Why Are $800 Billion in Document Assets Wasted Annually?

April 2006

by Michael K. Bergman

About BrightPlanet

BrightPlanet Corporation, Sioux Falls, SD, Washington, DC, and New York City, is a private venture-backed company founded in 1999, though its technology legacy extends to the early 1980s. BrightPlanet's mission is to obtain value from document assets. BrightPlanet is the leader in deep document content and the development of innovative ways to efficiently search, monitor and manage all Internet and internal content. BrightPlanet offers unique technologies for discovery, harvest, management, aggregation, qualification, and classification of this content.

BrightPlanet products include collaborative, high-productivity solutions for professional knowledge workers as well as comprehensive federated portal sites that provide single access to pre-qualified content for individual companies, public agencies and associations. Prominent customers include the intelligence community, state and federal agencies, and Fortune 2000 companies. Three patents are pending with four granted for various aspects of BrightPlanet's automation, discovery and content management technologies.

BrightPlanet's series of technology white papers were prepared by our development group under the direction of Michael K. Bergman, chief technology officer.

BrightPlanet, Deep Query Manager and DQM Publisher are trademarks of BrightPlanet Corporation. All other trademarks noted are owned by their respective parties.

BrightPlanet®

SUMMARY

It is a tragedy of no small import when $800 billion in readily available savings from creating, using and sharing documents is wasted in the United States *each year*. How can waste of such magnitude – literally equivalent to almost 8% of gross domestic product or more than 40% of what the nation spends on health care [1] – occur right before our noses? And how can this waste occur so silently, so insidiously, and so ubiquitously that none of us can see it?

Why is no one seeing the manifest $800 B annual waste in poor document use?

Let me repeat. The topic is $800 billion in annual waste in the U.S. alone, perhaps equivalent to as much as $3 trillion globally, that can be readily saved *each year* with improved document management and use. Achieving these savings does not require any Herculean efforts, simply focused awareness and the application of best practices and available technology. As the T.D Waterhouse commercial states, "You can do this."

This white paper is itself the result of an earlier white paper I authored also under BrightPlanet [2] sponsorship. Entitled, *Untapped Assets: The $3 Trillion Value of U.S. Enterprise Documents,*[3] that paper documented via scores of reputable references and data analysis tables the magnitude of the poor use of document assets within enterprises. The paper was perhaps the most comprehensive look to date at the huge expenditures document creation and use occupy within our modern knowledge economy, and first quantified the potential $800 billion annual savings in overcoming readily identifiable waste.

Simply documenting the magnitude of expenditures and savings was mind-blowing. But what actually became more perplexing was why the scope of something so huge and so amenable to corrective action was virtually invisible to policy or business attention. The vast expenditures and potential savings surfaced in the research quite obviously begged the question: Why is no one seeing this?

I then began a new series on my personal blog – mkbergman.com – to begin looking at factors as to why document use savings may fit other classes of "big" problems such as high blood pressure as a silent killer, global warming from odorless and colorless greenhouse gasses, or the under funding of cost-effective water systems and sanitation by international aid agencies. There seems to be something more difficult involving ubiquitous problems with broadly shared responsibilities.

The series began on my blog in October of last year and, at least for the purposes of this paper, concludes with this summary. Somehow, however, I suspect the issues touched on in this series are still poorly addressed and will remain a topic for some time to come.

The series looked at four major categories, with this summary's wrap-up:

---

[1] According to the U.S. Dept of Health and Human Services, the nation spent $1.9 trillion on health care in 2004; see http://www.cms.hhs.gov/NationalHealthExpendData/02_NationalHealthAccountsHistorical.asp#TopOfPage.
[2] See http://www.brightplanet.com/.
[3] Michael K. Bergman, "Untapped Assets: The $3 Trillion Value of U.S. Enterprise Documents," *BrightPlanet Corporation White Paper*, July 2005, 42 pp. The paper contains 80 references, 150 citations, and many data tables. See http://www.brightplanet.com/technology/whitepapers.asp.

![White Paper - BrightPlanet header banner]

I can regretfully conclude that I really haven't put my finger on the compelling reason(s) as to why broad, universal problems such as document use and management remain a low priority and have virtual no visibility ***despite the very real savings that current techniques and process can bring***. But I think some of the relevant factors are covered in these topics.

Document waste is one of those "big" problems that seems to defy resolution

The arguments in Part I are pretty theoretical. They firstly ask if it is in the public interest to strive for improvements in "information" efficiency, some of which may be applicable to the private sector with possible differentials in gains. They secondly question the rhetoric of "information overload" that can lead to a facile resignation about whether the whole "information" problem can be meaningfully tackled. One dog that won't hunt is the claim that computers intensify the information problem of private gain *v.* societal benefit because now _more stuff_ can be processed. Such arguments are diversions that obfuscate deserved and concentrated public policy that can bring real public benefits – and soon. Why else do we not see tax and economic policies that can enrich our populace by hundreds of billions of dollars annually?

Part II argues that barriers to collaboration, many cultural but others social and technical, help to prevent a broader consensus about the importance of document reuse (read: "information" and "knowledge"). Document reuse is likely the single largest reservoir of potential waste reductions. One real problem is the lack of top leadership within the organization to encourage collaboration and efficiencies in document use and management through appropriate training and rewards, and commitments to install effective document infrastructures.

Part III re-visits for prior failings and high costs in document or content initiatives within the enterprise. Perceptions of past difficulties color the adoption of new approaches and technologies. The lack of standards, confusing terminology, some failed projects, immaturity of the space, and the as-yet emergence of a dominant vendor have prevented more widespread adoption of what are clearly needed solutions to pressing business content needs. There are no accepted benchmarks by which to compare vendor performance and costs. Document use and management software can be considered to be at a similar point to where structured data was at 15 years ago at the nascent emergence of the data warehousing market. Growth in this software market will require substantial

Focused awareness and best practices can provide billions in dollar savings in document use and management.

improvements in TCO and scalability, among a general increase in awareness of the magnitude of the problem and available means to solve it.

Part IV looks at what might be called issues of attention, perception or psychology. These factors are limiting the embrace of meaningful approaches to improve document access and use and achieve meaningful cost savings. Document intelligence and document information automation markets still fall within the category of needing to "educate the market." Since this category is generally dreaded by most venture capitalists (VCs), that perception is *also* acting to limit the financing of fresh technologies and entrepreneurial ship.

The conclusion is that public and enterprise expenditures to address the wasted document assets problem remain comparatively small, with growth in those expenditures flat in comparison to the rate of document production. Hopefully, this white paper and the series that spawned it – plus, also hopefully, ongoing dialog and input from the community – can continue to bring attention and focus to the various ways that technology, people, and process can bring real document savings to our collective pocketbooks.

Table of Contents

I. THE NATURE OF INFORMATION AND ITS OWNERSHIP IN THE COMMONS

A column late last year by David Wessel in the *Wall Street Journal* argues that "Better Information Isn't Always Beneficial."[4] His major arguments can be summarized as follows:

1. Having more information available is generally good
2. Having some information available is clearly bad (to terrorists, privacy violations)
3. However, other information is also bad because it may advance the private (profit) interest but not that of society, and
4. Computers are worsening Argument #3 by reducing the cost of processing information.

Wessel claims that computers are removing limits to information processing that will force society to wrestle with practical issues of inequities that seemed only theoretical a generation ago. Though this article is certainly thought provoking, and therefore of value, it is wrong on epistemological, logical, and real-world grounds.

Epistemology

All of us at times confuse data or content with the concept of information. We often describe current circumstances with terms such as "information overload" or "infoglut." This confusion often extends to the economics literature in how it deals with the value of "information." Most researchers or analysts in knowledge management acknowledge this hierarchy of value in the knowledge chain:

data (or content) → information → knowledge (actionable)

This progression also represents a narrowing flow or 'staging' of volume. The amount of total data always exceeds information; only a portion of available information is useful for knowledge or action.

Rather than provide "definitions" of these terms, which are not universally agreed, let's use the example of searching on Google to illustrate these concepts:

- *Data* – the literally billions of documents contained within Google's search index
- *Information* – subsets of this data appropriate to the need or topic at hand. While this sounds straightforward, depending on how the user queries and its precision, the "information" returned from a search may have much

> It is a false argument that "improved" information should be seen through the lens of inequitable private gain.

---

[4] Sept. 22, 2005. See http://online.wsj.com/public/article/0,,SB112734060508547844-4Gx7dIqhluIyFxPiOh2mkB8mXrY_20060922,00.html?mod=blogs.

lower or higher percentages of useful information value, as well as a great range of total possible results

- *Knowledge* – Google obviously does not provide knowledge *per se*, but, depending on user review of the information from more-or-less precise search queries and information duplication or not, knowledge may come about through inspection and learning of this information.

The concept of staging and processing is highly useful here. For example, in the context of a purposeful document repository, initial searches to Google and other content aggregation sites – even with a query or topic basis – could act to populate that repository with *data*, which would then need to be mined further for useful *information* and then evaluated for supplying *knowledge*. Computers always act upon *data*, whether global in a Google case or local in a local repository case, and whether useful information is produced or not.

Wessel and indeed most economists co-mingle all three terms in their arguments and logic. By missing the key distinctions, fuzzy thinking can result.

A Philosophical or Political Polemic?

First, I will not take issue with Wessel's first two arguments above. Rather, I'd like to look at the question of Argument #3 that some information is "bad" because it delivers private *vs.* societal value.

Possible private gain from "improved" information does not negate public policy support.

His two economist references in the piece are to Arrow and Hirshleifer. As Wessel cites Hirshleifer: "The contrast between the private profitability and the social uselessness of foreknowledge may seem surprising," the late economist Jack Hirshleifer wrote in 1971. But there are instances, he argued, where "the community as a whole obtains no benefit ... from either the acquisition or the dissemination (by resale or otherwise) of private foreknowledge."

Yet Hirshleifer had a very specific meaning of "private foreknowledge," likely not in keeping with the Wessel arguments. The Hirshleifer [5] reference deals entirely with speculative investments and the "awareness" or not (knowledge; perfect information) of differing economic players. According to the academic reviewer Morrison [6]:

> *In Hirshleifer's terms, 'private foreknowledge' is information used to identify pricing errors after resource allocation is fixed. Because it results in a pure wealth transfer but is costly to produce, it reduces social surplus. … As opposed to private foreknowledge, 'discovery*

---

[5] J. Hirshleifer, "The Private and Social Value of Information and the Reward to Inventive Activity," *American Economic Review*, Vol. 61, pp. 561-574, 1971.

[6] A. D. Morrison, "Competition and Information Production in Market Maker Models," forthcoming in the *Journal of Business Finance and Accounting*, Blackwell Publishing Ltd., Malden, MA. See the 20 pp. online version, http://users.ox.ac.uk/~bras0541/12_jbfa5709.pdf#search='Hirshleifer%20private%20foreknowledge'

information' is produced prior to the time resource allocation is fixed, and because it positively affects resource allocation it generally increases social surplus. But even discovery information can be overproduced because optimal expenditures on discovery information will inevitably be subject to pricing errors that can be exploited by those who gather superior information. In cases of both fixed and variable resource allocation, then, excess search has the potential to occur, and private parties will adopt institutional arrangements to avoid the associated losses.

Hmmm. What? Is this actually in keeping with the Wessel arguments?

Wessel poses a number of examples where he maintains the disconnect between private gain and societal benefit occurs in the information commons. The examples he cites are:

- Assessing judges as to how they might rule on patent infringement cases
- Screening software for use in jury selections
- Demographic and voting information for gerrymandering U.S. congressional districts
- Weather insurance for crops production.

In virtually all cases of private gain from "improved" information, public benefits accrue as well.

These examples are what Wessel calls "the sort of information that Nobel laureate Kenneth Arrow labeled 'socially useless but privately valuable.' It doesn't help the economy produce more goods or services. It creates nothing of beauty or pleasure. It simply helps someone get a bigger slice of the pie."

According to Oldrich Kyn, an economics professor emeritus from Boston University, Joseph Stiglitz, another Nobel laureate, took exception to Arrow's thesis regarding information in the areas of market socialism and neoclassical economics [7] as shown by these Stiglitz quote excerpts:

The idea of market socialism has had a strong influence over economists: it seemed to hold open the possibility that one could attain the virtues of the market system—economic efficiency (Pareto optimality)—without the seeming vices that were seen to arise from private property.

The fundamental problem with [the Arrow—Debreu model] is that it fails to take into account ... the absence of perfect information—and the costs of information—as well as the absence of certain key risk markets ....

---

[7] See http://econc10.bu.edu/economic_systems/Theory/NonMarx_Socialism/Market%20Socialism/Stiglitz_marksoc_neoclass.htm.

The view of economics encapsulated in the Arrow—Debreu framework ... is what I call 'engineering economics' .... economics consisted of solving maximization problems ...The central point is that in that model there is not a flow of new information into the economy, so that the question of the efficiency with which the new information is processed—or the incentives that individuals have for acquiring information—is never assessed.... the fundamental theorems of welfare economics have absolutely nothing to say about ... whether the expenditures on information acquisition and dissemination— is, in any sense, efficient.

Stiglitz in his own online autobiography states:[8] "The standard competitive market equilibrium model had failed to recognize the complexity of the information problem facing the economy - just as the socialists had. Their view of decentralization was similarly oversimplified." Grossman and Stiglitz [9] more broadly observe "that perfectly informative financial markets are impossible and … the informativeness of prices is inversely related to the cost of information."

Care must be exercised in the application of the economists' standard of "perfect information" to actual real-world circumstances.

I am no economist, but reading the original papers suggests to me a narrower and more theoretical focus than what is claimed in Wessel's arguments. Indeed, the role of "information" is both central to and nuanced within current economic theory, the understanding of which has progressed tremendously in the thirty years since Wessel's original citations. By framing the question of private (profit) versus societal good, Wessel invokes an argument based on political philosophy and one seemingly "endorsed" by Arrow as a Nobel laureate. Yet as Eli Rabett commented on the Knowledge Crumb's Web site, "[the Wessel thesis] is a communitarian argument which has sent Ayn Rand, Alan Greenspan, Newt Gingrich and Grover Norquist to spinning in their graves."[10]

Logical Fallacies

Even if these philosophical differences could be reconciled, there are other logical fallacies in the Wessel piece.

In the case of assessing the performance of patent judges by crunching information that can now be sold cost-effectively to all participants, Wessel asks, "But does it increase the chances that the judge will come to a just decision?" The logical fallacies here are manifest:

- Is the only societal benefit one of having the judge come to a just decision or, also potentially, society learning about judicial prejudices singly or

---

[8] See
http://economia.unipv.it/marketing_high_tech/high_tech_letture/00.%20lezione%20iniziale/i%20nobel%20in%20economia/Web%20Devil%20Downloads/wwwnobel.se/economics/laureates/2001/stiglitz-autobio.html.
[9] S.J. Grossman and J.E. Stiglitz, "On the Impossibility of Informationally Efficient Markets," *American Economic Review*, Vol. 70, No. 3, pp. 393–403, June 1980.
[10] http://sciencepolicy.colorado.edu/prometheus/archives/science_policy_general/000583is_better_informatio.html

collectively or setting new standards in evaluating or confirming judicial candidates?

- No new information has been created by the computer. Rich litigants could have earlier gone through expensive evaluations. Doesn't cost-effective information democratize this information?
- Is not broad information availability an example of desired transparency as cited by Knowledge Crumbs?

Wessel raises another case of farmers now possibly being able to buy accurate weather forecasts. But he posits a resulting case where the total amount of food available is unchanged and insurance would no longer be necessary. Yet, as Mark Bahner points out, this has the logical fallacies of: [11]

- The amount of food available would NOT be "unchanged" if farmers knew for certain what the weather was going to be. Social and private benefits would also accrue from, for example, applying fertilizers when needed without wasteful runoffs
- Weather knowledge would firstly never be certain and other uncertainties (pests, global factors, etc.) would also exist. Farmers understand uncertainty and would continue to hedge through futures or other forms of insurance or risk management.

Logical fallacies can result from mixing economic theory with real-world circumstances.

The real logical fallacies relate to the assumption of perfect information and complete reduction of uncertainty. No matter how much data, or how fast computers, these factors will never be fully resolved.

Practical Role of the Computer

Wessel concludes that by reducing the cost of information so much, computers intensify the information problem of private gain *v.* societal benefit. He uses Arrow again to pose the straw man that, "Thirty years ago, Mr. Arrow said the fundamental problem for companies trying to get and use information for profit was 'the limitation on the ability of any individual to process information.'"

But as Knowledge Crumbs notes,[12] computers may be able to process more data than an individual, but they are still limited and always will be. Moreover there will remain the Knowledge Problem and the SNAFU principle to make sure that humans are not augmented perfectly by their computers. Knowledge Crumbs concludes:

> *"The issue with knowledge isn't that there is too much, it is that we lack methods to process it in a timely fashion, and processing*

---

[11] See http://markbahner.typepad.com/.
[12] See http://crumbtrail.org/mt

*introduces defects that sometimes are harmful. When data is
reduced or summarized something is lost as well as gained."*

*The dire specter
of computers
churning
through more
data is a
Luddite view of
their role in
improved
knowledge and
decisionmaking.*

The speed of crunching data or computer processing power is not the issue. Use
and misuse of information will continue to exist, as it has since mythologies were
passed by verbal allegory by firelight.

Importance to Document Assets

So, why does such a flawed polemic get published in a reputable source like the
Wall Street Journal? There are real concerns and anxieties underlying this
Wessel piece and it is always useful to stimulate thought and dialog. But, like all
"information" that the piece itself worries over, it must be subjected to scrutiny,
testing and acceptance before it can become the basis for action. The failure of
the Wessel piece to pass these thresholds itself negates its own central arguments.

Better that our pundits should focus on things that can be improved such as why
there is so much duplication, misuse and overlooking of available information.
These cost the economy plenty, totally swamping any of Wessel's putative private
benefits were they even correct.

Let's focus on the real benefits available today through computers and
information to improve society's welfare. Setting up false specters of computer
processing serving private greed only takes our eye off the ball.

II. BARRIERS TO COLLABORATION

Collaboration is important. BrightPlanet's earlier research paper on the waste
associated with enterprise document use (or lack thereof) indicated that $690
billion a year alone could be reclaimed by U.S. enterprises from better sharing of
information. That represents 88% of the total $780 billion wasted annually.

The issue of poor document use within the organization is certainly not solely a
technological issue, and is likely due more to cultural and people issues, not to
mention process. At BrightPlanet, we have been attempting a concerted
"document as you go" commitment by our developers and support people, and
have worked hard to put in place Wiki and other collaboration tools to minimize
friction.

But friction remains, often stubbornly so. At heart, the waste and misuse of
document assets within organizations arises from a complex set of these people,
process and technology issues.

Dave Pollard, the inveterate blogger on KM and other issues, provided a listing of 16 reasons of 'Why We Don't Share Stuff.'[13]   That thoughtful posting received a hail storm of responses, which caused Dave to later update that listing to 23 reasons under a broader post called 'Knowledge Sharing & Collaboration 2015'[14] (a later post upped that amount to 24 reasons). This is great stuff, and nearly complete grist for laying out the reasons – some major and some minor – why collaboration is often difficult.

I have taken these reasons, plus some others I've added of my own or from other sources,[15] and have attempted to cluster them into the various categories below.[16]

People, Behavior and Psychology

Barriers to collaboration – cultural, social and technical – help to prevent consensus about the importance of document reuse.

These are possible reasons why collaboration fails due to people, behavior or psychological reasons. They represent the majority (56%) of reasons proffered by Pollard:

- People find it easier and more satisfying to reinvent the wheel than re-use other people's 'stuff' (*)
- People only accept and internalize information that fits with their mental models and frames (Lakoff's rule) (*)
- Some modest people underestimate the value of what they know so they don't share (*)
- We all learn differently (some by reading, some by listening, some by writing down, some by hands-on), and people won't internalize information that isn't in a format attuned to how they learn (one size training doesn't fit all) (*)
- People grasp graphic information more easily than text, and understand information conveyed through stories better than information presented analytically (we learn by analogy, and images and stories are better analogies to our real-life experiences than analyses are) (*)
- People cannot readily differentiate useful information from useless information (* split)
- Most people want friends and even strangers to succeed, and enemies to fail; this has a bearing on their information-sharing behavior (office politics bites back) (*)
- People are averse to sharing information orally, and even more averse to sharing it in written form, if they perceive *any* risk of it being misused or misinterpreted (the better safe than sorry principle) (*)

---

[13] See Dave Pollard's blog from Sept. 19, 2005, http://blogs.salon.com/0002007/2005/09/19.html#a1278.
[14] See Dave Pollard's blog from Sept. 29, 2005 http://blogs.salon.com/0002007/2005/09/29.html#a1288..
[15] There have been some other interesting treatments of barriers to collaboration including that by Carol Kinsey Goman's "Five reasons people don't tell what they know" (see http://destinationkm.com/articles/default.asp?ArticleID=960) and Jack Vinson's "Barriers to knowledge sharing" (http://blog.jackvinson.com/archives/2005/09/20/barriers_to_knowledge_sharing.html).
[16] Pollard's initial 16 reasons are shown with a single symbol (*); the next 8 additions with a double symbol (**). All remaining reasons added by me have three symbols (***).

Management has a key role in overcoming collaboration barriers.

- People don't take care of shared information resources (Tragedy of the Commons again) (*)
- People seek out like minds who entrench their own thinking (leads to groupthink) (**)
- Introverts are more comfortable wasting time looking for information rather than just asking (sometimes it's just more fun spending 5 hours on secondary research, or doing the graphics for your PowerPoint deck by trial and error, than asking a good source to help you in less time) (**)
- People won't (or can't) internalize information until they need it or recognize its value (most notably, information in e-newsletters is rarely absorbed because it rarely arrives just at the moment it's needed) (**)
- People don't know what others who they meet know, that they could benefit from knowing (a variant on the old ""don't know what we don't know" – "we don't know what we don't know that they do") (**)
- If important news is withheld or sugar-coated, people will 'fill in the blanks' with an 'anti-story' worse than the truth (**)
- Experts often speak in jargon or "expert speak." They don't know they aren't communicating, and non-experts are afraid to ask (***).

Management and Organization

These are possible reasons why collaboration fails due to managerial or organization limits. They represent about one-fifth (20%) of the reasons proffered by Pollard:

- Bad news rarely travels upwards in organizations (shoot the messenger, and if you do tell the boss bad news, better have a plan to fix it already in motion) (*)
- People share information generously peer-to-peer, but begrudgingly upwards ("more paperwork for the boss"), and sparingly downwards ("need to know") in organizational hierarchy – it's all about trust (*)
- Managers are generally reluctant to admit they don't know, or don't understand, something (leads to oversimplifying, and rash decision-making) (*)
- Internal competition can mitigate against information sharing (if you reward individuals for outperforming peers, they won't share what they know with peers) (*)
- The people with the most valuable knowledge have the least time to share it (**)
- Management does not generally appreciate its role in overcoming psychology and personal behaviors that limit collaboration (***)

- Management does not appreciate the tremendous expense, revenue, profitability and competitiveness implications from lack of collaboration (***)
- Management does not know training, incentive, process, technology or other techniques to overcome limits to collaboration (***)
- Earlier organization attempts with CIOs, CKOs, etc., have not been sustained or were the wrong model for internalizing these needs within the organization (***)
- Organizational job titles still reinforce managerial *v.* expertise in status and reward (***)
- Hiring often inadequately stresses communication and collaboration skills, and does not provide in-house training if still lacking (***).

Poor infrastructure, process and training can be corrected to improve collaboration.

Technology, Process and Training

These are possible reasons why collaboration fails due to technology, process or training. They represent about one-eighth (12%) of the reasons proffered by Pollard, but also realize his original premise was on human or psychological reasons, so it is not surprising this category is less represented:

- People know more than they can tell (some experience you just have to *show*) and tell more than they can write down (composing takes a lot of time) (Snowden's rule) (*)
- People feel overwhelmed with content volume and complex tools (info overload, and poverty of imagination) (* split)
- People will find ways to work around imposed tools, processes and other resources that they don't like or want to use (and then deny it if they're called to account for it) (**)
- Employees lack the appreciation for the importance of collaboration to the success of their employer and their job (***)
- Most means for "recording" the raw data and information for collaboration have too much "friction" (***)
- There needs to be clear divisions between "capturing" knowledge and information and "packaging" it for internal or external consumption (***)
- Single-source publication techniques suck (***)
- Testing, screening, vetting and making new technology or process advantages is generally lacking (***).

Cost, Rewards and Incentives

These are also possible reasons why collaboration fails due to the cost and rewards structure, again about one-eighth (12%) of the reasons proffered by Pollard. Again, realize his original premise was on human or psychological reasons, so it is not surprising this category is less represented:

- The true cost of acquiring information (time wasted looking for it) and the cost of not knowing (Katrina, 9/11, bird flu, etc.) are greatly underestimated in most organizations (*)
- Rewards for sharing knowledge don't work for long (*)
- People value information they have paid for more highly than what they get free from their own people (thus the existence of the consulting industry) (from James Governor) (**)
- Document solutions are perceived to be high cost and risk (***)
- Performance pay is generally not linked to collaboration goals (***).

Insights and Quibbles

Use incentives and rewards to create a collaborative culture.

There are some 25 reasons provided by Dave and his blog respondents, actually closer to 40 when my own are added, that represent a pretty complete compendium of "why collaboration fails." Though I can pick out individual ones of these to praise or criticize that would miss the point.

The objective is neither to collect the largest numbers of such factors or to worry terribly about how they are organized. But there are some interesting insights.

Clearly, human behavior and psychology provides the baseline for looking at these questions. Management's role is to provide organizational structure, incentives, training, pay and recognition to reward the collaborative behavior it desires and needs. Actually, management's challenge is even greater than that since in most cases upper level managers don't yet have a clue as to the importance of the underlying information nor collaboration around it.

Like in years past, leadership for these questions needs to come from the top. The disappointments of the CIO and CKO positions of years past need to be looked at closely and given attention. The idea of these positions in the past was not wrong; what was wrong was the execution and leadership commitment.

Organizations of all types and natures have figured out how to train and provide incentives to employees for difficult duties ranging from war to first response to discretion. Putting in place reward and training programs to encourage collaboration – despite poor performance today – should not be so difficult in this light.

I think Dave brings many valuable insights into such areas as people being reluctant to reinvent the wheel but liking creative design, or without some sense of ownership a collaboration repository is at risk, or people are afraid to look stupid, or some people communicate better orally *v.* in written form, etc. These are, in fact, truisms of human diversity and skill differences. I believe firmly if organizations want to purposefully understand these factors they can still design

reward, training and recognition regimens to shape the behavior desired by that organization.

The real problem in the question of collaboration within the enterprise begins at the top. If the organization is not aware and geared to address human nature with appropriate training and rewards, it will continue to see the poor performance around collaboration that has characterized this issue for decades.

III. ENTERPRISE 'SOLUTIONS' ARE TOO EXPENSIVE

Past enterprise content "failures" are a major impediment to new solutions.

As noted by the Nobel laureate economist Herbert Simon more than 30 years ago:

> *What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention, and a need to allocate that attention efficiently among the overabundance of sources that might consume it. . . . The only factor becoming scarce in a world of abundance is human attention.[17]*

Document assets are poorly utilized at all levels and within all departments within enterprises. Prior sections addressed whether the root causes of this poor use were due to the nature of private v. public information or due to managerial and other barriers to collaboration. This part investigates whether high software and technology costs matched with poor performance is a root cause.

The Document Situation Within U.S. Enterprises

Document creation represents about $3.3 billion in annual costs to U.S. enterprises, or about 30% of gross national product, $800 billion of which can be reclaimed through better access, recall and use of these intellectual assets. For the largest U.S. firms, annual benefits from better document use average about $250 million per firm.[18]

Perhaps at least 10% of an enterprise's information changes on a monthly basis.[19] A 2003 UC Berkeley study on "How Much Information?" estimated that more than 4 billion pages of *internal* office documents are generated annually in the U.S. with archival value. The percentage of unstructured (document) data to the

---

[17] H.A. Simon, "Designing Organizations for an Information Rich World," in M. Greenberger (ed.), *Computers, Communications, and the Public Interest*, pp. 38-52, July 1971, The Johns Hopkins University Press, Baltimore, MD. Reprinted in: H.A. Simon, *Models of Bounded Rationality and Other Economic Topics*, Vol. 2.Collected Papers, The MIT Press, Cambridge, MA, May 1982.

[18] M.K. Bergman, "Untapped Assets: The $3 Trillion Value of U.S. Enterprise Documents," *BrightPlanet Corporation White Paper*, July 2005, 43 pp. See http://www.brightplanet.com/technologydocumentvalue.asp.

[19] Delphi Group, "Taxonomy & Content Classification Market Milestone Report," *Delphi Group White Paper*, 2002. See http://delphigroup.com.

total amount of enterprise data is estimated at 85% and growing.[20] Year-on-year office document growth rates are on the order of 22%.

Based on these averages, a 'typical' document may cost on the order of $380 each to create.[21] Standard practice suggests it may cost on average $25 to $40 per document simply for filing.[22] Indeed, labor costs can account for up to 30% of total document handling costs.[23] Of course, a "document" can vary widely in size, complexity and time to create, and therefore its individual cost and value will vary widely. An invoice generated from an automated accounting system could be a single page and be produced automatically in the thousands; proposals for very large contracts can take tens or thousands or even millions of dollars to create.

Documents – and the information contained within them – are a growing and critical source of enterprise knowledge and memory.

According to a Coopers & Lybrand study in 1993, 90 percent of corporate memory exists on paper.[24] A Xerox Corporation study commissioned in 2003 and conducted by IDC surveyed 1000 of the largest European companies and had similar findings:[25] [26]

- On average 45% of an executive's time was spent dealing with documents
- 82% believe that documents were crucial to the successful operation of their organizations
- A further 70% claimed that poor document processes could impact the operational agility of their organizations
- While 83%, 78% and 76% consider faxes, email and electronic files as documents, respectively, only 48% and 46% categorize web pages and multimedia content as such.
- Significantly, 90 to 97 percent of the corporate respondents to the Coopers & Lybrand and Xerox studies, respectively, could not estimate how much they spent on producing documents each year. Almost three quarters of them admit that the information is unavailable or unknown to them.

These statistics apply to the perhaps 20 million knowledge workers within US firms (though other estimates have ranged as high as 40 million).[27] [28] Of this number, perhaps nearly one million have job responsibilities solely devoted to

---

[20] P. Lyman and H. Varian, "How Much Information, 2003," retrieved from http://www.sims.berkeley.edu/how-much-info-2003 on December 1, 2003.

[21] M.K. Bergman, "A Cure to IT Indigestion: Deep Content Federation," *BrightPlanet Corporation White Paper*, December 2004, 40 pp. See http://www.brightplanet.com/technology/whitepapers.asp

[22] Cap Ventures information, as cited in ZyLAB Technologies B.V., "Know the Cost of Filing Your Paper Documents," *Zylab White Paper*, 2001. See http://www.zylab.com/downloads/whitepapers/PDF/21%20-%20Know%20the%20cost%20of%20filing%20your%20paper%20documents.pdf.

[23] Optika Corporation. See http://www.optika.com/ROI/calculator/ROI_roiresults.cfm

[24] As initially published in Inc Magazine in 1993. Reference to this document may be found at: http://www.contingencyplanning.com/PastIssues/marapr2001/6.asp

[25] J. Snowdon, *Documents – The Lifeblood of Your Business?*, October 2003, 12 pp. The white paper may be found at: http://www.mdy.com/News&Events/Newsletter/IDCDocMgmt.pdf

[26] Xerox Global Services, *Documents - An Opportunity for Cost Control and Business Transformation*, 28 pp., 2003. The findings may be found at: http://www.sap.com/solutions/srm/pdf/CCS_Xerox.pdf

[27] Nuala Beck, *Shifting Gears: Thriving in the New Economy*, Harper Collins Publishers, Toronto, 1993.

[28] pers. comm.., Guy Cresse, Aberdeen Group, November 19, 2001.

---

content management. In the largest firms, there are likely 300 employees or more whose sole responsibility is content management.

90% to 97% of enterprises can not estimate how much they spend on documents.

The High Cost of Searching and Organizing

The average knowledge worker spends 2.3 hrs per day – or about 25% of work time – searching for critical job information, with 60% saying search is a difficult process, made all the more difficult without a logical organization to content. A USC study reported that typically only 32% of employees in knowledge organizations have access to good information about technical developments relevant to their work, and 79% claim they have inadequate information about what their competitors are doing.[29]

According to the Gartner Group, the average enterprise spends from 60 to 70% of its application development budgets creating ways to access disparate data, importantly including documents.[30] IDC estimates that enterprises employing 1,000 knowledge workers may waste well over $6 million per year each in searching for information that does not exist, failing to find information that does, or recreating information that could have been found but was not.[31] As that report stated, "It is simply impossible to create knowledge from information that cannot be found or retrieved."

Forrester reported in 2002 that 54% of Global 3500 companies relied at that time on homegrown systems to manage content.[32] One vendor cites national averages as indicating that most organizations spend from 5% to 10% of total company revenue on handling documents; Cap Ventures suggests these ranges may be as high as 6% to 15%, with the further observation that 85% of all archived documents never leave the filing cabinet.

An A.T. Kearney study sponsored by Adobe, EDS, Hewlett-Packard, Mayfield and Nokia, published in 2001, estimated that workforce inefficiencies related to content publishing cost organizations globally about $750 billion. The study further estimated that knowledge workers waste between 15% to 25% of their time in non-productive document activities.[33]

Delphi Group's research points to the lack of organized information as the number one problem in the opinion of business professionals. More than three-quarters of the surveyed corporations indicated that a taxonomy or classification

---

[29] S.A. Mohrman and D.L. Finegold, *Strategies for the Knowledge Economy: From Rhetoric to Reality, 2000,* University of Southern California study as supported by Korn/Ferry International, January 2000, 43 pp. See http://www.marshall.usc.edu/ceo/Books/pdf/knowledge_economy.pdf.

[30] Gartner Group, as reported by P. Hallett, Schemalogic Corporation, at the 2003 Enterprise Data Forum, Philadelphia, PA, November 2003. See http://www.wilshireconferences.com/EDF2003/tripreport.htm.

[31] C. Sherman and S. Feldman, "The High Cost of Not Finding Information*," International Data Corporation Report #29127*, 11 pp., April 2003.

[32] J.P. Dalton, "Enterprise Content Management Delusions," *Forrester Research Report*, June 2002. 12 pp. See http://www.forrester.com/ER/Research/Report/Summary/0,1338,14981,00.html.

[33] A.T. Kearney, *Network Publishing: Creating Value Through Digital Content*, A.T. Kearney White Paper, April 2001, 32 pp. See http://www.adobe.com/aboutadobe/pressroom/pressmaterials/networkpublishing/pdfs/netpubwh.pdf

system for documents is imperative or somewhat important to their business strategy; more than one-third of firms that classify documents still use manual techniques.

Web Sprawl: The Proliferation of Corporate Web Sites

Web sprawl – the proliferation of enterprise Web sites – is a reprise of the 1980s problems with mushrooming PC use.

Another issue facing enterprises, especially large ones, is the proliferation of Web sites or "Web sprawl." This proliferation began as soon as the Internet became popular. Here are some anecdotal examples:

- As early as 1995, DEC (purchased by Compaq and then Hewlett Packard) had 400 internal Web sites and Sun Microsystems had more than 1,000 [34]

- As reported in 2000, Intel had more than 1 million URLs on its intranet with more than 100 new Web sites being introduced each month [35]

- In 2002, IBM consolidated over 8,000 intranet sites, 680 'major' sites, 11 million Web pages and 5,600 domain names into what it calls the IBM Dynamic Workplaces, or W3 to employees [36]

- Silicon Graphics' 'Silicon Junction' company-wide portal serves 7,200 employees with 144,000 Web pages consolidated from more than 800 internal Web sites [37]

- Hewlett-Packard Co., for example, has sliced the number of internal Web sites it runs from 4,700 (1,000 for employee training, 3,000 for HR) to 2,600, and it makes them all accessible from one home, @HP [38] [39]

- Providence Health Systems recently consolidated more than 200 sites [40]

- Avaya Corporation consolidated more than 800 internal Web sites globally [41]

- The *Wall Street Journal* recently reported that AT&T has more than 10 information architects on staff to maintain its 3,600 intranet sets that contain 1.5 million public Web pages [42]

- The Department of Homeland Security is faced with the challenge of consolidating more than 3,000 databases inherited from its various constituent agencies. [43]

---

[34] D. Strom, "Creating Private Intranets: Challenges and Prospects for IS," an *Attachmate White Paper* prepared by David Strom, Inc., November 16, 1995. See http://www.strom.com/pubwork/intranetp.html.

[35] A. Aneja, C.Rowan and B. Brooksby, "Corporate Portal Framework for Transforming Content Chaos on Intranets," *Intel Technology Journal* Q1, 2000. See http://developer.intel.com/technology/itj/q12000/pdf/portal.pdf.

[36] J. Smeaton, "IBM's Own Intranet: Saving Big Blue Millions," *Intranet Journal*, Sept. 25, 2002. See http://www.intranetjournal.com/articles/200209/ij_09_25_02a.html.

[37] See http://www.wookieweb.com/Intranet/.

[38] D. Voth, "Why Enterprise Portals are the Next Big Thing," *LTI Magazine*, October 1, 2002. See http://www.ltimagazine.com/ltimagazine/article/articleDetail.jsp?id=36877.

[39] A. Nyberg, "Is Everybody Happy?" *CFO Magazine*, November 01, 2002. See http://www.cfo.com/article/1%2C5309%2C8062%2C00.html.

[40] See http://www.cubiccompass.com/downloads/Industry/Healthcare/Providence%20Health%20Systems%20Case%20Study.doc.

[41] See http://www.proudfoot-plc.com/pdf_20004-USPR1002Avayaweb.asp.

[42] *Wall Street Journal*, May 4, 2004, p. B1.

[43] pers. comm.., Jonathon Houk, Director of DHS IIAP Program, November 2003.

Corporate IT does not even know the full extent of Web site proliferation, similar to the loss of centralized control when personal PCs entered the enterprise. In that circumstance it took changes in managerial mindsets and new technology such as the PC network by Novell before control could be reasserted. Similar changes will be necessary to corral the issue of Web sprawl.

It is not uncommon for the largest companies and agencies to have thousands of unconnected internal Web sites.

The Tyranny of Expectations

Vendor hype is one of the causes of misplaced expectations, but also wrong assumptions regarding benefits and costs.

One area where this can occur is in *time savings*. Vendors and customers often use time savings by knowledge workers as a key rationale for justifying a document initiative. This comes about because many studies over the years have noted that white collar employees spend a consistent 20% to 25% of their time seeking information; the premise is that more effective search will save time and drop these percentages. However, the fact these percentages have held stable over time suggests this is the "satisficing" allocation of time to information search. Thus, while better tools to aid better discovery may lead to finding better information and making better decisions more productively – an intangible and important justification in itself – there may not result a strict time or labor savings from more efficient search.[44]

Another area is lack of awareness about *full project costs*. According to Charles Phillips of Morgan Stanley, only 30% of the money spent on major software projects goes to the actual purchase of commercially packaged software. Another third goes to internal software development by companies. The remaining 37% goes to third-party consultants.[45]

The Poor Performance of Existing Software

Past software solutions have been expensive and have not scaled well.

Some of the causes of poor document content software performance include:

- *Poor Scalability* – according to a market report published by Plumtree in 2003, the average document portal contains about 37,000 documents.[46] This was an increase from a 2002 Plumtree survey that indicated average document counts of 18,000.[47] However, about 60% of respondents to a Delphi Group survey said they had more than 50,000 internal documents in their internal environment (generally the department level). Poor

---

[44] M.E.D. Koenig, "Time Saved – a Misleading Justification for KM," *KMWorld Magazine*, Vol 11, Issue 5, May 2002. See http://www.kmworld.com/publications/magazine/index.cfm.

[45] C. Phillips, "Stemming the Software Spending Spree," *Optimize Magazine*, April 2002, Issue 6. See http://www.optimizemag.com/article/showArticle.jhtml?articleId=17700698&pgno=1.

[46] This average was estimated by interpolating figures shown on Figure 8 in Plumtree Corporation, "The Corporate Portal Market in 2003," *Plumtree Corp. White Paper*, 30 pp. See http://www.plumtree.com/portalmarket2003/default.asp.

[47] This average was estimated by interpolating figures shown on the p.14 figure in Plumtree Corporation, "The Corporate Portal Market in 2002," *Plumtree Corp. White Paper*, 27 pp. See http://www.plumtree.com/pdf/Corporate_Portal_Survey_White_Paper_February2002.pdf.

scalability and low coverage of necessary documents is a constant refrain by early enterprise implementers

- *Long Implementation Times* – though average time to stand-up a new content installation is about 6 months, there is also a 22% risk that deployment times exceeds that and an 8% risk it takes longer than one year. Furthermore, internal staff necessary for initial stand-up average nearly 14 people (6 of whom are strictly devoted to content development), with the potential for much larger head counts [48]

- *Very High Ongoing Maintenance and Staffing Costs* – a significantly limiting factor to adoption is the trend that suggests that ongoing maintenance and staffing costs exceed the initial deployment effort. Based on analysis from BrightPlanet, the table below summarizes set-up, ongoing maintenance and key metrics for today's conventional approaches versus what BrightPlanet can do. These staffing estimates are consistent with a survey of 40 installations that found there were on average 14 content development staff managing each enterprise's content portal.[49] Current practices costing $5 to $11 per document for electronic access are simply unacceptable:

	DOCUMENT BASIS	INITIAL SET-UP			MAINTENANCE	
		Staff	Mos	$/Doc	Staff	$/Doc
Current Practice	37,000	6.2	5.4	$4.861	6.4	$11.278
BrightPlanet	250,000	1.0	0.8	$0.017	0.3	$0.078
BP Advantage	**6.8 x + up**	**6.2 x**	**6.7 x**	**280.4 x**	**21.4 x**	**144.6 x**

- *Lousy Integration Capabilities* – existing document software systems integrate and interoperable poorly with other IT data systems. Document content can not be treated in isolation for the total information needs of the organization

- *High TCO* – all of these factors combine into an unacceptable total cost of ownership. High TCO and risk are simply too great to raise the priority of document management sufficiently up within IT priorities, despite the general situational awareness that "infoglut" is costing the firm a ton.

The Result: An Immature Market Space

The lack of standards, confusing terminology, some failed projects, immaturity of the space and the as-yet emergence of a dominant vendor have prevented more widespread adoption of what are clearly needed solutions to pressing business content needs. Vendors and industry analysts alike confuse the market with competing terminology, each trying to carve out a unique "message" in this ill-

---

[48] Analysis based on reference 46.
[49] M. Corcoran, "When Worlds Collide: Who Really Owns the Content," *AIIM Conference*, New York, NY, March 10, 2004. See http://show.aiimexpo.com/convdata/aiim2003/brochures/64CorcoranMary.pdf.

formed space. Read multiple white papers or inspect multiple vendor Web sites and these difficulties become evident. There are no accepted benchmarks by which to compare performance and cost implications for content management. This limitation is especially acute because, given the confusion in the market, there are no independent sources to turn to for insight and quantitative comparisons.

These issues – in combination with high costs, risks and uncertainty of performance and implementation success – lead to a very immature market at present.

Clearly, the high costs of document management software matched with poor performance and unmet expectations is one of the root causes for the $800 billion annual waste in document use within U.S. enterprises. However, as other parts of this series point out, the overall explanation for this wasteful situation is very complex with other important contributing factors at play.

Document use and management software can be considered to be at a similar point to where structured data was at 15 years ago at the nascent emergence of the data warehousing market. Growth in this software market will require substantial improvements in TCO and scalability, among a general increase in awareness of the magnitude of the problem and available means to solve it.

IV. THE PROBLEM IS TOO CLOSE FOR FOCUS

This part looks at what might be called issues of attention, perception or psychology. Interesting observations in this area come from disciplines as diverse as sales, behavioral psychology, economics and operations research.

The SPIN Rationale

One explanation for this lack of attention can be described by the fact that document problems are still in the area of *implicit needs* as opposed to *explicit needs*. In other words, the perception of the problem is still situational but has not yet become concrete in terms of bottom-line impacts.

Sometimes the most important problems are so close at hand they are easily overlooked.

In Neil Rackham's SPIN sales terminology (Situation ⟶ Problems ⟶ Implications ⟶ Needs/pay-off),[50] the enterprise document market is still at a "situational" level of understanding. Decisions to buy or implement solutions are largely strategic and limited to early adopters that are the visionaries in their market segments. The inability to express and quantify the implications of not realizing the value of document assets means that ROI analysis can not justify a deployment and market growth can not cross the chasm.

---

[50] Neil Rackham, *SPIN Selling*, McGraw Hill, 197 pp., 1988.

The situation begins with the inability to quantify the importance of both *internal* and *external* document assets to all aspects of the enterprise's bottom line. Early adopters of enterprise content software typically capture less than 1% of valuable internal documents available; large enterprises are witnessing the proliferation of internal and external Web sites, sometimes exceeding thousands; use of external content is presently limited to Internet search engines, producing non-persistent results and no capture of the investment in discovery or results; and "deep" content in searchable databases, which is common to large organizations and represents 90% of external Internet content, is completely untapped. Indeed, the issue of poor document use in an organization can be seen in terms of the figure below:

Implicit problems around poor document use need to be converted into explicit impacts with ROI measurements.
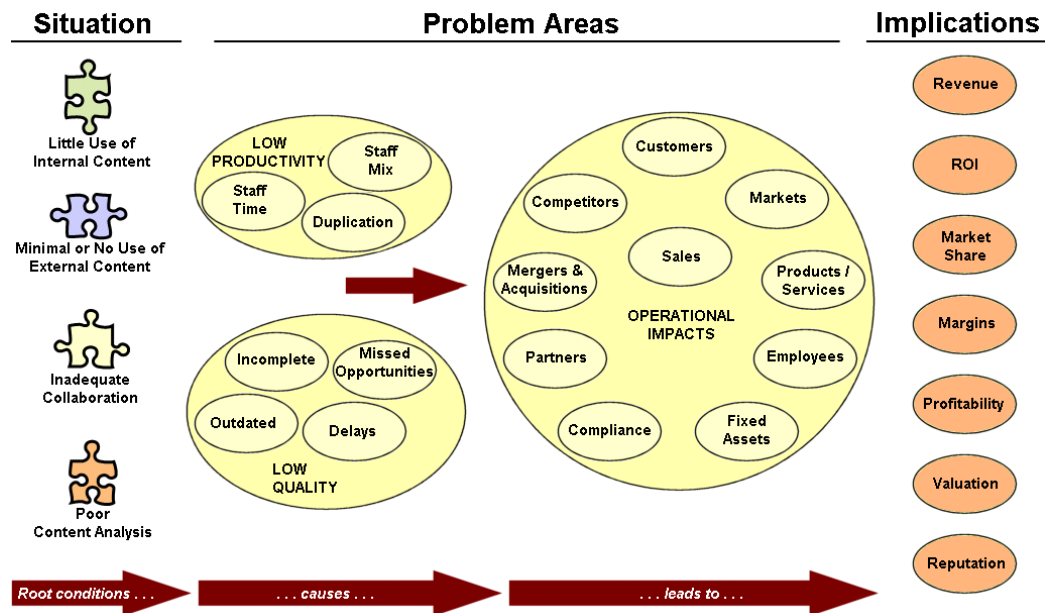
The diagram indicates that these root conditions or situations cause problems in low quality of decisions or low staff productivity. For examples, documents or proposals get duplicated without knowledge of prior effort that could be leveraged; opportunities are missed; or outdated or incomplete information is applied to various tasks. These root problems can impact virtually all aspects of the organization's operations: sales are lost; competitors are overlooked; compliance requirements are missed. These problems can lead to significant bottom-line implications from revenue and market share, to reputation and valuation and even indeed survival.

Thus, in the view of the SPIN model, the lack of attention to the issue of document assets can, in part, be ascribed to the sales or investigatory process.

Specific questions have not been posed that move the decision maker from a position of situational awareness to one of explicit bottom-line implications. There is undoubtedly truth to this observation. Sales of large document solutions to enterprises require a consultative sales approach and significant education of the market is required. As a first-order circumstance, this implies long sales lead times and the dreaded "educating the market" that most VCs try to avoid.

But there are even larger factors at play than a lack of explicitness regarding document assets.

The Ubiquitous and Obvious Are Often Overlooked

Put your index finger one inch from your nose. That is how close – and unfocused – document importance is to an organization. Documents are the salient reality of a knowledge economy, but like your finger, documents are often too close, ubiquitous and commonplace to appreciate.

The dismissal of the ubiquitous, common or obvious can be seen in a number of areas. In terms of R&D and science, this issue has been termed "mundane science" wherein most academic research topics exclude many of the issues that affect the largest number of people or have the most commonality.[51] In organizational and systems research, such issues have also been the focus of better, more rigorous problem identification and analysis techniques such as the "rational model" or the "theory of constraints" (TOC).[52]

Compounding the issue of the overlooked obvious is the lack of a quantified understanding of the problem. There is an old Chinese saying that roughly translated is "what cannot be measured, cannot be improved." Many corporate executives surely believe this to be the case for document creation and productivity.

More Specifically: Bounded Awareness

Chugh and Bazerman have recently coined a term "bounded awareness" for the phenomenon of missing easily observed and relevant data.[53] As they explain:

> *"Bounded awareness is a phenomenon that encompasses a variety of psychological processes, all of which lead to the same error: a failure to see, seek, use, or share important and relevant information that is easily seen, sought, used, or shared."*

---

[51] Daniel M. Kammen and Michael R. Dove, "The Virtues of Mundane Science," *Environment*, Vol. 39 No. 6, July/August 1997. See http://ist-socrates.berkeley.edu/~rael/Mundane_Science.pdf

[52] Victoria Mabin, "Goldratt's 'Theory of Constraints' Thinking Processes: A Systems Methodology linking Soft with Hard," *The 17th International Conference of The System Dynamics Society and the 5th Australian & New Zealand Systems Conference*, July 20 - 23, 1999, Wellington, New Zealand, 12 pp. See http://www.systemdynamics.org/conf1999/PAPERS/PARA104.PDF

[53] Dolly Chugh and Max Bazerman, "Bounded Awareness: What You Fail to See Can Hurt You," *Harvard Business School Working Paper #05-037*, 35 pp., August 25, 2005 revision. See http://www.people.hbs.edu/mbazerman/Papers/05-037.pdf

Is document waste the elephant – or gorilla – in the room?

The authors note the experiments from Simons [54] that extend Neisser's 1979 video in which a person in a gorilla costume walks through a basketball game, thumping his chest, and is clearly and comically visible for more than five seconds, but is not generally recalled by observers without prompting.

Chugh and Bazerman classify a number of these phenomena, with two most applicable to the document assets problem:

- *Inattentional blindness* – though most often applied to visual perception, this phenomenon refers to readily overlooking obvious and direct information when attention is drawn or focused elsewhere
- *System neglect* – this phenomenon is the tendency to undervalue a broader, pivotal factor to subsidiary ones, as in for example the effect of campaign finance-reform on specific political issues. In the document assets case, the general role of document access and management is neglected as a system over more readily understood specific issues such as search or spell checking. In other words, people tend to value issues that are more clearly seen as end states or outcomes.

Note the relation of these studies by behavioral psychologists to the SPIN terminology of the sales executive. Clearly, perceptual studies by scientists will lead to better understandings of market outreach.

Perceptions of Intractability?

An earlier installment in this series noted the high cost of enterprise content solutions, more generally linked to software that performed poorly and did not scale. In computer science, intractable problems are those which take too long to execute, the problem may not be computable, or we may not know *how* to solve the problem (*e.g.*, problems in artificial intelligence). *Tractable* problems can run in a reasonable amount of time for even very large amounts of input data. *Intractable* problems require huge amounts of time for even modest input sizes.[55]

Sometimes difficult problems are perceived as intractable ones, stymieing real efforts to address them.

At low scales, the efficiency of various computer algorithms is not terrible important because multiple methods can produce acceptable performance times. But at large scales whether a problem is tractable or not is not fixed: it depends critically on the efficiency of the algorithm applied to the problem. Let's take for example the issue of searching text items:

> *Take n to represent the number of keys in a list, and let O represent the order of the number of comparison operations required to find*

---

[54] See the various demos available at http://viscog.beckman.uiuc.edu/djs_lab/demos.html.
[55] Professor Constance Royden, College of the Holy Cross, course outline for CSCI 150, *Tractable and Intractable Problems*, Spring 2003. See http://mathcs.holycross.edu/~croyden/csci150spr03/notes/lec33_tractable.html.

an entry. For a small number of n items, the algorithm used is unimportant, and even a slow sequential search will work well. Sequentially searching the list until the desired match is found is O (n), or linear time. If there are 1000 items in a list, and there is an equal probability of searching for any item in the list, on average it will require n/2 = 500 comparisons to find the item (assuming all items already are on the list). A binary search works by dividing the list in half after each comparison. This is logarithmic time O (log n), much faster than linear time. For a 1000 item example it works out to about 10 comparisons. An O (1) operation, such as hashing, is applicable when some algorithm computes the item location and then retrieves it. On large lists it will significantly outperform a binary search, because it makes no comparisons. (It is a little more complicated than that because there may be collisions for the same address computed for different keys.) However, if the location is already known, even the hashing computation is unnecessary. This is what happens with direct addressing (the technique used by BrightPlanet), which will obtain the desired item in a single step.[56]

Poorly performing algorithms at large scales can require processing times for updates that take longer than the period between updates, and, thus, at least for that algorithm, are intractable at those scales.

This is one of the key and perceived problems to most document processing software at large scales – their computational inefficiencies do not allow updates to occur for the meaningful document volumes important to larger organizations. Whether the specific reasons are known by company managers and IT personnel, it is a widespread understanding – correct for most vendors – within the marketplace.

The bottom line is that most senior leaders do not understand the importance and magnitude of the document waste problem.

Since BrightPlanet's core index work engine is more efficient than other approaches (due, in part, to better sorting mechanisms as noted above, but also due to other factors), current perceived limits of intractability may not apply. However, these advances are still not generally known. Until broader understanding for how more contemporary approaches to document use and management are gained, perceptions of past poor performance will limit market acceptance.

Educating the Market

Thus, factors of awareness, attention and perception are also limiting the embrace of meaningful approaches to improve document access and use and achieve meaningful cost savings. These challenges may mean that the document intelligence and document information automation markets still fall within the

---

[56] R. L. Kruse, *Data Structures and Program Design,* Prentice Hall Press, Englewood Cliffs, New Jersey, 1987.

category of needing to "educate the market." Since this category is generally dreaded by most venture capitalists (VCs), that perception is *also* acting to limit the achievable improvements and cost savings available to this market.

But there is perhaps a very important broader question that remains open here: educating the market through the individual customer (*viz.* the SPIN sale) *vs.* educating the market through breaking market-wide bounded awareness. In fact the latter, much as what occurred with data warehousing 15-20 years ago, can create entirely new markets. This latter category should perhaps be of much greater VC interest with its accompanying potential for first-mover advantage.