies make an attempt to avoid killing one product to fatten a new one with less earning potential. Much of Microsoft's slow march toward online or cloud services is probably due to cannibalization issues. I don't know how Microsoft will deal with cannibalization. Oracle seems to be taking tentative steps toward cloud computing and I want to see how it deals with cannibalization of its flagship products' revenues.

I submit that when technology centric companies cannot handle the cannibalization challenge, I think that dead tree outfits will face the same battle with less armor. In fact, I think some of the online services offered by dead tree publishers are the equivalent of marching into battle in the 1st century BCE naked and with no weapon. The foe is today's equivalent of Julius Caesar. I wonder is Sergey Brin is related to him?

To wrap up, cannibalization is one of the inherent "rules" in online.The silent killers are cost and revenue shortfalls.

2 Business Process Logic

I have been jotting notes to myself as I put the finishing touches on Google: The Digital Gutenberg, my forthcoming monograph about Google and integrated information manufacturing system. One of the notecards, which I am converting to this narrative Web log reminder, had the phrase “business processes built on faulty logic” circled in red.

I don’t know where or who used this phrase, but I found it suggestive this cold morning. The plunging temperatures have frozen the acid runoff stream and my pond. This addled goose, therefore, must sit in his nest contemplating the mysteries of online. Too keep my web feet frost bite free, some observations.

Business Processes: Formed by Chance, Trial and Error, and What Clients Demand

I remember an interview I conducted with a guru from Thomson, the French electronics firm, in the late 1970s. I had to jog my memory, but I looked at my 1979 copy of the “Managing Innovation” study, which my boss William P. Sommers sold to an innovation-challenged Fortune 50 company. I was one intellectual ditch digger on that project. The French Ph.D. who answered my questions about innovation said something to the effect: “Who knows. We just do what’s been done around the lab here for years.” Whatever works, I suppose.

Tradition

The notion of tradition and business processes is deeply rooted in most of the organizations with which I am familiar. In the older companies, the methods are captured customized machines like the “Joe Herman machine” to make nails in the Keystone nail mill. Today MBAs use Excel to whip up financial methods. (We know how well that works.) Tangible machine or intangible method, once these constructs are in place, change becomes difficult.

Use What You Got

Using what “you got” is a phrase that I heard in the steel mill when I was 17, and I heard it last week in a meeting with an entrepreneur building a services Web site for professionals. “Use what you got” means the learnings, instincts, and tools available here and now. When creativity flashes, a business method is developed. If the method works even sort of works, the company is good to go.

Change Is Difficult

My hypothesis is that once something sort of works, most organizations in the United States don’t to discard the method that’s in place, comfortable, and known even if that method is not flawless. Companies prefer to invest in cosmetic features and marketing programs.

Some assertions to consider:

- Most business processes are ad hoc amalgamations. Those involved use what "you got". Some folks add learnings from the university of hard knocks. Someone hacks together a solution and it becomes concrete.
- Concretized business processes freeze some employees. Who wants a weapons system that goes bang unexpectedly?
- Changing a business process produces information excitement.

Adding computer technology to a business process makes some functions faster, but the business process may generate more baloney faster.

Stepping Back

A recent Web log post by Dennis Howlett here called "Honey, I Just Blew Up the ERP" had some useful information on my thoughts. The idea is that assumptions can contribute to lousy business processes is spot on. The author points to a problem in double entry bookkeeping. If one uses traditional accounting methods, one can get a bank into financial trouble. The "fix" is (and I am simplifying here) is to shift from traditional accounting to" counting to objects". In my opinion, I do not think that the certified and chartered accountants are going to shift to counting objects. Some accountants for Satyam may tally jail time, but that's as close as most of these fine souls will get to a revolution.

I do agree that assumptions influence business methods and rush those methods toward calcification. Mr. Howlett raises the idea of systems and methods that respond to inputs. Whether one wants to view the inputs as coming from a social system or from smart software, the idea is interesting. The business process adapts; it no longer does the same stupid stuff with every cycle. Conceptually this is a pretty interesting idea.

Adaptive Business Processes

Here are some thoughts I want to capture about adaptive business processes. Remember, you won't know how something is done or when the rules changed:

The person who owns the smart system controls the game. Who will know when the owner tweaks the smart system to deliver an advantage to the owner. The software is smart. The owner is smart. The customer may not be as smart.

As people become accustomed to crappy service like fewer mail delivery days, the likelihood of forced change decreases. An example is a SharePoint / MOSS licensee who knows that the system's search performance is lousy, but the cost, knowledge, and effort required to juice SharePoint share is too much work.

Change Becomes Really Difficult

Habit resists change. A system that people get used to is comfortable. Most people have habits and routines that are tough to change.

What about Online, Search, and Information?

One of the mysteries of online is why new services have so much trouble sustaining a user base and then expanding that user base without losing yesterday's customer today? Most Web site entrepreneurs use existing business methods and processes. The assumption is that these methods will work online as they did in the non online world.

Search in a general sense has been trapped by tradition for about 25 years. In the present economic climate, some of the search vendors now have to change. The changes will be traumatic for two reasons. Some customers may resist change. Second, the outfit trying to move a new direction may leave the customers behind. Also bad.

Search vendors can index the heck out of text. Now the users don't want to find a list of files containing the word

"smith". Users want answers that fit seamlessly into the existing and what may be deeply flawed business systems and methods. Think about how search systems integrate into flawed business processes.

For example, some content management systems find themselves in a really fragile space. On one hand, organizations are struggling to deal with the notion of Web marketing, Web sites, and Web content. Many CMS have to operate within muddled business processes. The CMS itself in most cases is a collection of half backed business methods that the system designers hacked together. Merging CMS with search is going to be an interest experiment in change. Some big name search companies are tackling this opportunity as I type these ruminations. Exciting for sure.

3 Free Versus Free

I have been thinking about Chris Anderson's article "The Economics of Giving It Away" here in the Wall Street Journal, Saturday, January 31, 2009. I spoke with a couple of people and read the thoughtful posts that are plentiful. A useful one is Staci Kramer's "'Long Tail' Author Anderson: Free Doesn't Work As A Standalone Business Model" here.

Mr. Anderson is a clever wordsmith and he's pretty good with numbers. I don't disagree too much with his analysis. I have encountered similar arguments over the years, and now I understand where the mind set that generates these interesting analyses.

I come from a different mental space, and I think the free - fee duality has several twists that keep the economists thinking and the pundits punditing. Economists and pundits have quite a track record in financial matters in the last nine months.

I want to capture some of the ideas that have not been developed in my previous monographs that touch upon this subject. Relevant information appeared in Publishing on the Internet: A New Medium for a New Millennium (Infonortics, 1996) and in New Trajectories of the Internet: Umbrellas, Traction, Lift, and Other Phenomena (Infonortics, 2001). Both of these publications are out of print. These two specialist monographs were written years ago and may have been among the first to address some of the free - fee issues. I will not repeat that information, confining my comments to the information I have in my notes and not in my for fee work. While not an idea approach from the point of view of today's reader, I am keeping within the guidelines I set for this Web log as a combination of recycled ideas and some new thoughts I don't want to let slip away.

Feel free to challenge these ideas. I am still struggling with them.

When a Person Will Pay

In the free - fee arena, I want to give an example of when you, gentle reader, will probably pay anything for specific information. A commercial database, now owned by Thomson (a dead tree publishing working hard to keep up revenues as it tries to reinvent itself), is Poisondex. The idea is that when you know what poison in a person's system, you can consult the database and get pointers for saving the person's life. Poisondex is a pay to use database. It is not free. There are situations in which you personally could access the database, but I will comment on those in a moment. Back to the main point.

Your three year old child is dying. You think she ate a household substance. The doctor tells you that he has to consult Poisondex before taking steps to save your child's life. The nature of poison is such that it is possible to accelerate its effects unless the doc has the specific information at hand. The doctor tells you that you have to pay to save your child's life. Let's say that the cost of the three minute database search is $1,000.

Will you agree to pay?

In my experience, most Americans, even trophy kids or azure chip consultants with progeny, will say, "Yes. Get the data."

I have asked focus groups about this situation over the years, and the answers range from "I will pay anything" to "Are you crazy? Of course, my wife and I will pay." Okay, so now we have established that you will pay for information, and most people don't ask, "How much?"

The reasons are:

BU Your child's life to you is priceless. You will literally pay anything. This is the reason why bad guys can get people to do almost anything by threatening a child. So much for security unless the victims have had special training.

BU You assume that the source is going to be accurate. You "trust" the doctor. You don't know enough about Poisondex to "trust" its data, but the doctor "certifies" that the information is what's needed to save your child's life. In fancy Dan talk, this is provenance of information. You have to know yourself as a subject matter expert or via a proxy that Poisondex will not include information that will increase the likelihood that your child will die.

BU Your decision process is necessarily constrained. In short, you don't have the luxury of time. You don't negotiate. You don't call around and get information from contacts whom you think "know" something useful to you at this decision point.

We have, therefore, established with some certainty that information has "value"; in this case, $1,000. How much more will you pay right now? You perceive that your child's life is worth more to you than any "cost" the doctor or the database imposes on you to save your child's life. We also know that time is limited, so you can't stand around and think about searching Google or calling your roomie from your days at Yale University.

You decide. You pay. You wait to see if the child lives.

For certain information, therefore, a market exists. If information is "free" but does not have the provenance to allow you to make a decision quickly, you will almost always go with the information for which a proxy or your own perception says, "This for fee information is probably better than this free information." You don't know whether the information in Poisondex is right. You don't know if the information in Yahoo's results list is right. You go with your perception.

When a Person Won't Pay

A person may not when he or she doesn't really care whether the information is right, fresh, or convenient. Examples of this type of information is what you find on a poster slapped on a wall announcing a concert. Who really cares whether the information is 100 percent accurate? If a person loves the band, the person may use the information as a reminder that you want to buy a ticket. Convenience, not life and death value make the band info a blow off.

Web content operates in a similar way. A person can run a query for movie times, restaurant hours, and weather in Chicago. The person may expect the information to be "sort of" accurate. If it is not, the person may have other low cost, reasonably low hassle ways to get the needed information. A person can, for example, whip out his or her smartphone and make a call to your friend in Chicago and ask about the weather.

The costs for getting these pieces of information are buried in a bunch of other costs that a person has

decided are important. Who thinks about this type of "extra" benefit to an online or mobile service? I don't. The idea that Internet access or a smartphone is important to a person means that certain functions have greater value than others. Maybe the person uses the smartphone to make business calls. The email and Web surfing are nice extras, and the person uses those once in a while. But the smartphone is a phone first. If it won't make calls, the person will dump the provider and look for an alternative. The person picks a package that meets his or her perceptions of value and what he or she thinks is needed.

Note that the subjective nature of the subjective decision influences the willingness to put up with:

BU Information that is wrong

BU What context is "right"

BU The hassle associated with getting and verifying the information.

Research can provide some useful sign posts for the relative importance of these factors. But I believe that for certain types of information, getting a "free" source is okay or "good enough".

Free is a matter of perception. Here's why:

In the early days of online, commercial vendors would provide services to universities. A student could walk into a library, run a query on the system, and get information. The librarian smiled. The information was "just there", and it was beyond the ken of most users of an online system in the late 1970s to know much beyond entering a few key words and printing out abstracts.

The college students since the 1970s perceive the computer terminal in the university library as a system that provides information. The student did not care about the cost. Tuition made then many services "free". When a student paid for dorm meals, the student could or could not eat. The cost of a meal was of little interest to her. With a loan or mummy paying, the student was operating in a "free" environment. Money was for clothes, booze, and road trips. The student would pay for pizza. But pay for abstracts from the library, not much chance.

The database vendors did everything in their power to expose students to their online services. LexisNexis and Westlaw made access to their really expensive legal information services widely available. Libraries gave classes to teach students how to use such must-have databases as ABI/INFORM. Heck, I even showed up and explained how to wring real nuggets from the business databases with which I was associated.

From the point of view of the database producers and the commercial online services, these students were tomorrow's customers. Until the Web took off, who knew that online access would be like water or electricity. In the 1970s, online information was really expensive but users in universities and libraries were insulated from the cost.

From the point of view of the universities and libraries, these services were not free, but the vendors would make information available on an educational discount. Institutions wanted to give students access to the hot new online services. Everyone hoped that students would grow up to use these services and maybe create new and better services, which some students did.

The key point is that since the 1970s, users in college were addicted intentionally to free electronic information. Free was part of marketing. Free was an institutional cost. Free was not even the issue. Learning how to use electronic information was an educational focus.

Who really paid?

That's easy to answer. Prior to the Internet, commercial entities, organizations, and the government paid for online information. The vendors would cut deals with these outfits and then the organizations would permit access to the systems. The vendors wanted a one password, one user rule, but in reality, most organizations shared that one password. After all, if one were working in a corporate library that had a single password, would anyone in his or her right mind want to run queries for the five people who wanted "everything about GM" or "data about light bulbs" on the spur of the moment. I would run a training program for those who were power users, and teach those early adopters to run their own queries. Some searching had to be intermediated. The boss would just expect a librarian or power user to produce information. The word "intermediation" is important in online and it means that a person listens to your information request and then goes and gets it or does the work. Consultants are intermediaries, but that word is viewed as inappropriate to the worthies who sell their expertise most of which is updated using Google and other online services, asking colleagues, or just inventing based on prior learnings.

This cost shifting has long been a part of online. Google and other online advertising plays are little more than variants.

The point of this is that for about 40 years, online users were taught that online information was free information. As part of that process, most of the users did not pay for their drug — information and its kick of immediacy, efficiency, omnipotence, or whatever other notion everyone under the age of 40 has in their head.

So what's free? The answer is, "Anything in electronic form."

Free information lacks the context for the Poisondex example. There are quite a few fancy economic buzzwords to describe this situation. I prefer to point to companies that flop because people won't pay for their digital information regardless of form.

For What Does a Person Pay?

This is an easy question. One pays to satisfy a specific need for the information context in which one finds oneself. Here's how this works.

A lawyer's child wants a particular song. If she is an iTunes customer, she determines whether to go the mall (bummer), calling a friend to get a copy, or just click and buy. The price point is low enough to lets her act in a way that meets her needs. She may not like the integrated iTunes experience, but it is convenient. She may have her own credit card. Her dad may pay. After all, that's his job. She has other things to do.

The key is that if a person has money, that person may pay because it is easier, faster, more convenient. When the person has no money, then the person will decide to skip the info or take what's available without cost.

If a person is hip to the ways of online, some people will download the song for free. The complex actions needed to locate and obtain a pirate song are second nature to you. Anyway, kids learned in middle school that online information is abundant and comes out of computers like water from a tap. Some think that any one dumb enough to put information online should know that it will be suckable. If the song is corrupt, may a person can get someone to rip it for him or her. The person getting the song may not know if the tune is a keeper. Some folks rationalize that if he or she really likes a song, he or she might buy a concert ticket to hear the artist perform–someday, maybe. The digital information is, therefore, a free sample like the cigarettes that once were handed out on street corners as a promotion.

The reasons people pay for information include (and this is not an exhaustive list):

BU Convenience. It's easier, faster, more reliable to pay for specific information than go through hoops to get it for free

BU Bundle. A person pays once and then gets other online information. All-you-can eat buffets work the same way.

BU A person is law abiding (afraid?)

BU A person himself or herself wants to be paid for your work so he or she pays even though he or she doesn't really want to pay.

When Will a Person Pay?

In my notes, I have captured some situations in which people will pay for electronic information:

BU A company does not want to take certain information because the consequences outweigh the value of the information. Wall Street outfits, in most cases, will pay for Bloomberg, Thomson Reuters, and Financial Times information. (There are exceptions. Click here for one alleged situation. Ah, MBAs. What a delight.)

BU The customer perceives the information as having value commensurate with the cost. An example would be a person paying to see a netcast of a mixed martial arts fight.

BU The user has access to lots of different information as part of one month fee; for example, a mobile phone service may allow customers to access certain high value information as part of the deal. An example would be registering to get online access to certain information as a subscriber. The cost of this information disappears into the maw of the mobile phone company.

BU The user gets something like a hard copy or some other fungible or psychic reward for paying. An example would be buying the USA Today $5 inauguration edition of MacPaper.

BU A company or organization pays for information on the hoof with a digital saddle. A consultant to tell you what you need to know. You buy a package, image, snootiness, class, whatever.

Database producers, publishers, and others in the sprawling "information business" have discovered that their business models don't work very well in the online arena. These companies are like old gladiators who have to fight Googzilla barehanded. The fight isn't even interesting to me. Googzilla has thick scales, breathes fire, and possesses some really big incisors.

Who is a pirate? Maybe our kids?

What about the Middle Ground?

Now we come to the core of the free - fee discussion. The user has money. The information has a price. What allows the user to make the decision to pay the price and access the information.

The answer is, "Context." In certain situation, the factors (value, time, need, perception) combine to ring the cash register. If a factor shifts, the customer may not buy. Because electronic information does not "disappear" or "get used up" when transferred, it's different from a printed book.

The argument about free - fee is tricky. Most of the people talking about free or fee don't articulate their motives. Here are some examples:

BU Open source software. The supporters want to get out of the proprietary software trap, replacing it for a nerd trap with consulting services needed to make the stuff useful.

BU Sales. The free software is like the drug dealer giving the 10 year old some crack. Get them hooked young and put them to work making money for the drug dealer. The example rankles some white shoe information mavens, but think about those “addicted” to online information.

BU Time on one’s hands. Some people volunteer like my mom who racked up 50,000 hours as a hospital volunteer. Others write Web logs, make lists of the order of battle for World War II in Europe, or write Amazon book reviews, among other displacement activities.

BU Build reputation. Some companies offer shareware or freeware to be noticed and get a reputation for quality or cleverness or whatever.

BU Learn. Information in a Wikipedia or Knol post provides a person with a way to learn and show that expertise. If the person is goofy, then the article may be goofy, but the person still learned how to use the system. That learning may have been the whole point of the exercise in the first place.

BU Agitate. Check out Reddit.com for examples.

You get the idea. The information is a means, not an end, for some. Digital information is a magic substance: lubricant, commodity, high value, and possibly indestructible until someone pulls the plug.

As a result, the boundary between free and fee is fuzzy, really fuzzy. Some companies think that if they give away their commercial information, their market will grow. This may work for general interest information but it doesn’t work too well for narrow, complicated data. The reason is that the market is too small. If a commercial information product can’t make it with its information in a narrow segment, it probably will fail by giving it away to a larger group. No one will care. Specialist information almost has to have a price tag in order to have credibility. Think about the Poisondex example and your child in this context.

Shifting Who Pays

When I first encountered the pay for traffic plays that made Google possible, I realized that traditional business models were in for rough sailing. The reason is that the trophy child user gets information in a way that looks free to her, but shifts the cost to an advertiser who will pay and pay and pay. Furthermore this model operates like the old AT&T. You wanted to make a call that worked, you bought AT&T. Google is in this position now.

The point is that the Google users don’t pay for basic services. A Google user can pay for services that push the buttons for value, need, context, etc. For that reason Google is a disruptive company, and it is too late to slow the GOOG. Competitors are metaphorically behind the eight ball. Hoping that Google will self destruct may be some companies best shot at survival in my opinion.

If a user perceives information as free, that user may have little interest in digging out who foots the bill and implications of that relationship.

Disintermediation

Disintermediation is alive and well. True, libraries have been disintermediated. Users can get their own online access and run their own queries. Anyway, every time I meet a 20 something I learn that the person honestly believes himself or herself to be an online search expert. This is like the freshmen in a English 101 perceiving themselves as good writers. Or math majors who see themselves and user interface designers.

What's happened is that a value chain in information exists. Some links in the chain have been chopped out. The value chain still exists. If a person works as an intelligence agency manager, that person may depend on intermediaries which are filling the librarian role. The same applies to financial analysts. The reason intermediation exists is because the people in the chain have different contexts. Some tasks are more important than others, so some online information work gets delegated. Not even the wizards at a hot online start up can do everything informational themselves.

The grousing about Disintermediation comes from those links chopped out of the value chain. Newspapers and news being disintermediated. I have written about other examples as well. The object is to make oneself and one's knowledge have value.

This is the heart of the knowledge value strategy first articulated by Taichi Sakaiya.

Net Net

Information is a complex notion. Paying for information thrusts an addled goose like me into the murky boundaries between free and fee. The crispness of digital zeros and ones gives way to a subjective world in which the value of information and its price vary by the observers' contexts. There is a reason that theoretical physics has made significant contributions to some of Google's systems and methods.

I will leave it to the wizards who are more informed than I to simplify the effect of quantum theory for pricing online information.

4 The Bits Are Bits Fallacy

In a meeting last week, a young wizard said, “Bits are bits.” The context for this statement was a meeting to move an organization’s databased information and unstructured text online. The idea was that the task was trivial.

In fact, the task was a mixture of trivial and non-trivial sub tasks. So, bits are not the same because a zero and one may not behave like grains of salt. The ones and zeros may look the same, but one of the mysteries of online is that many factors bedevil the would be online entrepreneur. Google, for example, wants out of its AOL deal. Obviously the bit wizards at Google know that AOL bits are not Google bits here. But Google dumped some serious coinage into the online company direct mail spam made famous.

Bits are bits just like penguins.

Here’s my list of factors, which is not complete and represents my thoughts to myself:

- Digital objects have stages. The source may be transformed, indexed, tokenized, and manipulated by two or more sub sub systems. Get these processes wrong, and weird behaviors become apparent. What’s wrong? Who knows. A person or persons have to figure it out, find a fix, and implement it. As this process goes forward, it becomes apparent that the bits are a tad mischievous

- A fancy search system cannot locate a document or other object. Indexing systems may skip malformed documents, indexes may not update, and other issues annoy users. What went wrong? Who knows. A person or persons have to figure it out, find a fix, and implement it.

- A document returns a 404 or file not found error. The document used to exist because it is in the index. Now the document has gone walkabout. What’s wrong? Who knows. A person or persons have to figure it out, find a fix, and implement it.

Causes

I wish I had a fool proof way to prevent errors caused by this “bits are bits” fallacy. Much of he frustration generated by search, content management, and business intelligence systems have their roots wrapped tightly around the facile assumption that electronic information is no big deal. Electronic information is a big deal and for many organizations electronic information may be their undoing. The reason? Many assume that once a file is in electronic form, the rest is easy.

Nope. the rest is hard, often expensive, and many times conceptually beyond the mental stage on within the would be online product will perform. Those with technical knowledge often know how to ameliorate the “bits are bits” issue. In my experience, the boss explains what is required. The tech staff asks some questions. The boss answers, The tech staff shake their heads and do what they can to meet the

requirements. Often–well, many times–bosses don't listen or defer to a committee with MBAs, art school graduates, and consultants. The results are often stunning in their impact on users.

Here are some search related examples:

- One government agency involved in police work has an insecure system. One could access the system and search for information that is not generally available to the public.
- On government agency spends $6 million for a news service that is used by fewer than 60 people. Most users rely on Google, but the clunker system keeps on getting funding.
- One government agency used a search system and discovered unauthorized content on secure servers. Some of the content was rich media if you get my drift.

I know of commercial examples as well. take a look at the news media's Web sit4es. Which one is user friendly? The Chicago Tribune's for fee service. The New York Times' system? What about the Wall Street Journal? if bits were bits, these outfits would not face some of the challenges that plagues these sites.

These sites sort of work, but they limp along.

Resolving the Mystery

My view for the purposes of this short article may strike you as radical. But here goes. My hunch is that organizations will be more and more eager to get out of the information technology business. The shift is evident and will accelerate. The big winners will be vendors who offer solutions that allow the customer to have some hardware in its offices, but this will be for psychological purposes, a bit like a child with a blanket. The executive don't want the burden of making technical decisions. Out source the systems and pay attention to the real business of the organization–trying to stay in business.

instead of a highly diverse technical ecosystem, I think customers will support a three or four vendor ecosystem. I think life will be increasingly difficult for many hardware and software vendors, consultants, and executives. These folks will find themselves victims of the bits are bits fallacy. Once customers find out that bits are not bits, the revelation will force some changes that accelerate the emergence of the limited ecosystem.

Finally, I think technology will fall from the mountain top it has held for the past 40 years. Technology will be important, jut not the driver. No Web 2,0 consultant swizzle will take the place of running a profitable business. That takes human ingenuity and technical competence. The bits are bits folks will find life becoming difficult.

My hunch is that the bits are bits fallacy has contributed to the present economic climate. We have ourselves to blame. The bits, at least until Skynet becomes a reality, do our bidding. Uninformed decision making coupled with the dismissive attitude toward technical issues has:

- Made search dissatisfying
- Created information systems that users can't manipulate
- Produced business intelligence outputs that are stupid or just plain wrong

- Fostered an information technology environment in which change costs so much the customer cannot change.

Bits are, therefore, not bits.

5 The Power of Information Flows

I have some edits to stuff into the Outlook section of my new study Google: The Digital Gutenberg, but I saw another write up about the buzz over a Wall Street Journal editor's comment that "Google devalues everything." (Man, those categorical affirmatives are really troubling to this old, addled goose. Everything. Right.) The story in TechDirt has a nifty sub head, "From The No Wonder No One Uses It Department". You can read the story here. I agree with the whining about the demise of traditional media's hegemony. For me the most interesting comment in this article was:

> The value of the web and Google is that it lets people look at many sources and compare and contrast them qualitatively. Putting up a paywall is what devalues the content. It makes it harder to access and makes it a lot less useful. People today want to share the news and spread the news and discuss the news with others. As a publisher, your biggest distributors should be your community. And what does the WSJ want to do? Stop the community from promoting them. I can't think of anything that devalues their content more.

TechDirt and the addled goose are standing feather to feather.

I do, however, want to pull out my musty notes from a monograph I have not yet started to write. As you may have noticed, the title of my essay is "mysteries of online," and this is the fifth installment. I am recycling ideas from my 30 plus years in the digital information game. If you are not sure about the nature of my observations, you will want to read the disclaimer on my About page. Offended readers can spare me the jibes about the addled nature of my views in this free publication.

The future of traditional media. The ground opened and the car crashed. The foundation of the road was gone, eroded by unseen forces. Source: http://gamedame.files.wordpress.com/2008/05/car_sinkhole.jpg

What's under the surface of the dead tree executive's comments and also driving in part the TechDirt observations are some characteristics about electronic information. More people have immersed themselves in easy and painless online access. As a result, the flow of "real time" has become the digital amphetamine that whips up excitement and in some cases makes or breaks business models. I want to summarize several of the factors that are now mostly overlooked.

Information Has Force

The idea is that in the post Gutenberg era, digital information can carry quite a wallop. Some people and institutions can channel that force. Others get flattened. My father, for example, cannot figure out what a newstream is on my ArnoldIT.com Overflight service. He simply tunes out the flow because his mental framework is not set up to understand that the flow is the value. One can surf on the flow; one can drown in the flow.

Will these kids read dead tree newspapers, magazines, and textbooks as I did in the 1950s?

Information Erodes

Just like a sandblaster cleaning an old car's chassis, information has can abrade. Fire information flows at a hierarchical organization and the traditional information supports become weaker. The structure collapses. In my research the notion of a matrix-style organization is a reaction to information flows. With a matrix or other "flat" set up, acceleration occurs. Some acceleration is positive; some, not so good. Check out the collapse of BearStearns for negative.

Information Morphs

In the digital information world, one piece of information is interesting. But excitement results when two or more pieces of information fuse, mutate, and flow. My high school English teacher Edwardine Sperling would have a real problem with doing "library research" in today's world. She wanted everyone in 1959 to conduct information collection, analysis, and synthesis one way. No more. The notion of "real time" information makes traditional methods less useful. The methods are not eliminated, but options exist. Different tools are necessary to keep track of fluid information that mutates and replicates almost as if information had its own inner drive.

Information Destabilizes

When information is no longer "fixed" or even "relatively fixed", entities and processes based on permanence begin to settle. You have seen pictures–like the one above–that show a car driving along a familiar street swallowed by a sink hole. That erosion and collapse when a certain stress is applied is what's behind the assertion that Google is devaluing "everything." Google is not the cause. Google is a surfer on information flows. The Wall Street Journal dinosaur's statement is that of a 96 pound weakling on the beach griping about the high school quarterback with the girls clinging to his muscled frame. The fix is to go to the gym. Yapping won't solve the problem. While these traditional media wizards complain, the foundations of their business methods and processes are weakening. Organizational osteoporosis, not concrete and steel, ensures their fragility. Even a minor what to the ear will send these institutions reeling like a Sigma Chi on a bender.

Information weakens the "bones" of the traditional information business model. Fragility leads to breakage. Breakage can lead to withering away.

Information Is Magnetic

When information is charged by moving through individuals, systems, and methods, it becomes magnetic. Early adopters and young people respond to "weak" magnetic forces. That's why social systems like Facebook.com are a mystery to most of the old geese around my pond in Harrod's Creek. As the magnetic force of the information increases, then more people get "pulled" into the system. Look at Google's Web search market share. Google has a heck of a pull. The paper Financial Times or Wall Street Journal has a weaker pull and it operates on a demographic that is not in step with the weak forces in information. You can't "get" magnetism. Magnetism in digital information arises from flows. The analogy with field theory is not perfect, but you have to "generate" an information force. You can't wish it into existence in my experience.

What's the Fix?

To wrap up, when one looks at these "mysteries" of digital information as fundamentals, then it become easier to understand why the Wall Street Journal maven's remark is evidence of the core perceptual problem in traditional media. These "mysteries" also explain why TechDirt and I see eye to eye on the challenges traditional media face. I don't know the folks at TechDirt, but it's clear that digital information has imprinted our perceptual framework in somewhat similar ways. The gulf between how I see information and how the traditional media perceive information is wide and possibly unbridgeable.

6: Revenue Sharing

I was finishing the revisions to my monetization chapter in Google: The Digital Gutenberg and ran across notes I made in 1996, the year in which I wrote several articles about online for Online Magazine. One of the articles won the best paper award, so if you are familiar with commercial databases, you can track down this loosely coupled series in the LITA reference file or other Dialog databases.

Table 1: Terms in This Section

Term	Definition
database	A file of electronic information in a format specified by the online vendor; for example Dialog Format A or EBCIDIC
database producer	An organization that creates a machine-readable file designed to run on a commercial online service
online revenue	Cash paid to a database producer generated when a user connected to an online database and displayed online or output the results of a search to a file or a hard copy
online vendor	A commercial enterprise that operated a time sharing service, search system, and customer support service on a fee basis; that is, annual subscription, online connect charge, online type or print charge
publisher	An organization engaged in creating content by collecting submissions or paying authors to create original articles, reports, tables, and news
revenue	Money paid by an organization or a user to access an online vendor's system and then connect and access the content in a specific database; for example, Dialog File 15 ABI/INFORM

My "mysteries" series has evoked some comments, mostly uninformed. The number of people who started working in search when IBM STAIRS was the core tool are dwindling in number. The people who cut their teeth in the granite choked world of commercial online comprise an even smaller group. Commercial online began with US government funding in the early 1960s, so Ruby loving script kiddies are blissfully ignorant of how online files were built and then indexed. No matter. The lessons form foundation stones in today's online world.

Indexing and Abstracting

Aggregators collect content from many different sources. In the early days of online, this meant peer reviewed articles. Then the net gathered magazines and non-peer reviewed publications like trade association magazines. Indexing and abstracting in the mid 1960s was a backwater because few publishers knew much about online. Permission to index and abstract was often not required and when a publisher wanted to know why an outfit was indexing and abstracting a publication, the answer was easy. "We are creating a library reference book." Most publishers cooperated, often providing some of the indexing and abstracting outfits with multiple copies of their publications.

Some of the indexing and abstracting was very difficult; for example, legal, engineering, and medical

information posed special problems. The vocabulary used in the documents was specialized, and word lists with Use For and See Also references were essential to indexing and abstracting. The abstract might define a term or an acronym when it referenced certain concepts. When abstracts were included with a journal article, the outfit doing the indexing and abstracting would often ask the publisher if it was okay to include that abstract in the bibliographic record. For decades publishers cooperated.

The reason was that publishers and indexing and abstracting outfits were mutually reinforcing operations. The published collected money from subscribers, members, and in some cases advertisers. The abstracting and indexing shops earned money by creating print and electronic reference materials. In order to "read the full text", the researcher had to have access to a hard copy of the source document or, in some cases, a microfilm instance of the document.

No money was exchanged in most cases. I think there was trust among publishers and indexing and abstracting outfits. Some of the people engaged in indexing and abstracting crated products so important to certain disciplines that courses were taught in universities worldwide to teach budding scientists and researchers how to "find" and "use" indexes, abstracts, and source documents. Examples include the Chemical Abstracts database, Beilstein, and ABI/INFORM, the database with which I was associated for many years.

Pay to Process Content

By 1982, some publishers were aware that abstracting and indexing outfits were becoming important revenue generators in their own right. Libraries were interested in online, first in catalogs for their patrons, and then in licensing certain content directly from the abstracting and indexing shops. The reason for this interest from libraries (medical, technical, university, public, etc.) was that the technology to ingest a digital file (originally on tape) was becoming available. Second, the cost of using commercial online services which would make hundreds of individual abstract and index databases available was variable. The library (academic or corporate) would obtain a password and a license. Each database incurred a charge, usually billed either by the minute or per query. Then there was online connect charges imposed by outfits like Tymnet or other services. And there were even charges for line returns on the original Lexis system. Libraries had limited budgets, so it made sense for some libraries to cut the variable costs by loading databases on a local system.

By 1985, full text became more attractive to users. The reason was that A&I (abstracting and indexing) services provided pointers. The user then had to go find and read the source document. The convenience of having the bibliographic information and the full text online was obvious to anyone who performed research in anything other than a casual, indifferent manner. The notion of disintermediation expanded first in the A&I field because with full text, why pay to crate a formal bibliographic record and manually assign index terms. The future was full text because systems could provide pointers to documents. Then the document of interest to the researcher could be saved to a file, displayed on screen, or printed for later reference.

The shift from the once innovative A&I business to the full text approach threw a wrench into the traditional reference business. Publishers were suspicious and then fearful that if the full text of their articles were in online systems, subscription revenues would fall. The publishers did not know how much risk these systems poses, but some publishers like Crain's Chicago Business wanted an up front payment to permit my organization to crate full text versions of certain articles in the Crain publications. The fees were often in the five figure range and had additional contractual obligations attached. Some of these original constraints may still be in operation.

Negotiating an online deal is similar to haggling to buy a sheep in an open market. The authors were often

included among the sheep in the traditional marketplace for information.

Revenue Sharing

Online vendors like Dialog Information Services knew that change was in the air. Some vendors like Dialog and LexisNexis moved to disintermediate the A&I companies. Publishers jockeyed to secure premium deals for their full text material. One deal which still resonates at LexixNexis today was the New York Times's arrangement with LexisNexis for the New York Times's content. At its height, the rumor was that LexisNexis paid more than $1 million for the exclusive that put the New York Times's content in the LexisNexis services. The New York Times decided that it could do better by starting its own online system. Because publishers saw only part of the online puzzle, the New York Times's decision was a fateful one which has hobbled the company to the present day. The New York Times did not understand the cost of the infrastructure and the importance of habituated users who respond to the magnetism of an aggregate service. Pull out a chunk of content, even the New York Times's content, and what you get is a very expensive service with insufficient traffic to pay the overall cost of the online operation. Publishers making this same mistake include Dow Jones, the Financial Times, and others. The publishers will bristle at my assertion that their online businesses are doomed to be second string players, but look at where the money is today. I rest my case.

To stay in business, online players cooked up the notion of revenue sharing. There were a number of variations of this business model. The deal was rarely 50 - 50 for the simple reason that as contention and distrust grew among the vendors, the database companies, and the publishers, knowledge of costs was very difficult to get. Without an understanding of costs in online, most organizations are doomed to paddling upstream in a creek that runs red ink. The LexisNexis service may never be able to work off the debt that hangs over the company from its money sucking operations that date from the day the New York Times broke off to go on its own. Dow Jones may never be able to pay off the costs of the original Dow Jones online service which ran on the mainframe BRS search system and then the expensive joint venture with Reuters that is now a unit in Dow Jones called Factiva. Ziff Communications made online pay with its private label CompuServe service and its savvy investments in high margin database and operations that did business as Information Access. Characteristic of Ziff's acumen, the Ziff organization exited the online database business in the early 1990s and sold off the magazine properties, leaving the Ziff group with another fortune in the midst of the tragedy of Mr. Ziff's health problems. Other publishers weren't so prescient.

With knowledge in short supply, here were the principal models used for revenue sharing:

Tactic A: Pool and Payout Based on Percentage of Content from Individual Publishers

This was a simple way to compensate publishers. The aggregator would collect revenues. The aggregator would scrape off an amount to cover various costs. The remainder would then be divided among the content providers based on the amount of content each provider contributed. To keep the model simple (it wasn't) think of a gross online revenue of $110. Take off $10 for overhead (the actual figure was variable and much larger). The remainder is $100. One publisher provided 60 percent of the content in the pay period. Another publisher provided 40 percent of the content in the pay period. One publisher got a check for $60 and the other a check for $40. The pool approach guarantees that most publishers get some money. It also makes it difficult to explain to a publisher how a particular dollar amount was calculated. Publishers who turned an MBA loose on these deals would usually feel that their "valuable" content was getting short changed. It wasn't. The fact is that disconnected articles are worth less in a large online file than a

collection of articles in a branded traditional magazine. But most publishers and authors today don't understand this simple fact of the value of an individual item within a very large collection.

I was fascinated when smart publishers would pull out of online services and then try to create their own stand alone online services without understanding the economic forces of online. These forces operate today and few understand them after more than 40 years of use cases.

Tactic B: Minimums

The second model is that the publisher demands a minimum. Some publishers then want to participate in the pool as well. Different tactics are possible. The idea is that the publisher would get some money and then a piece of the action. This worked for special cases, but most publishers were not able to negotiate these types of deals. The reason goes back to what happens when a large amount of information on a specific topic is aggregated. Publishers and authors think their work is of monumental importance., A tiny fraction of what's published is important. An even smaller portion is monumental. Some monumental articles like Albert Einstein's essays in the early 20th century were not monumental until several decades drifted past. The online savvy company would focus on getting as many sources as possible, confident that the information would appear in multiple sources. Even if the information was "secret" at one time, with multiple sources, the "secret" would be in the online file. This is the core of open source intelligence, and it works. Most people don't understand this principle thoroughly today.

Tactic C: A Quid Pro Quo

A third approach to revenue was to shift from hard cash payouts to other benefits. These ranged from managing fulfillment of photocopies (a major hassle for a traditional publisher) and sharing the revenue derived from hard copy sales to providing access to the online file containing the publisher's information without a charge. In the commercial online world, the publisher who understood online could use an expensive online service without paying. The money that would have be paid to use the service was, therefore, preserved. The payment to the publisher was this cash conservation. Other publishers wanted information. The database vendor would provide briefings, often coincident with a contract cycle. The idea was that the publisher was part of a grand experiment and getting information straight from those who knew how online worked. Other variations were possible, but the goal was to compensate the content provided without distributing cash.

Tactic D: Mixing and Matching

Most online services and database producers would mix and match these compensation schemes. The advent of the Internet and the ad model has marginalized these tactics, however. But if you poke around, you can find vestiges of these three approaches.

What about the Service Bureaus?

Keep in mind that in the early days of online, the systems were hard to use. The main vendors were Dialog Information Services, SDC (Systems Development Corp.), the European Space Agency, LexisNexis, and a handful of others. These outfits operating the plumbing, provided online connectivity, and handled maint3enance of the search systems. The database producers were contractors to the service bureaus.

The relationship between the online vendors and the database producers was similar to that between the database producers and the publishers. There was one important difference. The database producers knew

how online worked and understood how little control over the customers and the revenue from the online usage of their files.

Distrust became the handmaiden of database producers. Publishers distrusted the database producers and the online vendors. Database producers distrusted publishers and the online vendors. Online vendors distrusted database producers and publishers. The result of this phase change meant that by the early 1990s, no one felt comfortable in what was once an exclusive, in-crowd.

Little wonder that when a graphical browser, the Internet, and simple electronic publishing became available, the stampede online began. Service bureaus still exist, but we call them managed service providers or data center vendors. The economics are still murky, but most are on some sort of pay as you go basis. The nonlinear pricing surprises still exist, but these are part of the economics of online. The cost spikes are not more tolerable than they were 40 years ago because options exist. Options did not exist 40 years ago. That's one plus of today's online world.

Few people even today understand the complexity of running an online service. When people tell me that their systems people can run an online service, I get out of Dodge City. Online services are specialized, complex, and black holes of money if a trophy generation wonder makes a trivial mistake. Don't believe me. Look at your company's data center costs and now project that into a public online environment with tens of thousands of users every day. What about millions of users? If you can't see the cost issues, don't read this Web log. Stop. Get into gardening or house painting.

Wrap Up

The learnings that I keep on my notecards are easy to summarize:

- Traditional media business models and the original online business models no longer work reliably in today's online world. Google's power comes from its plumbing and its business model. You can't have one without the other no matter what a trophy generation parvenu consultant asserts.
- Distrust is part of the landscape. Combined with ignorance of the economics of online, the mixture is volatile and potentially dangerous in terms of legal liability.
- Misunderstanding of the consequences of trivial decisions about online lead to technical, managerial, customers, and financial problems. To the uninformed, these problems seem random, disconnected, and unpredictable. The problems are not what they seem as long as one has the appropriate knowledge. Without that knowledge, information problems are unfixable by trial and error because the person or company will run out of time and money in most cases before the solution is implemented. Look at the present efforts of America Online. Ask.com, Microsoft, the New York Times, and Yahoo. Lots of effort; minimal or negative progress in terms of cost control, customer uptake, and revenue.

I have skipped over most of my note cards. This seems to be a fertile business topic. I may provide other thoughts from my notecards. Feel free to provide factual anecdotes, verifiable examples, and data on this topic. Telling me I am stupid or an idiot is okay. I admit that I am an addled goose, but I prefer information not the ad hominem approach to my opinions.,

7 Errors, Quality, and Provenance

This installment of "Mysteries of Online" tackles a boring subject that means little or nothing to the entitlement generation. I have recycled information from one of my talks in 1998, but some of the ideas may be relevant today. First, let's define the terms:

- Errors–Something does not work. Information may be wildly inaccurate but the user may not perceive this problem. An error is a browser that crashes, a page that doesn't render, a Flash that fails. This notion of an error is very important in decision making. A Web site that delivers erroneous information may be perceived as "right" or "good enough". Pretty exciting consequences result from this notion of an "error" in my experience.

- Quality–Content displayed on a Web page is consistent. The regularity of the presentation of information, the handling of company names in a standard way, and the tidy rows and columns with appropriate values becomes "quality" output in an online experience. The notion of errors and quality combine to create a belief among some that if the data come from the computer, then those data are right, accurate, reliable.

- Provenance–This is the notion of knowing from where an item came. In the electronic world, I find it difficult to figure out where information originates. The Washington Post reprints a TechCrunch article from a writer who has some nerve ganglia embedded in the companies about which she writes. Is this provenance enough or do we need the equivalent of a Ph.D. from Oxford University and a peer reviewed document. In my experience, few users of online information know or know how to think about the provenance of the information on a Web page or in a search results list. Pay for placement adds spice to provenance in my opinion.

So What?

A gap exists between individuals who want to know whether information is accurate and can be substantiated from multiple sources and those who take what's on offer. Consider this Web log post. If someone reads it, will that individual poke around to find out about my background, my published work, and what my history is. In my experience, I see a number of comments that say, "Who do you think you are? You are not qualified to comment on X or Y." I may be an addled goose, but some of the information recycled for this Web log are more accurate than what appears in some high profile publications. A recent example was a journalist's reporting that Google's government sales were about $4,000, down from a couple of hundred thousand dollars. The facts were wrong and when I checked back on that story I found that no one pointed out the mistake. A single GB 7007 can hit $250,000 without much effort. It doesn't take many Google Search Appliance Sales to beat $4,000 a year in revenue from Uncle Sam.

The point is that most users:

- Lack the motivation or expertise to find out if an assertion or a fact is correct or incorrect. Instead of becoming a priority, in my opinion, few people care too much about the dull stuff–chasing facts. Even when I chase facts, I can make an error. I try to correct those I can. What makes me nervous are those individuals who don't care whether information is on target.

- See research as a core competency. Research is difficult and a thankless task. Many people tell me that they have no time to do research. I received an email from a person asking me how I could post to this Web log every day. Answer: I have help. Most of those assisting me are very good researchers. Individuals with solid research skills do not depend solely upon the Web indexes. When was the last time your colleague did research among sources other than those identified in a Web index.

- Get confused with too many results. Most users look at the first page of search results. Fewer than five percent of online users make use of advanced search functions. Google, based on my research, takes a "good enough" approach to their search results. When Google needs "real" research, the company hires professionals. Why? Good enough is not always good enough. Simplification of search and the finding of information is a habit. Lazy people use Web search because it is easy. Remember: research is difficult.

What's the Mystery?

The mystery is that content which is factually correct may be perceived as wrong or off base if the presentation of the data is not consistent, easily understood, distraction fee.

The challenge of online is to have solid data and achieve consistency. The systems must be easy to use and make it possible for users to compare and contrast like data. A handful of systems present side by side results, but these systems are not used.

The bafflers, in my opinion, are:

- Online makes it easy to spot certain types of errors. For example, a missing value in a D&B credit report. Viewed on paper, the gap may not be noticeable. Online the mistake can often be spotted easily if one takes the time to look. At the same time, the blurring homogenization of pages of text make it tough to focus and spot errors. Flipping and multi tasking exacerbate the problem.

- Good information when converted to online form may become unusable information. Examples of this range from spotting a bogus Web page pumped up on SEO steroids to crammed interfaces with lots and lots of headlines, intrusive pop-ups, and crazy color schemes. I find www.popurls.com hard to use even less useful than print outs of the lists. Spotting an error online is quite difficult due to the colors, blue and black with white headlines. The interface invites hip hopping around.

- A news story or article broken up across three or more Web pages makes it hard for me to flip back and check what the author said on a previous page. The ads, the surveys, and the automatic videos–these erode attention. If a user does not exert considerable effort, the information may be right and be ignored or looked at with half an eye.

Is There a Fix?

Search without search may provide some benefits for certain device users. For general users, I am not sure how the rigor of checking sources, comparing data points, and digging into multiple, high value resources can be inculcated in most Web users. Some people have an affinity for research. Others, in my opinion, don't know how poor their research and information processing skills are. Others are content with the first Google listing in a results list.

The big fix will come from a company that "becomes" the Internet. In that scenario, one organization can use its view of the datasphere to make decisions about what's relevant and what's reliable. If this sounds scary, think in terms of benign reference librarian. This person has the expertise to judge accuracy, reliability, and provenance. Few doubted the librarian in grade school I attended from 1950 to 1958.

One of the mysteries of online then is the answer to this question, "Who will become the librarian who takes care of our meta-information needs in the 21st century? Send me your candidates, please.

8 Duplicate Content

In print, duplicates are the province of scholars and obsessives. In the good old days, I would sit in a library with two books. I would then look at the data in one book and then hunt through the other book until I located the same or similar information. Then I would examine each entry to see if I could find differences. Once I located a major difference such as a number, a quotation, or an argument of some type, I would write down that information on a 5×8 note card. I had a forensics scholarship along with some other cash for guessing accurately on objective tests. To get the forensics grant, I had to participate in cross examination debate, extemporaneous speaking, and just about any other crazy Saturday time waster my "coaches" demanded.

Not surprisingly, mistakes or variances in books, journals, and scholarly publications were not of much concern to some of the students who attended the party school that accepted an addled goose with thick glasses. There were rewards for spending hours looking for information and then chasing down variances. I recall that our debate team, which was reasonably good if you liked goose arguments, were putting up with a team from Dartmouth College. I was listening when I heard a statement that did not match what I had located in a government reference document and in another source. The opponent from Dartmouth had erroneously presented the information. I gave a short rebuttal. I still remember the look of nausea that crossed our opponent's face when she realized that I presented what I found in my hours of manual checking and reminded the judges that distorting information suggests an issue with the argument. We won.

For most people, the notion of having two individuals with the same source is an example of duplicate information. Upon closer inspection, duplication does not mean identical in gross features. Duplication drills down to the details of the information and to the need to determine which item of information is at variance and then figuring out why and what is the most likely version of the duplicate.

That's when the fun begins in traditional research. An addled goose can do this type of analysis. Brains are less important than persistence and a toleration for some dull, tedious work. As a result, finding duplicative information and then figuring out variances was not something that the typical college sophomore spends much time doing.

Enter computer systems.

Now the issue of duplicates and variances takes an interesting twist. There is now a mindless system to look at information and find identical information. In the happy world of computer systems, a duplicate is an object that is identical in every respect, right down to the punctuation in the last sentence of the last paragraph. One variance means that the two information objects (documents) are different.

The result of this is that documents that are alike in most respects but different in one or more features are considered different. The problem is that a human can look at two versions of a story on two different online services and recognize that the stories are "about" the same subject. A quick scan of the two articles will allow most online news sucking humans to determine if these documents are pretty much alike. If there is a difference such as a picture in one article and no picture in another, even a speed reader will spot

the difference. If the facts are generally in line in each article, the reader concludes that the two articles are the “same”.

Yikes, the news sucking human has identified a duplicate and that process does not match what the computer does. The human operates at a higher level of abstraction than the average computer. A human discards certain variances and hooks into important differences as the human perceives them. Remember, this is not a scholar who will be using a different mental flight path.

No one wants to read dozens of identical “hits” that are essentially the “same” article repeated over and over. So, online systems have devised methods to eliminate duplicates from certain hit lists. Other systems identify duplicates and then group them together so the news sucking human can read one version of the story and then dig into other variants as time and inclination permit.

There are, therefore, different meanings to the notion of duplicate content.

One one hand, there is the computer system that can easily match two information objects bit for bit and determine if the documents are identical. There are short cuts that operate to make this a relatively speedy process today.

Then there is a human who looks at two articles and can determine, often without much thought, that the articles are duplicates. But humans can make some mistakes because in certain documents that look identical in broad features may have some important minor variances. These “minor variances” can mean the difference between going to jail or avoiding jail, the perp walk, and the awkward reentry explanations.

Yikes again. Duplicate content, variances, and similarities are suddenly not such a backwater or semi conscious perceptual operation. Like so many digital information issues, duplicates often play no role in production. When those involved turn their attention to the issue, duplicates and the task of deduplication takes center stage. A steep learning curve presents itself. Duplicates quickly become a cost consideration. In most electronic publishing systems, duplicates and duplicate detection become the focal point of trade offs, compromises, and short cuts.

Once again, the uninformed, the trophy generation computer science grad, and the Peter Principle-in-action manager find themselves in scramble mode.

What You Can Buy

A search on Googzilla will reveal that a number of companies offer deduplication tools. Read the fine print. Deduplication for structured data works reasonably well. Deduplication for unstructured or semi structured data is a different animal. Companies often bundle deduplication routines with other content transformation products. You will want to get trial versions of the software and run tests on content with known duplicates. The performance and efficacy of the vendor’s system can then be determined. The more casual your tests, the greater the risks of answering questions in an adversarial situation accurately. The more stringent your tests, the more informed one’s answer may be. Guessing is probably not a good idea.

The ArnoldIT.com approach is to use our custom tools. We know what these scripts can and cannot do. There may be “better” systems available, but when the time window is open one or two inches, the system we know makes my comfort level creep up a notch. Dedeuplication is dependent on definitions, content collection characteristics, and context. One size does not fit every content foot. Custom tweaking is sometimes required.

Points to Consider

I am dipping into my notes from the 1979 to 1980 period so spare me the complaints that the information is old. (Check out the editorial policy here if you are a newcomer to this Web log.)

- What is the policy for duplicates for the particular online system that will be deployed? eDiscovery has one angle of attack; a news service has another.
- What is the method that will be used to identify duplicates, near duplicates, and distinct objects? Keep in mind that money, technical expertise, software and machine resources are needed to implement the grand plans for duplicates.
- Will the user understand what he / she is looking at when deduplicated results are displayed? Confused users are not a positive. On the other hand, if you get the duplicate method wrong, the user may not go away quietly. Think lawyers suing an organization.
- What's is the policy regarding errors? Two document identical in every respect except for one values in a table of payments are in what category? A duplicate or a mistake? The answer to this question depends upon the context of the parties involved.
- What does the system do with duplicates? Delete them? Archive them? There are risks associated with some approaches to archiving and retention.
- When an error becomes known, how will the error be corrected? The policy for this method used requires some careful thought.

Wrap Up

The issue of duplicative information is not easily confined. The decisions about duplicates made for the Business Dateline database may not be appropriate for today's online user. Most organizations do not pay much attention to version control, emails that have been doctored by their recipients, text in Word files that the user thinks has been stricken from the "final" version, copies of images in asset management systems, and consultant reports that have been purchased and then copied near and far within an organization so dozens or scores of instances of an "eyes only" document are floating around. There are duplicate Web sites and single Web sites with identical content. The list of duplicate issues can be extended.

The challenges range from basic definitions of duplicates to matters of policy to software methods. Hovering over the issue is a human's ability to look at two documents and say, "Why are you showing me duplicates?"

9 Time

Electronic information has an interesting property: time distortion. The distortion has a significant effect on how users of electronic information participate in various knowledge processes. Information carries humans along much as a stream whisks a twig in the direction of the flow. Information, unlike water, moves in multiple directions, often colliding, sometimes reinforcing, and at others in paradoxical ways that leave a knowledge worked dazed, confused, and conflicted. The analogy of information as a tidal wave connotes only a partial truth. Waves come and go. Information flow for many people and systems is constant. Calm is tough to locate.

In the good old days of cuneiform tablets, writing down the amount of wheat Eknar owed the king required specific steps. First, you had to have access to suitable clay, water, and a clay kneading specialist. Second, you needed to have a stylus of wood, bone, or maybe the fibula of an enemy removed in a timely manner. Third, you had to have your data ducks in a row. Dallying meant that the clay tablet would harden and make life more miserable than it already was. Once the document was created, the sun or kiln had to cooperate. Once the clay tablet was firm enough to handle without deleting a mark for a specified amount of wheat, the tablet was stacked in a pile inside a hut. Forth, the access the information, the knowledge worker had to locate the correct hut, find the right pile, and then inspect the tablets without breaking one, a potentially bad move if the king had a short temper or needed money for a war or a new wife.

In the scriptorium in the 9th century, information flow wasn't much better. The clay tablets had been replaced with organic materials like plant matter or for really important documents, the scraped skin of sheep. Keep in mind that other animals were used. Yep, human skin worked too. Again time intensive processes were required to create the material on which a person would copy or scribe information. The cost of the materials made it possible to get patrons to spit out additional money to illustrate or illuminate the pages. Literacy was not widespread in the 9th century and there were a number of incentives to get sufficient person power to convert foul papers to fair copies and then to compendia. Not just anyone could afford a book. Buying a book or similar document did not mean the owner could read. The time required to produce hand copies was somewhat better than the clay tablet method or the chiseled inscriptions or brass castings used by various monarchs.

Yep, I will have it done in 11 months, our special rush service.

With the invention of printing in Europe, the world rediscovered what the Chinese had known for 800, maybe a thousand years. No matter. The time required to create information remained the same. What changed was that once a master set of printing plates had been created. A printer with enough capital to buy paper (cheaper than the skin and more long lasting than untreated plant fiber and less ink hungry than linen based materials) could manufacture multiple copies of a manuscript. The out of work scribes had to find a new future, but the impact of printing was significant. Everyone knows about the benefits of literacy, books, and knowledge. What's overlooked is that the existence of books altered the time required to move information from point A to point B. Once time barriers fell, distance compressed as well. The world became smaller if one were educated. Ideas migrated. Information moved around and had impact, which I discussed in another Mysteries of Online essay. Revolutions followed after a couple hundred years, but the mindless history classes usually ignore the impact of information on time.

If we flash forward to the telegraph, time accelerated. Information no longer required a horse back ride, walk, or train ride from New York to Baltimore to close a real estate transaction. Once the new fangled electricity fell in love with information, the speed of information increased with each new innovation. In fact, more change in information speed has occurred since the telegraph than in previous human history. The telephone gave birth to the modem. The modem morphed into a wireless USB 727 device along with other gizmos that make possible real time information creation and distribution.

Time Earns Money

I dug out notes I made to myself sometime in the 1982 – 1983 time period. The implications of time and electronic information caught my attention for one reason. I noted that the revenue derived from a database with weekly updates was roughly 30 percent greater than information derived from the same database on a monthly update cycle. So, four updates yielded a $1.30, not $1.00. I wrote down, “Daily updates will generate an equal or greater increase.” I did not believe that the increase was infinite. The rough math I did 25 years ago suggested that with daily updates the database would yield about 1.6 percent more revenue than the same database with a monthly update cycle. In 1982 it was difficult to update a commercial database more than once a day. The cost of data transmission and service charges would gobble up the extra money, leaving none for my bonus.

In the financial information world, speed and churn are mutually reinforcing. New information makes it possible to generate commissions.

Time, therefore, not only accelerated the flow of information. Time could accelerate earnings from online information. Simply by u9pdating a database, the database would generate more money. Update the database less frequently, the database would generate less money. Time had value to the users.

I found this an interesting learning, and I jotted it down in my notebook. Each of the commercial database in which I played a role were designed for daily updates and later multiple updates throughout the day. To this day, the Web log in which this old information appears is updated on a daily basis and several times a week, it is updated multiple times during the day. Each update carries and explicit time stamp. This is not for you, gentle and patient reader. The time stamp is for me. I want to know when I had an idea. Time marks are important as the speed of information increases.

Implications

The implications of my probably third-hand insight included:

- The speed up in dissemination means that information impact is broader, wider, and deeper with each acceleration.
- Going faster translates to value for some users who are willing and eager to pay for speed. The idea is that knowing something (anything) first is an advantage.
- Speed is not enough. Customers addicted to information speed want to know what’s coming. The inclusion of predictive data adds another layer of value to online services.
- Individuals who understand the value of information speed have a difficult time understanding why more online systems and services cannot deliver what is needed; that is, data about what will happen with a probability attached to the prediction.

> Knowing that something has a 70 chance of taking place is useful in information sensitive contexts.

Let me close with one example of the problem speed presents. The Federal government has a number of specialized information systems for law enforcement and criminal justice professionals. These systems have some powerful, albeit complex, functions. The problem is that when a violation or crime occurs, the law enforcement professionals have to act quickly. The longer the reaction time, the greater the chance that the bad egg will tougher to apprehend increases. Delay is harmful. The systems, however, require that an individual enter a query, retrieve information, process it and then use another two or three systems in order to get the reasonably complete picture of the available information related to the matter under investigation.

The systems have a bottleneck. The human. Law enforcement personnel, on the other hand, have to move quickly. As a result, the fancy online systems operate in one time environment and the law enforcement professionals operate in another. The opportunity to create systems that bring both time universes together is significant. Giving a law enforcement team mobile comms for real time talk is good, but without the same speedy and fluid access to the data in the larger information systems, the time problem becomes a barrier.

Opportunity in online and search, therefore, is significant. Vendors who pitch another fancy search algorithm are missing the train in law enforcement, financial services, competitive intelligence, and medical research. Going fast is no longer a way to add value. Merging different time frameworks is a more interesting area to me.