Enterprise Search

Can Rich Text Analysis Make Search Work as You First Intended?

Prepared by Stephen E. Arnold, ArnoldIT.com
for Siderean Software, July 2007

Version 1.1

Enterprise Search: Can Rich Text Analysis Make Search Work as You First Intended?

With more than two-thirds of enterprise software applications running into cost problems, is enterprise search any different? Most vendors argue that search puts money on the bottomline. Our research suggests this is true—"sometimes". What can you do to make your enterprise search project a winner and a revenue turbo-charger?

Your organization needs a search system. Employees must find specific information to do their jobs. Marketing professionals want to find the files used in a brochure published two months ago. The company's attorney wants to track down emails to respond to a legal matter.

You also know that each of these search requests can trigger confusion or spark a manual search for a particular document or item of information.

You are not alone. Our research indicates that enterprise search systems are acceptable at best and often a source of frustration for many system users in an organization.[1]

This white paper tackles a murky, poorly understood aspect of enterprise search. In this analysis you will learn:

- What managers responsible for search can do to improve or "turbo charge" their existing enterprise search system
- The trade offs suggested by vendors, usually "rip and replace" or "in-place upgrades" to your search system
- The costs associated with improving your search system
- A series of recommendations you can use as a way to identify pitfalls and specific best practices appropriate for your specific situation.

As you work through this white paper, you will find it helpful to examine the condensed business analysis for acquiring an enterprise search system. Remember, that cost remains on your organization's books. You will also find a financial model that you can use to guide you in understanding the costs associated with adding important new features to your search engine.

Vendors alter their license fees frequently. Hardware prices continue to drift downwards. Outsourcing opportunities change frequently. What does not change are the tasks required to improve a search system.

1.The author of this white paper is Stephen E. Arnold who wrote the first, second, and third editions of Enterprise Search Report. He also is the author of The Google Legacy. Information about his firm's research activities is located at www.arnoldit.com

What Almost Everyone Faces

Search is ubiquitous. With more electronic information flowing into an organization's storage devices, an employee cannot do work unless he or she can locate a particular document or a specific fact about a customer's order.

Without search, there is not much work an employee, manager, or company president can perform. When search fails, everyone falls back on voice telephone calls, digging through paper files, and asking colleagues, "What did Smith order last month?"

In our research, we uncovered a little-known fact. Most organizations with more than $300 million in revenue have five or more search systems. None of them scores high in user satisfaction. In our follow up interviews, employees want system to work "like Google".

There is a major problem with the assumption that enterprise search works "like Google." The Web search available from www.google.com depends on links and user clicks to pinpoint the most relevant page among a list of documents.

In an organization, most documents may be looked at once or twice. To compound the problem for popularity-based search systems, documents rarely contain links nor do other documents include hyperlinks to other documents. But the real problem is the employee's need to have access to information pigeon-holed in an organization's Oracle or SQLServer database, the enterprise applications from SAP and IBM, and sometimes the legacy systems that have been running payroll for more than 25 years.

Today enterprise software systems include a search tool, often Autonomy or Fast Search & Transfer technology. Microsoft builds search into most of its current products. Databases from DB2 to MySQL include search technology. In some organizations, departments may have a search appliance, a version of the open source solution Lucene. The list goes on.

Bottomline? Many systems, low user satisfaction, and work-arounds galore. In one $500 million dollar safety supply company, the customer support staff relied on "sticky notes" on the edges of monitors, walls, and desktops to keep the "must have" information instantly available. User satisfaction with that company's search system was average. Usage was low, a signal that the enterprise search solution is actually a problem. The senior vice president of technology told us

> We have a search system for customer support. No one uses it. We have a search system for our general marketing and administrative information, and I can't find documents I wrote and sent a client two days ago. When I want to pull data from our warehouse, I have to get a person from IT [information technology] to help me pull the data. We're spending millions and none of these systems does what we need.

You can hear the frustration in this executive's statement. Our analysis of user assessment of search systems suggests that a rating of average may translate to awful. Many people don't use the search system; others want to offer a reasonable judgment about an expensive system their fellow employees have deployed.

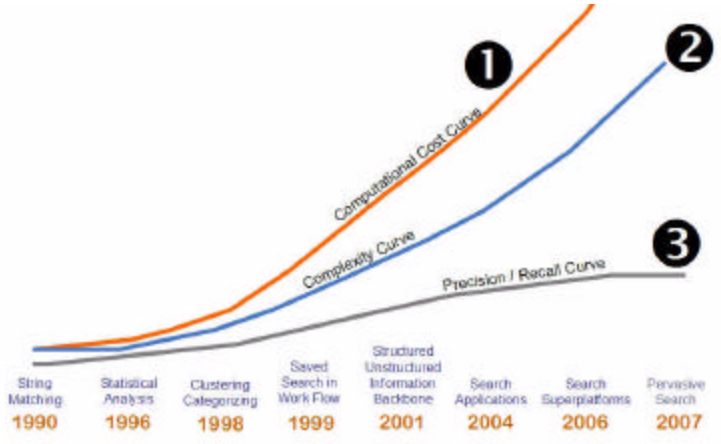
In one focus group, a program manager said,

> I have to use two different systems. Even then I don't always find what I need. It's easier to find information on the Internet via Google than it is to find information in this organization.

As information technology departments struggle to keep systems up and running, there's a growing need to "fix" search. The idea of removing one or more of the existing systems and starting over again is, as one systems professional at a drug company in Philadelphia, PA said,

> Sure we talk about getting a new system. But we don't have the resources or the time to pull this off. My colleagues and I keep the present systems up and operational. Search and retrieval is very important here, but we have to live with what we have.

What can an overburdened IT staff do to improve search? How can the CFO use available funds and deliver a shot to the bottomline when search improves revenues? Because the cost for search has grown over time, what can we do to deliver better results without breaking our budget? What solution is available to keep the status quo intact while delivering better usability and improved search results?



Curve one shows that over time, the cost of search infrastructure, maintenance, and programming goes up. In fact, it has risen more quickly than the savings from the decline in server and similar hardware. Curve two shows that the complexity of today's systems is going up, due in part to user expectations (the "Google effect") and exogenous factors like regulatory requirements. Curve three shows that in terms of precision and recall ("findability") most systems are improving but at a more leisurely pace.

The Reality

Several years ago, organizations would license another enterprise search system. The logic was, "Maybe this vendor's product will deliver the results we need." In the last two years, the differences among major vendors of enterprise search system have undergone an interesting transformation.

Our analysts track more than 50 enterprise search systems. The differences between systems is often difficult for experts to discern. Even some solutions based on the open source Lucene engine an open source solution like Lucene deliver precision and recall comparable to the six- and seven-figure systems available from such companies as Autonomy, Endeca, Fast Search & Transfer, IBM, Microsoft, Oracle, and SAP, among others. To add to the confusion, virtually every vendor of enterprise search systems asserts that their systems:

- Classify every document
- Operate without the need for human intervention in the indexing process
- Deliver processing performance sufficient to keep pace with the changes and additions to the content processed by the search system
- Allow quick and easy customizing of the search results, system parameters, security settings, and other routine housekeeping tasks.

These assertions are accurate if—and this is a big if—appropriate resources are made available. Many licensees discover that resources means additional license fees, consulting fees, hardware and infrastructure costs, and sometimes full time engineers to deal with the system that delivers improved search.

Let's step back and look at an important shift from keyword searching to concept searching. keyword

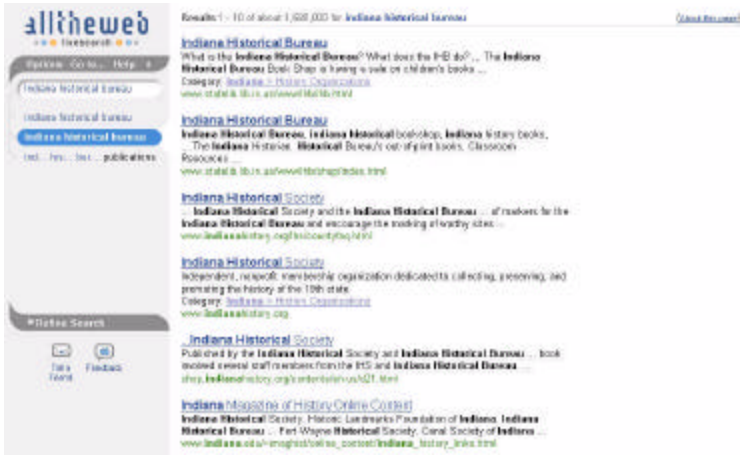
search, particularly in an organization, is not able to deliver what users expect, need, and demand.

Enter Rich Text Processing

Rich text processing is a relatively new category of search software. The idea is to move beyond keyword indexing. Search and retrieval since the 1960s has have the ability to take a user's query—for example, *Indiana Historical Bureau*—and find all occurrences of those words in the search system's index.

The problem is that a string match returns a laundry list of results including that string. Here's the query Indianapolis Museum of Art as retrieved by Yahoo's newest search system from Fast Search & Retrieval SA. You can try this yourself at http://livesearch.alltheweb.com.

Here's the results from Yahoo, arguably one of the world's most visited sites:



A laundry list of search results returned by Fast Search & Transfer's system as implemented on Yahoo's redesigned AllTheWeb.com service. The user must point-and-click through pages of results. No person can make sense of the more than 1.5 million items in this laundry list.

Contrast the Yahoo system with this approach implemented by Siderean for Oracle .

The system provides a bird's-eye view of the content or an "overlook" of the information. Oracle Corporation uses Siderean to allow users to see what's new, understand the content available, and perform a keyword search in a Web 2.0 interface that a user can rearrange to suit his / her needs.

The Yahoo system is what might be described as Web 1.0 or "old school". The Resource Connection approach is a Web 3.0 or "new school" approach to the challenge of locating information needed to answer a question.

The Resource Connection approach exploits rich text processing in three ways:

1. The system provides groups of information, not a long list of results. Research into

user behavior makes clear that most people look at a page, maybe two, of results. The RTP (rich text processing) approach allows the user to get a good sense of what's available at a glance

2. Each of the groupings—for example, "by Subject", "by Geographical Area", "by Time Period" and so on—make it possible to spot the specific category and the slice of data you want to examine. The idea is to reduce the amount of time required to scroll through a list. A moment's inattention increases the possibility the user will miss an important reference.

3. A search box is available so once the user has an "overlook" of the information available, the query can then be more sharply focused. What's clear is that this Resource Connection provides instructional materials from a wide range of sources, and the user knows this before wasting precious minutes trying to figure out what's in the system and what's excluded

4. Additional information about the number of items in a category is immediately available. Anyone trying to prepare a lesson plan knows that it is helpful to know that there are 899 resources available in mathematics without having to grind through irrelevant results to find the useful items.

In summary, rich text processing performs the type of value-adding most often associated with getting information from a person familiar with a particular subject. The idea is that instead of looking for strings of words, you can access information by categories, context, and relationships.

It's worth reemphasizing. RTP does not eliminate the need to perform a keyword search. RTP offers users more context, more information. The arguments for RTP go beyond convenience and have a financial benefit as well. The next-generation text processing systems can slash the time required to find an answer.

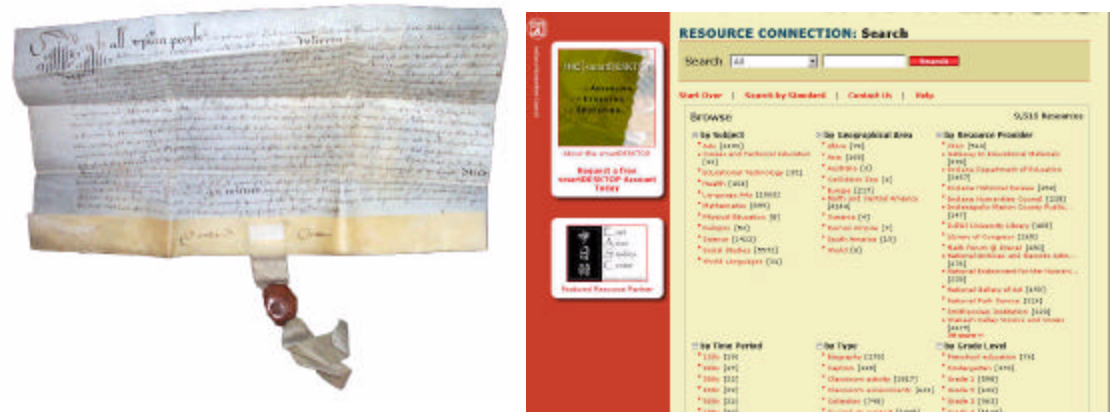
What's behind RTP?

When you dive into search, you are entering unknown waters. The marketing professionals invent buzzwords, hypnotize you with PowerPoint presentations, and scatter glittering generalities like pixie dust.

Let's go back in time. Visualize the scriptorum or library at Mont St. Michel off the coast of France. A small number of monks worked in the scriptorum, carefully handling scraped calf- or lambskin documents. The location of each item, its author, its title, and its date were entered in a log kept at the elbow of the person in charge of the collection. Some entries included additional information about the topics covered in the scroll, other items related to the scroll, and words or phrases that helped clarify a particularly broad title such as Ablavius' *Epistulae Ablabii praef. praet. et Constantini imperatoris de iure civitais Orcistenorum.*

These entries were precursors of RTP. The idea is very simple. A document—say, a report in your organization's marketing department—may not contain the name of the actual product. What's in the report labeled *Project Snowbird: New Product Analysis* is a treasure trove of customer research, competitive analysis, and financial projections. But two years after Project Snowbird has become your firm's biggest seller, the connection between the current product name and Project Snowbird may be

lost.



On the left is a manuscript typical of those kept in scriptorum until the invention of printing in Europe. On the right, is an interface based on content processed by the Siderean rich text processing system. The hyperlinks are more functional than the paper tab pasted to the source document. Both perform similar functions, but the point-and-click interface is a 21st-century, state-of-the-art faceted system. The paper tab is a medieval hyperlink.

We're not in France in the 9th century. There are no monks sitting down the hall, painstakingly creating keywords, phrases, and connections among the documents in your company's data centers.

Enter RTP.

The idea is to make software perform most, if not all, of the intellectual work performed by our hard-working monks. The RTP functions can run as a process within your existing search-and-retrieval system or be implemented as a value-adding process running as an external process. The implementation is less important than adding additional information about the document; for example:

- Linking the word *snowbird* to the concept *new product* and possibly adding the name of the product itself
- Identifying the type of document; for example, *market study*, *business plan*, and maybe *competitive analysis*
- Linking the author (hypothetically Janet Smith) and her unit, in this example *Marketing Department*
- Determining the date on which the document was created; for example, February 1, 2007 and the date on which the document was added to the digital collection, May 15, 2007
- Including a security identifier; for example, confidential, which means only employees with a specific level of access may know about the document's existence and view the document itself
- Document identification number, essentially a social security number for this particular report
- The categories or pigeon holes in the organization's taxonomy to which the document "belongs"; for example, R&D > Electronics > Consumer > Clock, for example.

Each RTP system can perform most, if not all, of these types of metadata or metatag indexing. *Metadata* means information about the information.

Metatagging involves associating a document with its metadata. Recall that a search engine indexes every word in a document. RTP adds additional indexing to a document. Instead of a tireless monk, the metatagging is handled by tireless software running on your computer. The best RTP software "understands" the content and concepts of a document and automatically indexes the document using these insights.

Why "Rip and Replace" Won't Work

You've justified the cost of your present enterprise search system. (For a refresher, take a look at the cost-benefit analysis in Annex 1 to this white paper.) In today's cash-constrained business environment, starting over from ground zero is neither practical nor desirable.

From a practical standpoint, your organization has a substantial investment in software, customization, and "on the job" training in the management of your search system. The task of reindexing the content in your organization, resetting security flags, and hooking your user interface into the search system is daunting.

From a desirable standpoint, a "rip and replace" has an impact on user behavior and work load for the specialists and engineers involved in the project.

Many organizations are looking for ways to enhance an existing search system. In this white paper, we provide cost analyses of two different ways to make search better. You can build out enhanced search using the tools available from your present search engine vendor. The alternative is to look at products and services from vendors who have more up-to-date specialist subsystems. The advantage of using a specialist subsystem is that your basic search engine remains unchanged, although you will want to update interfaces to reflect the new search and discovery options. The disadvantage is that you will have to manage software from two vendors. As you will learn, the cost of the specialist vendor's subsystem may save you significant amounts of money. In the first 12 months alone, you may realize a savings of up to 50 percent which can be a hundred thousand dollars or more. Over the life of the system, the cost savings may reach a million or more. You will also get the financial payoff from an enhanced search system. Users spend less time hunting for what they need, and some business deals will close because the needed information was available at the click of a mouse.

What you will find is that most vendors of enterprise search systems assert that their systems perform RTP, classification, and rich indexing as a *standard* feature. Our research indicates that the word standard is often used with little regard for its dictionary meaning.

Standard, according to Dictionary.com means usual, common, or customary. RTP is, by definition, part of the standard functions provided by a search engine. In reality, RTP is almost always subject to two important caveats:

1. A comprehensive tagging, entity extraction (identify the names of people, places, and things), and relationship discovery system is an extra-cost item. Not only must a licensee obtain additional software, additional professional services are required to implement the industrial-strength RTP functionality

2. The hardware, storage, and software used for keyword indexing is typically not able to handle the additional processes. This means an additional capital expense. Regardless of the amount of money involved, time and complexity make RTP a significant change in a mission-critical system.

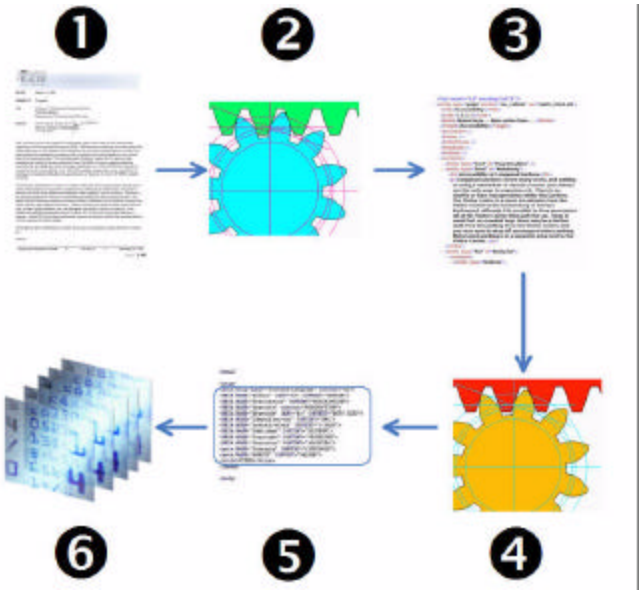
Let's look at some of the costs that many vendors of search systems with RTP "baked in" tuck into the nooks and crannies of their license agreements.

Hidden Costs of RTP

The table below provides you with a list of the costs that accompany RTP projects. Three of these warrant additional discussion because these three have a tendency to balloon upwards. In a few cases, the cost expansion has been so severe that the budget breaks.

Integration: Hooking In and Fine Tuning

Integration means making different bits and pieces into a cohesive whole. RTP systems are not usually a standard search engine function. Even though companies such as industry leaders like Autonomy, Endeca, Fast Search & Transfer, IBM, and Oracle, among others, assert that metatagging is embedded in their systems, make certain you know how each of these companies defines embedded. You will discover that RTP is an after-thought, sometimes licensed from a third-party provider or cobbled together from open source software. If the integration is successful, the generation of metatags to allow for entity extraction, concept tagging, and synonym expansion is automatic. Keep in mind that if RTP is buried within complex processes, you may not be able to fine tune the system to meet your needs. Each organization has its own vocabulary, jargon, and notions of how to best express its products features and benefits. A fully automatic system is like one-size fits all leisure suits from the 1970s. A clever marketing idea at one time, but obviously out-of-step for the 21st century.



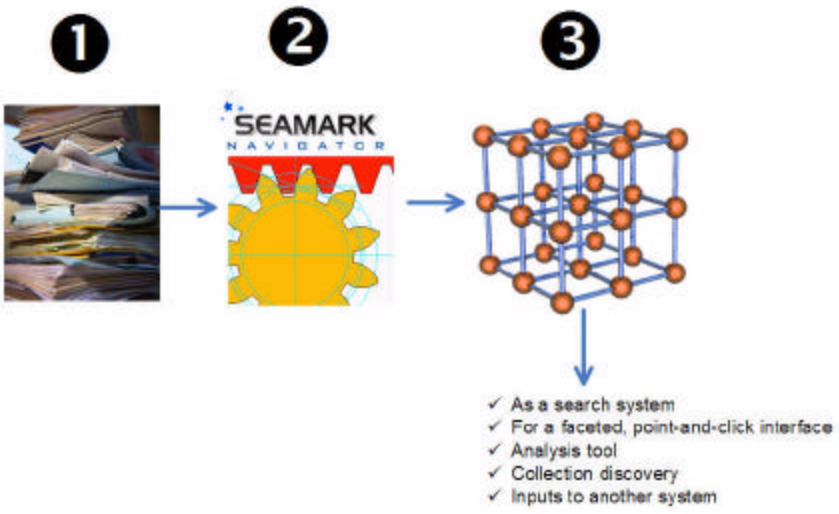
RTP involves several steps. [1] a source document is made available to the processing system, [2] the source document is converted (transformed) to XML, [3] an XML instance is written to the storage system, [4] the metatag processing system extracts from and discovers the metatags for that document, [5} the metatags are generated in XML, and [6] the search engine index adds these additional descriptors to the index and makes the tags available to the query processing subsystem and generates the faceted interfaces displayed to the user.

Smart Software

One of the hallmarks of the human mind is that it can effortlessly identify patterns. Granted some people like trained editors at a major publishing company may be better at this than others, but in general people can make connections as they receive data.

Software is not human in its capacity to manipulate data. Programmers, algorithms, and the knowledgebases are getting better, but more baby steps are needed before the Star Wars's computers become possible.

RTP is a series of processes that attempt to wrest additional meaning from text. Some of the processes use very traditional look up processes. A word list or knowledgebase is available to the system. When the processing subsystem identifies an acronym like EBITA, it looks in a list of synonyms (now called a knowledgebase) and finds out that EBITA means "Earnings before Interest and Taxes" and is generally used in finance. What happens is actually very simple. The RTP system generates a metatag that says, "This term means earnings before interest, taxes, and amortization" and a metatag that indicates the document in which the term appears contains information that can be classified as having a financial reference. Obviously a document with a large number of financial terms is a candidate for receiving a metatag that puts it in the Finance category of the taxonomy.



The Siderean system generates a graphical representation of the document's content and makes it possible to access a single document or a group of related documents by clicking on a cateogry or combining a traditional search with point-and-click exploration.

It's important to keep in mind that some RTP is not truly intelligent. The knowledgebases contain data the system uses when performing Rich Text Processing. Lower cost systems leave it up to you to create knowledgebases. More costly systems may include a knowledgebase tailored to your industry. There's a cottage industry producing word lists, dictionaries, thesauri, and databases stuffed with geographic and other factual information. Two key points are:

- You may have to invest considerable time and money in building and maintaining these knowledgebases
- An RTP that doesn't include a knowledgebase or a mechanism to use them may look like a great bargain only to turn into a less-than-useful system.

Other RTP subsystems are "intelligent". Many of these systems are using mathematical techniques that are several hundred years old. The reason is that only recently have computer systems become cheap and powerful enough to run these algorithms. Other systems use new algorithms in layers. This means that once one RTP process has been completed, another algorithm uses the data from the first process to identify other data. The process may be repeated dozens, even hundreds of times in order to produce

metatags that capture the names of people, places, and things in a document; the categories to which the document belongs; the most probable documents to which the document being processed is related; and so on.

One of the surprising aspects of RTP is that many systems are marketed as being radically different in their approach, power, and "intelligence". A quick look under the hood of these systems reveals that the "differences" are very minor. In fact, when one looks closely at the technology and the engineers involved in these systems, you will discover as we did that smaller companies licensed their algorithms to these industry leaders. EPI Thunderstone, Inxight (now part of Business Objects), Outside In (formerly Stellent and now Oracle) enable the brand name RTP systems. More surprising still is that a relative handful of experts in text and linguistics from a handful of top flight universities provide most of the cutting-edge science in these systems. The centuries old proverb *Caveat emptor* applies directly to RTP. *Buyer, beware*. What you see may be quite different upon closer examination.

These "intelligent systems" use many different techniques to perform RTP. Two approaches perform much of the "magic" in RTP and warrant brief comment.

The first is pattern recognition. The RTP examines each document, makes note of certain patterns, and then uses the data about the patterns to make a judgments about meaning. For example, if a series of documents contains numerous references to Inconel, a high-grade steel, power generation, and contains names of 100 companies with most documents referencing 10 of these firms, the system has enough clues to generate metatags that pinpoints these documents as belonging to the category *Power Generation* and to associate the most-mentioned companies as being related to this category and the product name *Inconel*. The idea, therefore, is to count and rank.

The second approach is to look at the way the document, paragraphs, and sentences are formed. Without getting bogged down in such jargon as latent semantic indexing or linguistic analysis, let's simplify this process. The RTP system chops a document into parts when it generates an XML instance. If we zoom into a sentence, the RTP will identify the subject of the sentence and its verb. If words like White House and stock market are used, the RTP will understand that these two words belong together. The system will use algorithms to read the sentence and identify these phrases and relate them to concepts like president or the name of an office.

As you might imagine, algorithms make errors, so you will not find 100 percent accuracy in these language-centric processes. But for most purposes, a system that gets 75 to 80 percent of the sentence's meaning correct, the metatags are extremely useful. A document containing the terms Bush, White House, and Washington, DC, will almost always be properly tagged and categorized.

In general, when the documents are technical or contain what might be thought of as research-oriented information, the RTP systems do their best work. When the source documents are jargon-filled and highly colloquial, the RTP systems stumble.

What you have learned is that RTP involves one of these two approaches. Some companies include both approaches in the RTP systems. You have also learned that neither approach is perfect. The value of the systems, to some extent, depends on the source documents themselves. Within the last two or three years, more powerful computer chips at extremely competitive prices have made it possible to offer intelligent software to individuals and organizations. Rapid progress is being made in this area of RTP, but vendors are prone to exaggeration.

Complexity

The final point can be difficult to see. It's woven into the fabric of modern search-and-retrieval systems. These systems are complex, and this means that a system may be only as good as the vendor's commitment to giving the licensee a software system that delivers what the client wants.

Advanced text processing is complex. Among the factors creating the challenges most organizations face each day are:

1. Amount and type of content to be processed. Let's face it, Google and Yahoo can choose what their systems can index. Organizations can't. Until a system is up and running, most organizations have no concrete data about how much processing their documents require. For example, brief email and Word files are rarely a problem. But when the source documents are mutli-megabyte Portable Document Format files, large PowerPoints, and terabytes of content in databases—processing time and costs become evident.

2. Taxonomy and knowledgebase tasks. Once the rich text processing system is installed, you will be able to examine initial outputs and determine if you need to fine-tune the system, the taxonomy, or the knowledgebase the system uses. Training collections are very useful in getting the basic system up and running. But many issues arise only when a large volume of content is available for analysis.

3. Customization. Most modern systems work "out of the box" and deliver useful results. However, once the system is in operation, you will find that you need to "tweak" or modify the system's operation to reflect what your organization needs to derive maximum value from its content.

The puzzle is, "How do you control your rich text processing costs *before* you operate the system?" Let's look at what ArnoldIT.com learned about one next-generation rich text processing system.

Siderean: Navigating Content and Costs

In the course of our work for the first three editions of the Enterprise Search Report, we interviewed dozens of licensees of the major enterprise search systems and many rich text processing systems. In addition, we've participated in many off-the-record conversations in the hallways at conferences and even talked with developers who make their living customizing search and text processing systems.

What we've learned is that there are several cost "hot spots" in a rich text processing implementation. Furthermore, we've gathered basic data about the amount of time required to set up, develop knowledgebases, and customize systems. We've aggregated our data and developed assumptions about how much a "typical" organization is likely to spend to perform the following tasks:

- Refine an existing taxonomy or knowledgebase
- Set up and tune a rich text processing system
- Develop "connectors" to deliver the content either to another enterprise application such as a keyword search system or an analytics program able to develop reports about the people and other facts tagged by the rich text processing system
- Maintain the system.

As you know, assumptions can vary widely from organization to organization. We've reproduced ours in the Annex "Rich Text Processing Cost Assumptions" that accompanies this white paper.

We've also used data gathered in the course of analyzing the rich text processing system developed by Siderean, Inc. The Siderean approach is directly congruent with major industry trends in extracting meaning from text and making it possible to examine content, find needed answers, and explore topics

using “hooks” and “handles” generated by the Siderean system.

The Major Players’ Costs

A licensee can obtain from Autonomy, Endeca, Fast Search & Transfer rich text processing functionality. Each of these companies offers a variety of tools, add-ins, and components to:

- Extract entities
- Generate metatags about a document, including author, date and time of creation, and file size
- Determine and assign a document to one or more categories, index a document with its concepts, and make these elements available to the user in a range of interfaces, including visualizations of various types.

Many organizations believe that obtaining technology from a single source eliminates many variables in implementing mission-critical systems. However, the use of a single vendor’s software creates two well-known challenges.

First, once the system has been installed, it becomes difficult to shift to another system. “Lock in” limits the licensee’s flexibility and ability to implement functions not supported by the vendor’s system.

Second, the single-source makes it difficult to obtain price competitive bids for certain work. For example, most mainstream vendors of search systems use proprietary technology. Even a licensee with exceptional technical skills may not be able to implement certain changes. The alternative, as many organizations discover, is that the vendor’s professional services unit performs the work at a negotiated price. Without a competitive bid, the fees for these services can often exceed the original license fee itself. Some vendors have added former consultants armed with several years’ experience at a major consulting firm like Boston Consulting Group and an MBA from a prestigious institution to justify the fee structure.

Siderean: Semantics and Cost Controls

The Siderean approach stands apart from the approach taken by other vendors of search and companies such as IBM, Microsoft, and Oracle who “bundle” many services with other products.

First, Siderean’s technology has been engineered to make use of the information a customer has available. For example, if there is a search system installed, Siderean’s technology operates as a component of that system, integrating with the keyword search system to give users more ways to access the processed content. The search box remains the same. What changes is the inclusion of suggested content, point-and-click categories that expose additional related information, and automated mechanisms to highlight the most recent content on a particular topic.

Second, Siderean’s approach allows the licensee to install the system in the organization’s data center, access the Siderean functionality in a secure, hosted environment, or combine the two approaches to deliver the best combination of cost and performance.

Third, Siderean brings to bear a semantic approach. What the company has developed is a system that “understands” the meaning of content the system processes. Siderean’s system can “learn” as it processes documents, and it can be set up to generate categories and other “hooks” and “handles” automatically.

Siderean has achieved what we have termed *balance*. The system can be deployed in as little as three working days. Alternatively, the system can be customized to perform very specialized functions required for military intelligence and legal discovery applications.

Our tests and data reveal that Siderean delivers rich text processing at a cost that is consistently half that of most search systems. When compared to the costs of enterprise platforms, Siderean's cost savings are even greater.

Our tests reveal that the functionality of the Siderean system requires no compromise in interface design. We did discover, however, that Siderean's option for a hosted installation or a local installation, like other systems, may require additional processing capacity under two conditions: [a] document sizes grow larger over time and [b] the number of documents processed for updates increases. These are normal growth issues outside the control of the rich text processing system, and the model assumes that the baseline processing will be two million initial documents with 10,000 new or changed documents per week. Your situation may vary from this baseline, and the cost data will require adjustment.

The key point is that the Siderean system is a next-generation rich text processing system. It features fast configuration, ability to use existing knowledgebases or "discover" the terms and categories, and a "snap in" design that adds semantic functions to almost any keyword indexing system.

The Cost Analysis

The analysis per the assumptions in the Annex reveals that the Siderean system delivers rich text processing at an estimated cost of about $240,000 versus the estimated cost of the enterprise search implementation of more than $500,000. In this analysis, this represents a cost savings over start up and first year operation of more than 50 percent. That's a more significant savings than is carried out through the life of the project with no compromise in functionality.

The major savings are in the type and amount of professional services required to implement rich text processing with the enterprise search system vendors' tools. Enterprise search systems have added features and functionality over time, often by using open source solutions, acquiring technology via buy outs, or licensing technology from tool vendors. The result is a "box of parts," not an integrated system. The engineering and business planning work adds to the integration, deployment, and tuning costs.

As a result, organizations with a search system from a high-profile vendor like Autonomy, Endeca, Fast Search & Transfer, Google, IBM, Microsoft, or Oracle may end up paying a premium for the needed functionality. Because of the complexity of these enterprise search systems, the ongoing costs continue to run higher than costs for next-generation systems. Finally, the complexity of these "box of parts" solutions is higher, which means that troubleshooting takes longer. What many licensees reported to us in the course of our research for editions one, two, and three of the Enterprise Search Report was:

- Costs became difficult to control due to the number of unknowns about dependencies in the system
- Some problems could only be resolved by waiting for a specific engineer who was often not available or priced at a premium by the enterprise search vendor
- Certain issues could not be "fixed". In effect, the original developers were no longer available, and the enterpriser search vendor's engineers did not know how to address certain problems in specific deployments of the subsystem.

Business Consulting or Engineering Consulting?

One other key cost factor is the confusion among more than half of those interviewed in the course of our research for the Enterprise Search Report. Many enterprise search firms have turned to general business consulting to boost their revenues.

Most of the vendors of enterprise search have been unable to expand their revenues by licensing their products to new customers. The consequence is that some enterprise search companies derived more than half their revenues from non-search activities or from selling consulting services to existing customers and new licensees.

As a result, the typical enterprise search installation in a company with more than $300 million in revenues can easily hit $1 million in first year set up and operation. Many licensees assumed that the MBAs were essential to the search process. Others groused because scarce resources were allocated for work flow studies, starving the project manager for technical and programming resources essential to get the system up and running smoothly.

In short, with enterprise search vendors pressured with low cost systems such as those offered by from Tesuji.com (a vendor of open source search), and squeezed by "free" systems from IBM-Yahoo, the licensee of a mainstream enterprise search system may find itself paying more to get the features required.

The mumbo-jumbo of MBAs speaks clearly to some. But to those responsible for getting a complex technical system deployed, the engineers who deliver code and solutions without PowerPoint are what's needed.

Siderean is a customer-centric engineering firm. Its systems can be deployed in days, not months. When technical support is required, Siderean includes services that other firms classify as "consulting services."

Cost Variations

The model in the Annex to this white paper presents an analysis based on aggregate data. Your situation is likely to be different. Using the data in the Annex as a baseline, the table below provides you with a guideline for determining the costs of your system using the Siderean technology and your existing enterprise search system. Note that if you do not have an enterprise search system, the Siderean rich text processing system includes a keyword search system as a standard component.

Table 1: Baseline Comparisons

	Enterprise Search	Siderean
Taxonomy	$165,000	$8,000
Collection Analysis	$8,000	$5,600
Transformation	$59,600	$12,400
Team	$2,160	$2,160
Other / Contingency	$141,190	$73,040

Table 1: Baseline Comparisons

	Enterprise Search	Siderean
First year operations	$82,200	$63,800
Platform upgrade	$86,550	$81,950
Total	**$544,700**	**$246,950**

If we assume unexpected changes (additional hardware, storage, or troubleshooting), the total first year costs could rise 15 to 100 percent. In this situation, the table below provides a way to compare potential cost exposure at three levels of risk:

First, let's look at the costs of errors for the enterprise search implementation at 25, 50, and 100 percent:

Table 2: Cost Exposure Analysis

	Enterprise Search Cost Risk	Next Generation Cost Risk
Baseline	$544,700	$246,950
25% exception	$680,875	$308,688
50% exception	$817,050	$370,425
100% exception	$1,089,400	$493,900

What's clear is that if additional customization or document connectors are required, the cost exposure from the next-generation is less than the cost exposure from the enterprise search system offering rich text processing as an add on.

Second, when one recalls that the Siderean system includes a keyword search function, it is possible that some licensees can reduce their dependence on their existing enterprise search system.

Finally, with the added productivity from the rich text processing interface to enterprise content, licensees are likely to experience a shorter time to reach break even and enjoy a larger return on their investment in rich text processing.

The cost exposure is significant. To sum up, you get reduced risk, lower total cost of operation, and the rich text functionality that can deliver additional return-on-investment payoffs.

ArnoldIT Observations

Our analysis of rich text processing systems reveals that organizations are increasingly concerned about keyword search systems. Laundry lists of results are not useful in most enterprise contexts. Furthermore, employees expect a keyword system to work "like Google", and the content in organizations does not lend itself to a voting or popularity algorithm.

Consequently, there is an increasing demand for systems that can process content and make information available via different types of interfaces, with different tools to aid the employee in exploring content, and with the option for exposing information in personalized presentation formats.

Siderean's rich text processing delivers enhanced indexing, categorization, and other rich text

processing features. Furthermore, it delivers these important features at a lower initial and operating cost than systems available from better-known vendors of search and enterprise software.

Most importantly, if an organization discovers that its initial assumptions about its content are incorrect, the Siderean system poses lower risk of a crippling cost overrun at no loss of functionality.

Other key points we noted are:

- Siderean has implemented a semantic system that reduces the amount of MBA and manual processing associated with "box of parts" rich text processing systems
- The Siderean system can be operated as a managed service or as an installation on the licensee's servers
- Siderean's metatagging supports a wide range of interface options ranging from a point-and-click Yahoo-style interface to a Web 2.0 drag-and-drop approach
- Siderean includes a keyword system so some licensees may be able to dispense with a bulky, expensive enterprise search system
- Siderean's processing time is comparable to or faster than the other systems ArnoldIT tested, with speeds of 6000 documents per minute observed on our test corpus.

As with any complex technology, our analysts identified four important caveats:

- Content within organizations is extremely diverse and complex. Each licensee, therefore, may encounter specific file formats or specific content issues that are unique to that organization; for example, the need to delete classified material from the index at a particular date and time. Some specialized security functions may have to be addressed by specialists trained in regulatory compliance, security requirements, and similar exogenous factors.
- Users move "up the learning curve" quickly when new features and functions become available. It is important to manage expectations and set specific checkpoints to add features to a system. This is particularly true when functionality jumps up several orders of magnitude with a rich text processing system.
- Content growth in organizations can range from 1.5 to 2.0 growth per year. This means that the baseline servers and storage devices will require upgrades to handle this type of content growth.
- Organizations change, sometimes rapidly. Large-scale systems are difficult to turn on a dime. If you must make a significant change due to a reorganization or a merger, you will want to document your system and ensure that you have people experienced with the system on staff. Your team does not have to code the rich text processing subroutines, but you want to have appropriate knowledge of what the system is doing and how it operates.

We recommend that organizations looking for a rich text processing system give Siderean a long, hard look and a test drive.

Annex 1: Table of Costs for Traditional Rich Text Processing: Aggregate of Systems such as Autonomy, Endeca, Fast Search & Transfer, IBM iPhrase, Microsoft MOSS, and Oracle SES 10g

This analysis shows projected start up costs and first year operational costs for a rich text processing system implemented with add-ons available from mainstrean vendors of enterprise search. Note that these are composite numbers based on assumptions derived from research conducted for editions one, two, and three of the Enterprise Search Report. The content volume is an initial two million objects with 50% growth over the first year. The hardware included in this analysis is known to have the processing capacity required for this flow and weekly updates. Your specific situation may affect fees for services and infrastructure. Use these assumptions as a starting point for your business case.

(NOTE: Consultants cost more than in-house employees on an hourly basis.) If you make other assumptions about time and cost, your analysis will vary from this model's.

Table 3: Traditional RTP Set Up Cost Assumptions

	Assumptions	Values	Units	Quantity	Skill Category	Cost per Unit	Est. Cost	Notes
Set Up Costs								
Taxonomy: Develop Taxonomy/Training Set	Taxonomy Specialist	500	Hours	2	specialists	$100	$100,000	Contractor provides two part-time people for three months
	Taxonomy Tool License Fee	-	Fixed Cost	1	n/a (product)	$65,000	$65,000	Use Data Harmony or equivalent
Collection Analysis	Content Audit							
	Analyze search collections	40	Hours	2	specialists	$40	$3,200	Review database content not in unstructured content index
	Identify additional content	40	Hours	1	analyst	$75	$3,000	Identify content appropriate for inclusion

Table 3: Traditional RTP Set Up Cost Assumptions

	Assumptions	Values	Units	Quantity	Skill Category	Cost per Unit	Est. Cost	Notes
	Report	24	Hours	1	analyst	$75	$1,800	Develop security and transformation recommendation for additional content
Transformation[1]	Metacontent server	-	Fixed Cost	1	n/a (product)	$5,400	$5,400	Assume outright purchase; PowerEdge 1950 plus storage device
	Custom functions	40	Hours	2	n/a (product)	$75	$6,000	Assume perl or python scripts plus CSS editing
	Tuning	16	Hours	2	n/a (product)	$75	$2,400	Assume minimal tuning due to experience with core search engine
	Work flow analysis	80	Hours	2	analysts	$200	$32,000	MBA level work to "hook" tags into Web screens for employees; assume two custom processes and one default process
	Work flow scripts	40	Hours	2	programmers	$75	$6,000	Scripts hook queries and displays to particular work flows
	Debugging	24	Hours	1	programmer	$75	$1,800	Modifications needed to make system stable
	System training	40	Hours	2	programmers	$75	$6,000	Running selected documents identified in "analyze search collections" through system to update taxonomy and other knowledgebases

Table 3: Traditional RTP Set Up Cost Assumptions

	Assumptions	Values	Units	Quantity	Skill Category	Cost per Unit	Est. Cost	Notes
Professional Team (Employee)	IT Manager	8	Hours	1	manager	$75	$600	Licensee's internal staff
	Project Manager	12	Hours	1	manager	$50	$600	Licensee's internal staff
	System administrator	8	Hours	1	programmer	$40	$320	Licensee's internal staff
	Budget analyst	8	Hours	1	analyst	$40	$320	Licensee's internal staff
	Contract specialist	8	Hours	1	analyst	$40	$320	
Sub-total Start-up (Scenario A)							$234,760	$2,160
Other Additional	Contingency	$234,760	Subtotal		n/a	25%	$58,690	Cash available for modifications is 25% of total
	License fee	$330,000	Fixed Cost	1	n/a (product)	25%	$82,500	Assume 25% of annual license fee to RTP module
	Total Start-up (Scenario A)						$375,950	

1. Existing document processing does not include new metatags

Annex 2: Table of Operating Costs for Traditional Rich Text Processing: Aggregate of Systems such as Autonomy, Endeca, Fast Search & Transfer, IBM iPhrase, Microsoft MOSS, and Oracle SES 10g

This is a continuation of Annex 1.

Table 4: Traditional RTP First-Year Operating Cost Assumptions

	Assumptions	Description		Quantity	Skill Category	Cost per Unit	Est. Cost	Notes
Operational Costs								
Taxonomy Modifications in Production	Taxonomy Specialist	192	Hours	1	specialist	$75	$14,400	Assume two days per month
	License word lists	-	Fixed Cost	1	n/a (product)	$4,000	$4,000	
	Collection Analysis							
	Analyze new content	192	Hours	1	specialist	$75	$14,400	Assume two days per month
	Transformation							
	Connector	16	Hours	1	programmer	$75	$1,200	Contractor creates / tweaks connectors
Professional Team (Employee)	IT Manager	96	Hours	1	manager	$75	$7,200	One day per month
	Project Manager	192	Hours	1	manager	$50	$9,600	Two days per month
	System administrator	769	Hours	1	programmer	$40	$30,760	8 days per month; 64 hrs / month
	Budget analyst	8	Hours	1	analyst	$40	$320	1 day per year

Table 4: Traditional RTP First-Year Operating Cost Assumptions

	Assumptions	Description		Quantity	Skill Category	Cost per Unit	Est. Cost	Notes
	Contract specialist	8	Hours	1	analyst	$40	$320	1 day per year
	Sub-total Operations (Scenario A)						$82,200	
	Contingency	$82,200	Subtotal		n/a	25%	$20,550	Cash available for modifications is 25% of total
	Platform Upgrade	$330,000	Fixed Cost	1	n/a (product)	20%	$66,000	Assume 20% of annual license fee to RTP module
Total Operations (Scenario A)							$168,750	
Grand Total	Total Start Up and First Year Operations (Scenario A)						$544,700	

Annex 3: Table of Costs for Next-Generation Rich Text Processing: Aggregate of Systems such as Autonomy, Endeca, Fast Search & Transfer, IBM iPhrase, Microsoft MOSS, and Oracle SES 10g

Newer systems allows certain cost reductions in implementing and managing rich text processing. The cost analysis below identifies these savings using highlighting. It is possible to over-spend for a next-generation rich text processing system by engaging in extensive customization of the vendors' systems. This model assumes that the customization is focused on integrating the metatagged objects into interfaces without relying on manually-written scripts to perform stored searches. (NOTE: Consultants cost more than in-house employees on an hourly basis.)

Table 5: Next-Generation RTP Set Up Cost Assumptions

Set Up Costs	Assumptions	Description	Unit	Quantity	Skill Category	Cost per Unit	Est. Cost	Notes
Taxonomy: Develop Taxonomy/Training Set	Taxonomy Specialist	80	Hours	1	specialists	$100	$8,000	Next-generation systems use embedded knowledge-bases and heuristics to "learn" new content; thus reducing the time an expert must spend to create a tax-onomy
	Taxonomy Tool License Fee	-	Fixed Cost	1	n/a (product)	$0	$0	Use Data Harmony or equiv-alent
Collection Analysis	Content Audit /							
	Analyze search collec-tions	20	Hours	1	specialists	$40	$800	Automated routines gener-ate candidates, allowing more efficient content review
	Identify additional content	40	Hours	1	analyst	$75	$3,000	Identify content appropri-ate for inclusion

Table 5: Next-Generation RTP Set Up Cost Assumptions

Set Up Costs	Assumptions	Description	Unit	Quantity	Skill Category	Cost per Unit	Est. Cost	Notes
	Report	24	Hours	1	analyst	$75	$1,800	Develop security and transformation recommendation for additional content
Transformation[1]	Metacontent server	-	Fixed Cost	1	n/a (product)	$5,400	$5,400	Assume outright purchase; PowerEdge 1950 plus storage device
	Custom functions	8	Hours	1	n/a (product)	$75	$600	Assume perl or python scrips plus CSS editing
	Tuning	8	Hours	1	n/a (product)	$75	$600	Assume minimal tuning due to experience with core search engine
	Work flow analysis	8	Hours	1	analysts	$200	$1,600	MBA level work to "hook" tags into Web screens for employees; assume two custom processes and one default process
	Work flow scripts	8	Hours	1	programmers	$75	$600	Scripts hook queries and displays to particular work flows
	Debugging	16	Hours	2	programmer	$75	$2,400	Modifications needed to make system stable
	System training	16	Hours	1	programmers	$75	$1,200	Running selected documents identified in "analyze search collections" through system to update taxonomy and other knowledgebases
Professional Team (Employee)	IT Manager	8	Hours	1	manager	$75	$600	Licensee's internal staff

Table 5: Next-Generation RTP Set Up Cost Assumptions

Set Up Costs	Assumptions	Description	Unit	Quantity	Skill Category	Cost per Unit	Est. Cost	Notes
	Project Manager	12	Hours	1	manager	$50	$600	Licensee's internal staff
	System administrator	8	Hours	1	programmer	$40	$320	Licensee's internal staff
	Budget analyst	8	Hours	1	analyst	$40	$320	Licensee's internal staff
	Contract specialist	8	Hours	1	analyst	$40	$320	Licensee's internal staff
	Sub-total Set-up (Scenario B)						$28,160	
	Header Needed Here							
	Contingency	$28,160	Sub-total		n/a	25%	$7,040	Cash available for modifications is 25% of total
	License fee	$330,000	Fixed Cost		n/a (Product)	20%	$66,000	Assume 20% of annual license fee to RTP module
Total Set-up (Scenario B)							$101,200	

1. Existing document processing does not include new metatags

Annex 4: Table of Operating Costs for Traditional Rich Text Processing: Aggregate of Systems such as Autonomy, Endeca, Fast Search & Transfer, IBM iPhrase, Microsoft MOSS, and Oracle SES 10g

This is a continuation of Annex 3.

Table 6: Next-Generation RTP First-Year Operating Cost Assumptions

	Assumptions	Description		Quantity	Skill Category	Cost per Unit	Est. Cost	Notes
Operational Costs	Taxonomy Modifications in Production							
	Taxonomy Specialist	96	Hours	1	specialists	$75	$7,200	Assume one day per month
	License word lists	-	Fixed Cost	0	n/a (Product)	$0	$0	
Collection Analysis	Analyze new content	96	Hours	1	specialist	$75	$7,200	Assume one day per month
Transformation	Connector	16	Hours	1	programmer	$75	$1,200	Contractor creates / tweaks connectors. This is gated by changes in systems elsewhere in the organization
Professional Team (Employee)	IT Manager	96	Hours	1	manager	$75	$7,200	One day per month
	Project Manager	192	Hours	1	manager	$50	$9,600	Two days per month
	System administrator	769	Hours	1	programmer	$40	$30,760	8 days per month; 64 hrs / month
	Budget analyst	8	Hours	1	analyst	$40	$320	1 day per year

Table 6: Next-Generation RTP First-Year Operating Cost Assumptions

	Assumptions	Description		Quantity	Skill Category	Cost per Unit	Est. Cost	Notes
	Contract specialist	8	Hours	1	analyst	$40	$320	1 day per year
	Sub-total Operations (Scenario B)						$63,800	
	Header Needed Here							
	Contingency	$63,800	Sub-total		n/a	25%	$15,950	
	Platform Upgrade	$330,000	Fixed Cost		n/a (Product)	20%	$66,000	
	Grand Total Operations (Scenario B)						$145,750	
Grand Total	Total Start Up and First Year Operations (Scenario B)						$246,950	