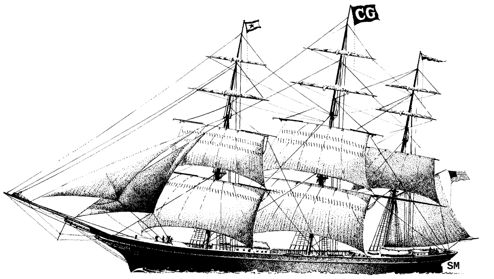
THE CLIPPER GROUP
Explorer™



Navigating Information Technology Horizons℠

Published Since 1996 Report #TCG2005048 July 28, 2005

Enterprise Search Adds a User Dimension to Business Information Organization

Analyst: Anne MacFarland

Management Summary

Information has always been organized to make it more usable, in the form of journals, account books, and file cabinets. Business has now moved to the cyber-equivalents of databases (and their junior version, spreadsheets) and files. These may be satisfactory to the applications that generated the data, and the systems administrators that manage the environment, though the volume of data is getting troublesome. They do not always satisfy either the users who want information rather than data or the applications that seek to leverage information they did not create. There are processes that can integrate data from different sources, and map contexts from one data source to another – but first the data must be found.

Application-based search and indexing capabilities can suffice, but suffer from the expectations for which they were created. While federated application-level search may work where only a few applications are involved, the coordination and integration of results into usable coherence can be troublesome as the number of information sources scales. **Enterprise Search is a class of applications that have been designed to deal with multiple file and database formats and multiple data sources, with the architecture to quickly seek, intelligently analyze, and promptly present the information in a useful form and timely manner, tuned to the needs of the user.** This alone is worth exploring, but there is more.

Web search engines, as separate applications or as part of a Web page or portal, are so familiar to us all that we may not realize that enterprise search can do far more for the enterprise than cater to the sudden need to know or find something**. In an enterprise of vast but finite documentation and describable, persistent needs for certain kinds of information, a search engine can crawl during idle times, winnow out new information, order and de-duplicate it, and deliver it to users in a push, as desired. Search can monitor business process or workflow application data and alert managers when target events, or deviations from norms, occur**. Search optimizes content management and is an inherent part of database operations and many workflow applications. Used opportunistically on an enterprise scale, search can identify similar projects across the enterprise and aid in enterprise coordination of efforts. Enterprise search can do some tasks on an episodic basis, and use compute-intensive functions, like full-text analytics, only when and where needed. In doing so much, enterprise search can become a tool for managing both information and the business itself.

This enterprise use of search represents a different way of getting at data than the familiar, static organizations of databases and file systems. While databases, particularly relational databases, give a dimensional view of data, the real-time push, pull and filter capabilities of search can tailor the data more easily and exactly to the relevance of the moment of each user. Because relevance is in the eye of the beholder, it is important to take a look at your enterprise's use of information, and at its functions and culture, to determine how search can effectively benefit your enterprise. Read on for more details.

IN THIS ISSUE

Search and the Enterprise

The value of information comes from its use. While larger data sets generate better statistics, and more sources create more confidence in the result, terabyte count alone does not make information useful. Increasingly, enterprises are looking to leverage the value of their information by using it in multiple contexts across the organization, and, where appropriate, sharing it with partners**. Search enables this kind of use more directly than any other form of information organization**. It also can be used as a tool to enhance traditional forms of information organization (see Exhibit 1, below) – or remedy their disorganization.

Most search engines have a similar basic structure (as shown in Exhibit 1). They differ in their scope, in how well they scale, in how efficient they are, and, most significantly, in all the amenities and special features that they offer. Many of the differentiators come from special characteristics featured in the areas of indexing, classification, contextualizing, conceptualizing, and enriching; and in the filtering, and ranking of results to meet the need of the individual user.

Scaling

Replicable query, crawl, and indexing components, deployed across scale-out architectures, allow a search engine to scale as needed to handle more concurrent queries, a greater volume and variety or data sources, and more indexing, analysis and filtering elements. With such componentization, query serving and indexing become two separately-synchronized, asynchronous processes like the vending machine stockers and the vending machine buyers.

Process Latency

The use of modular, scale-out architectures can also reduced the latency of the response to a query by allowing the search against the index to be segmented and parallelized. There is a more

Exhibit 1 –
Different Dimensions of Data Organization

	Structured Data	**Unstructured Data**	**Graphics and Representational Data**	**Search as a Dimension of Data Organization**
Example	Data Base Application Content Legacy Data Structured Documents	Files Documents E-mail Stock feeds	Medical Records CAD Exhibits Data visualization News media clips	Results (lists, graphs or alerts)
Tools	ETL Data Mining EAI Other Analytics, Business Intelligence products	The classification and taxonomy-generation of Knowledge Management and Content Management applications	Application-specific capabilities Some industry-specific tools are available, for instance in health care, where there are standardized imaging formats	All of those at left, but generally not with the complexity of dedicated applications Digital Assets Management (DAM) Digital Rights Management (DRM)
Search as a Tool	Queries Navigation Analytics Exploration	Metadata search Text Mining Entity Extraction	Search for patterns or anomalies Metadata Searches Navigation	
Point of Organization	Done as part of capture or import from other data source	Can be done at capture, import or asynchronously, as needed	Generally done at capture/creation	*Opportunistic: Done when needed*

persistent latency issue, however, in the process of data indexing and analysis. Much of the need for information, these days, is for the latest information, in as near real-time as possible. Most news feeds, for instance, lose value if they are not fresh. Transaction data is most valuable if it is so immediately available that it can provide a reality check between a decision and its implementation, for instance in stock trades and other opportunistic environments. Processing information to be available to a search engine takes time.[1]

More latency can be added to the process when the results are filtered for a particular user. For instance, if a user is searching marketing campaigns involving rebates, and wants results ranked by the bump each gave to relevant product revenue over the course of the campaign, normalized to a per week amount, it may take a little time.

Authorization and Security

If enterprise search engines are working with other than publicly-available (or organization-wide) information, they must respect local authorization limitations. Users should understand that search does not transcend restrictions on information access – though the enterprise may act to make some information more widely accessible. The information that is collected about searches should also be kept within the enterprise and under IT administrative control.

For distributed enterprises, the desire to leverage a large body of enterprise content, one that spans corporate firewalls, can cause concerns. External use of information, for instance in ebusiness websites, is usually segregated outside the corporate firewall. However, often, internal users want to include external information sources within their search, a process that would involve crawling across the firewall. Federated search addresses this problem. However, federated search with diverse search engines can lead to inferior, least-common-denominator results. In general, the ideal search engine for an enterprise can work anywhere, and that can support any characteristic that the enterprise requires. With a common structure and metadata model, search can then be aggregated across security domains, and masked to provide the appropriate user experience.

Shopping for Search

All search capabilities can seem very attractive in the abstract. If you start by surveying survey search products, you can be quickly overwhelmed. If you consider only your immediate pain points, you may buy a point product that cannot expand to meet broader enterprise needs. Supporting a complex search ecosystem within an enterprise should not be undertaken lightly. So the first thing you should consider, when looking at enterprise search, is your own enterprise – *what you will be using search for*, what *volume and kind of queries you envision*, and *what data sources you wish to search*. Finally, you should think about the presentation modes that your users will require, for there are many ways that results can be presented, besides a too-long list.

Your Own Enterprise

Consider your domain, be it an enterprise, ecosystem of organizations, or industry.

- ***How is your enterprise structured? How firmly is it bounded****?* Leaving aside the passion-arousing issues of security (touched on above) and culture, is information usefully shared across business units? Do trends within your organization and within your industry lead you to think this will change?
- ***Is your institution rule-delimited or consensus-driven?*** This will influence your preferences in the area of indexing and analysis. Information-centric industries, like insurance and banking, often have an appreciation of the taxonomic approach to information, while process- or service-industries may see taxonomy development as a bewildering chore, and may want to go the entity extraction route.
- ***What is the collaborative profile of your organization?*** Is collaborative innovation formally supported, or is collaboration more of an accessory to the reporting structure of meetings, conferences, and reports, and the work of business accomplished by "single-threaded" independent initiatives? Is change (in process or in product) done out of necessity, or is rapid evolution and adaptation part of a pervasive "keep fresh" game plan? This varies by industry, but also can be affected by the attitudes and strategies of senior management. A lot of the need for broad search is born from the need for effective collaboration and innovation. If your organization is a "meet the moment" kind of player, the ad hoc organization that enterprise search permits may be particu-

[1] This is like going to a library to borrow a best seller, only to find it not yet on the shelves because it had not been entered into their business systems.

larly valuable – and an extensible, customizable approach to search may be very important.

Search can be an important part of transforming your organization to a more integrated whole, but it is a tool, not a cure-all. What you look for, particularly in indexing and classification, depends on how your enterprise works.

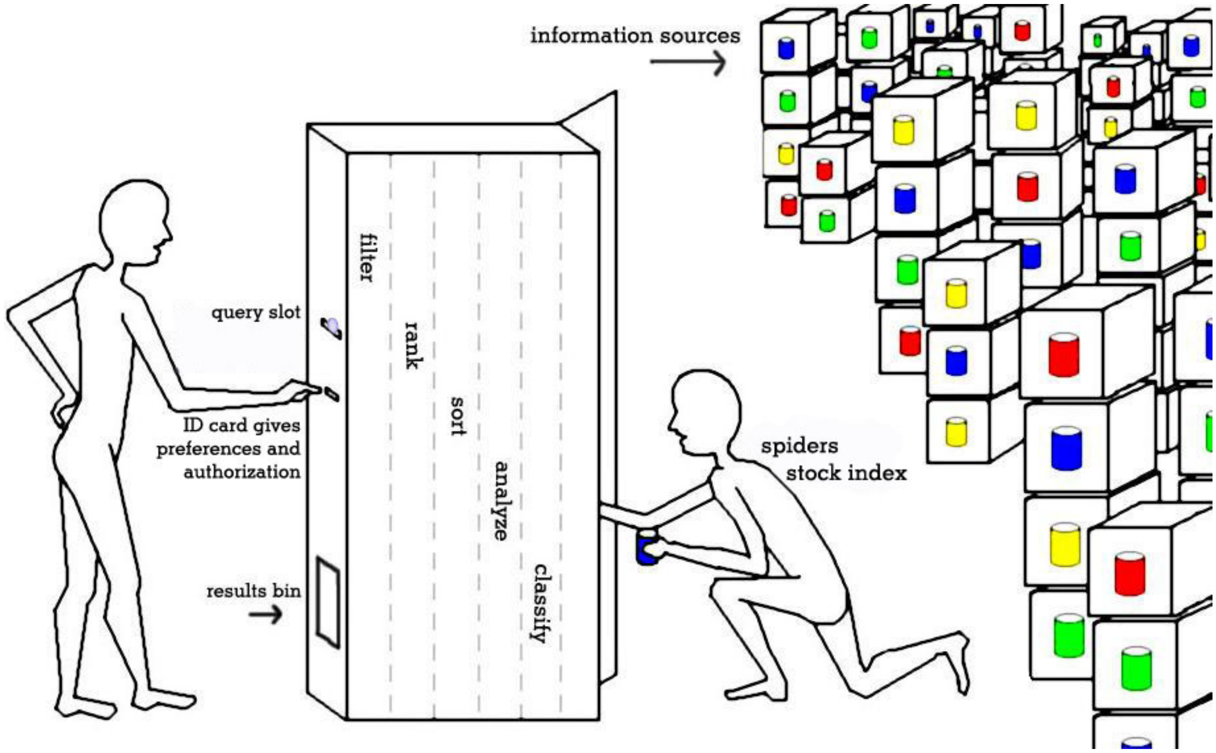
For What Will You Use Search?

- ***Do you wish to use search to enrich an application or specific process?*** Search is an essential part of many applications. Many enterprise search products are targeted at the production applications of a particular industry, for example life sciences, or at a particular horizontal function, such as content management. So think carefully about what scope of search will do the most, both short-term and long-term, for your business.
- ***Are you using search to enhance organizational functionality such as reporting?*** A routine of search can push information about classes of events as they happen, enhancing corporate governance (and, to some extent, proving the negative). Search can be used to ferret out non-compliance and non-standard operations Search may be used in conjunction with a portal-style publish to synchronize focus, philosophy, and messaging across the enterprise. Search also may be used as an expertise and experience locator, to promote collaboration, but this may require exporting of some information to a neutral repository, unfettered by local authorization constraints. Search can work hand-in-hand with e-learning systems to provide on-demand learning. The challenge in these broad uses of search is to identify tangible, countable benefits that justify a more broadly targeted initiative.
- ***What kinds of things will your searchers be looking for – facts and documents, or concepts and patterns?*** Will they be looking for an event or threshold, or lack thereof? Are they knowledge workers, or are they doers who need organizational support to be more effective? On a more pragmatic note, what kinds of exposures does your enterprise have to litigation, and what kinds of data will you have to search to prove the negative on a discovery mission?
- ***Are you using Enterprise Search to support ebusiness****?* As with a vending machine (see analogy in Exhibit 2, on the next page), the enterprise hosting the search can influence the index (the stocking) and the results. If search is to be used by customers at a site where revenue optimization is the goal, the domain of information is limited. A properly tuned search engine can optimize the opportunities to up-sell, cross-sell, and feature particular products, based on a user profile. Quick availability of fresh information is highly important, while proper personalization and respect for the privacy of customers who want anonymity will drive use of (asynchronously-derived) customer profiles.
- ***Is your organization primarily concerned with content dissemination or publishing?*** Search not only gives the organization a much more dynamic overview of its information assets, but also can support push publishing to communities and individuals who have profiled their interests.
- ***Is your organization primarily interested in research?*** A well-designed search engine can be a draw to researchers, and a similarly well thought out search-and-push strategy can keep research synchronized and knowledge shared.

Your Information Domain

- ***What is the extent and nature of the information your enterprise uses****?* (See, again, Exhibit 1.) Is it self sufficient, using only internally-generated information? Are there self-sufficient departments within the enterprise? Do you have partners and suppliers with whom you share information, and, if so, how? Through a portal? Through a neutral space or application? Through the vagaries of email, perhaps as attachments? Does your enterprise's use of information include external information, or a need for Web search capabilities, or is this something you wish to preclude?
- ***How intelligently organized are the sources of information you wish to survey****?* Databases have inherent structure and many ways for querying and exporting information, but some databases have text fields and blobs that must be mined to expose their information content. Much of an enterprise's data is in unstructured files like email and documents, but most of the facts that underlie traditional business analytics are in production database structures that are well indexed by the application.
- ***Consider your unstructured information.*** If it is web-enabled, HTTP provides some metadata, but you may need more metadata to use it well. Internal files may only have file system metadata, which may be inadequate for your needs. How much of your data is structured, and how much is files that will require full-text

Exhibit 2 — Basic Search Engine Architecture: A Vending Machine Analogy



search? How dense is that full text with significant concepts, facts, measurements, and other deliverables? How much of your assets are graphics and images? Do you, or does your industry, have a well-maintained taxonomy? How quickly is it evolving? Does your institution have a body of slang, code names, and other unique features?

- ***What search functions do you already have in your databases, operating systems, file structures and applications*?** Are they sufficient? Will they be sufficient for the organization two years hence? Will you need more micro-analysis? More macro-analysis? Are you using more external sources of information or do you expect to? **What you seek are the shortfalls of information access and the barriers inherent in your informational structures. For it is there that enterprise search can be most useful to your organization**.

Results presentation

The last, but not the least, consideration is the matter of *results presentation*. If the information is not presented to users in a good way, buy-in and/or use will be low. Results can be filtered in many ways. They will, of course, be filtered so that users only see results within their authorization. It may be useful to build a relevance profile[2] for certain power users to increase the efficiency of their searches.

Do you want the search engine to notify you of new information found? How do you want your results prioritized? If the search is broad, do you want results clumped by sub-category? Do you want the results graphed to axes of particular relevance to your position, so that actionable patterns will jump out at you? These are just some of the options being brought to market.

Answers to all these questions will determine what features you will want to look for in enterprise search. If you consider these questions with your eyes firmly on your business, you can prioritize and characterize the needs of individual departments and processes as well as the ongoing needs of the enterprise as a whole. This profile can be used to test the fit of various options.

Search Engine Basics

A basic search engine inherently is modular. Generally, there are two less compute-intensive areas of functionality and a core of more compute intensive processes. (See Exhibit 2, on the next

[2] For instance, some users may be interested in sales collateral and marketing documents, while other users may wish to avoid such material.

page.) The two lightweight processes are the *crawlers*, also known as *spiders*, that find and index the information to be searched, and the browsers used for fielding queries, that are often sited on an independent portal. The classification, analysis and filtering that comprise the query processing are more involved and more highly integrated. Very scalable search engines break out these lighter-weight processes and run them as separate components, allowing scalability to be added where it is needed, i.e., to the crawlers for a wider search domain and to the query interface to support more concurrent users.

The browser used for fielding queries is usually part of a separate environment, perhaps a portal or a pop-up window, opportunities for personalizing the browser interface to make it more useful to the business role of the user abound. A good enterprise search engine can support different front-end interfaces, such as HTML, XML (Web Services), Java, or C++.

Search engines also have an administrator's dashboard and a developer workbench, as well as a business control panel, whereby a person with business, not technical, knowledge can tune the search engine to meet the needs of the business.

The following is a look at the typical elements that are available today, bearing in mind that no product has all the amenities, and that new ones are added frequently.

Query-Side Amenities

- ***Basic query capabilities*** include Boolean searches, limitation by time or source or data, and phrase detection[3], wild cards and support for multiple data types.
- ***Add-ons*** include approximate matching, anti-phrasing[4], multi-language support and natural language comprehension; ambiguity resolution algorithms for things like homonyms and acronyms; the word normalization of lemmatization or de-stemming; proper name recognition and obscenity filtering.
- ***More advanced features*** include query pattern analyzers, navigation, query focus interactions, the ability to specify a context, the ability to analyze the query semantically (by its meaning as a whole) as well as syntactically (as a litany of words), and the ability to navigate through these processes to perfect a query. These supplements, usually presented as brokers, librarians, etc., may be well worth the additional burden of intelligence they add to the client-side or browser software, if they make queries more productive.

Information Source-side Amenities

The basic capabilities for crawlers or spiders include support for a range of data sources, formats, and languages. Indexing is often done at this tier, as it is usually done just once, and scaling the hardware dedicated to this tier is cheaper than adding cost to query processing. Some search engines, in the interests of keeping the costs and response-time down, limit crawler capabilities to the keep-it-simple of *find*. The crawler server tier can also be expanded for faster ingestion.

Indexing and Classification

This is where many the questions asked earlier become relevant. There is *a spectrum of indexing and classification schemes*, ranging from the traditional classifications, like library classification schemes with their vehemently negotiated taxonomies, to the more self-generating continuously evolving mechanisms architypified by *Wikipedia*. The former are quick to use but can be difficult to keep current. The latter are current and trendy but need oversight to maintain quality. Traditional indexing uses words and word phrases to derive meaning. Bayesian logic and other semantic approaches are more involved, but offer the ability to discover patterns that the user could not anticipate. Because they organize based on the information put in, they evolve much more easily. They are more responsive to the consensus of the source data - which is good as long as you are aware of the data domain limitations, and the risks of subjective bias. Combining both kinds of analysis creates better relevancy.

XML and its variants offer many opportunities to enhance information at the application level with metadata elements to document context and relationships. And, of course, a search engine should be able to use a variety of approaches, find all, and quickly do an incremental index of new information. This is particularly important in information-push situations.

Similarity recognition is a key part of any classification system – but real life offers countless vectors of similarity. It is important to think about your organization and industry, and whether wither will change significantly in the near term. The greater the rate of market change, the more appropriate the semantic and entity extraction and consensus ends of the indexing and classification

[3] Typo mitigation and spelling correction
[4] The removal of superfluous words

spectrum may appeal.

Filters and Ranking Mechanisms

There are many popular filtering and relevance criteria, and presentation modes, of which the following are only a sample.

- ***Security filters*** are obviously important to use of search with non-public information. Exporting information to a repository does not automatically remove the need for authorization and authentication. Through deft use of encryption, ways are being developed to search information for congruencies, without revealing particulars.[5] There is much to be developed in this area, and a lot of cultural norms that may have to be revisited.
- ***Personalization*** uses knowledge of the inquirer to limit results, not only to the person's authorization, but also to the person's role or preferences.
- ***Authority*** limits results to respectable or relevant sources. A search for technical or medical information often benefits by specifying sources.
- ***Completeness*** ranks an exact index match before results that merely contain the term, often out of context.
- The ability to ***rank by newest*** information, or to push new information pro-actively, is what lets enterprise search be used effectively for corporate oversight and governance, if the queries are well crafted and the triggers to push the information are chosen well. A lot of the effectiveness of the newest filter depends on the completeness, regularity, granularity, and veracity of the information sources.
- ***Geographic rankings*** are useful when one is looking for results in a limited geographic area for a job search, or for a restaurant when one is visiting an unknown city.
- ***Popularity*** is the ranking algorithm used by many Web search engines. It is generally useful to spot trends (including trends within an organization), but can hide unpopular specifics.
- ***Shopping style ranking*** by a particular characteristic (lowest price, quickest availability, etc.) can be used to find resources and opportunities within the organization, if business attributes are used as characteristics.
- ***Binning, or classifying search results in subcategories,*** can reveal unexpected relationships and also can show where information sources are sparse or spotty. The categories can be hierarchical, Boolean, or multi-dimensional. They can be works of art that reflect and enhance an organization's strategy. Alternatively, they can be a simple set of taxonomy elements.
- ***Visualization*'s** graphic depiction of data has been used effectively in many industries from automotive design to life sciences to present abstruse information in a way to make interesting stuff easier to identify. When you think of areas of business information like product information and related sales and distribution strategies, visualization may prove to be the tool that allows the enterprise to farm complexity for insights.
- ***Custom tuning options*** include absolute and relative query and document boosting that would seem scandalous in an academic setting but may be appropriate in an environment with a targeted focus. Just as the shopping-style option will get the user quickly to his or her characteristic of choice, so tuning the search engine can enhance the efficiency of a commerce website – or highly-focused research project.

Plan of Attack

- ***Consider your organizatio****n* – its ecosystem and its markets, and how much they are likely to change in the next few years
- ***Consider the data sources you use*** and whether they need custom or particular kinds of classification and indexing to be used for your purposes.
- ***Consider the volume, kind, and purpose of the queries*** that will be performed on these data sources.
- ***Consider the profile of search capabilities*** already in your technology infrastructure.

Determine whether there are unmet needs, particularly looking to the future and whether they would best be met by a *general-purpose* product, a *customizable* solution, a product with specific *special features* such as query navigation or visualization of results, or an *industry-specific* product. In search, one size or shape does not please all. For example, a governmental organization looking to track the long-term effects of Agent Orange will have different needs from a start-up technology company looking for market niches in which to bloom.

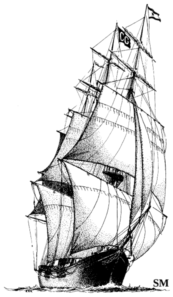
Will enterprise search eclipse other forms of

[5] IBM's Entity Analytic Solutions are exploring this area.

data organization, like databases? Obviously not for transaction processing, where locked-down security is essential, speed of the essence, consistency of process is greatly to be desired. Nor does it replace content management systems where much business knowledge and logic is built into the source-side data organization. However, it can enhance these environments, and stem the use of data structures such as databases for uses not suited to their strengths, like static data or tables of text. Search may be enhanced by other technology-enabled forms of data organization, like Geographic Information Systems, that organize information about geographic areas and points of interest as maps. An enterprise search engine is basically a consummate librarian, who knows where to look for information, and knows what new sources of information have become available. However, better than a librarian, the search engine also has the ability to analyze the information and present it to the user in a custom-optimized form.

Conclusion

The existence of enterprise search can provoke a fresh perspective on how enterprise processes, both production and management, might be enhanced. Business success is often a matter of putting the right information in the right place at the right time. Search, used as part of your infrastructure, gives a very effective way to get it there. Think of how enterprise search might benefit your enterprise.



About The Clipper Group, Inc.

The Clipper Group, Inc., is an independent consulting firm specializing in acquisition decisions and strategic advice regarding complex, enterprise-class information technologies. Our team of industry professionals averages more than 25 years of real-world experience. A team of staff consultants augments our capabilities, with significant experience across a broad spectrum of applications and environments.

- ***The Clipper Group can be reached at 781-235-0085 and found on the web at www.clipper.com.***

About the Author

***Anne MacFarland* is Director of Infrastructure Architectures and Solutions for The Clipper Group.** Ms. MacFarland specializes in strategic business solutions offered by enterprise systems, software, and storage vendors, in trends in enterprise systems and networks, and in explaining these trends and the underlying technologies in simple business terms. She joined The Clipper Group after a long career in library systems, business archives, consulting, research, and freelance writing. Ms. MacFarland earned a Bachelor of Arts degree from Cornell University, where she was a College Scholar, and a Masters of Library Science from Southern Connecticut State University.

- ***Reach Anne MacFarland via e-mail at Anne.MacFarland@clipper.com or at 781-235-0085 Ext. 128. (Please dial "128" when you hear the automated attendant.)***

Regarding Trademarks and Service Marks

The Clipper Group Navigator, **The Clipper Group Explorer**, **The Clipper Group Observer**, **The Clipper Group *Captain's Log***, **The Clipper Group Voyager**, and *"clipper.com"* are trademarks of The Clipper Group, Inc., and the clipper ship drawings, *"Navigating Information Technology Horizons"*, and *"teraproductivity"* are service marks of The Clipper Group, Inc. The Clipper Group, Inc., reserves all rights regarding its trademarks and service marks. All other trademarks, etc., belong to their respective owners.

Disclosure

Officers and/or employees of The Clipper Group may own as individuals, directly or indirectly, shares in one or more companies discussed in this bulletin. Company policy prohibits any officer or employee from holding more than one percent of the outstanding shares of any company covered by The Clipper Group. The Clipper Group, Inc., has no such equity holdings.

Regarding the Information in this Issue

The Clipper Group believes the information included in this report to be accurate. Data has been received from a variety of sources, which we believe to be reliable, including manufacturers, distributors, or users of the products discussed herein. The Clipper Group, Inc., cannot be held responsible for any consequential damages resulting from the application of information or opinions contained in this report.