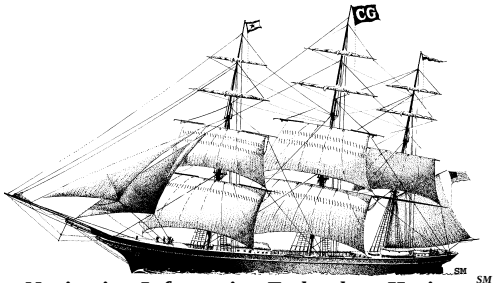
Clipper Notes™
VENDOR-AGNOSTIC EXPLANATIONS AND ADVICE FOR THE INFORMATION TECHNOLOGY BUYER



Navigating Information Technology Horizons℠

New in 2007 Report #TCG2007090 Updated May 5, 2008

How to Derive Value Effectively from All Your Business Information

Analyst: Anne MacFarland

Management Summary

When watching the Red Sox play baseball in Fenway Park, one notices that the most devoted fans do more than stare at the field, with an occasional glance at the scoreboard. They prefer to be within range of a TV monitor, to get the replays that Fenway has decided not to show. The real addicts will access additional statistics on their PDAs, and will listen to the talk shows that follow the game to get a fuller sense of what has occurred. This is a rich base of knowledge.

Now, imagine what would happen if you chose then to follow baseball by only a key performance metrics – maybe the standings, the endurance of the starting pitchers and the attendance figures for home games. It might be adequate, but it would certainly not be a full enough view to satisfy the baseball fan, let alone someone who was trying to run the team.

Most businesses have the complexity of a competitive sport. They produce many kinds of information as a part of their processes, and amass other data sources that are relevant to the business. It becomes an overwhelming amount of information. Disciplined use of information, including the development of key performance indicators for various parts of the business, has brought rationality to the art of management – but has narrowed the scope of how we look at our organizations to only the data that we can address in a disciplined way. It is safe to assume that a lot of opportunity is going unaddressed.

We must extend that discipline and use it to address business data that is not so easy – the often-crucial documentation of e-mail, the unstructured information of documents and external information sources, and the information still in paper form that must be captured. We must leverage the opportunities provided by sensors like RFID and video – including the digitized surveillance that deliver a different view of operations than *ex post facto* reports. We must use information not just in presentation mode, as if they were paper, or as search results, but also to enrich existing processes to address new opportunities in a flexible, well-focused way. We must address them in a way that can transcend the traditional stovepipes of file systems and databases.

More comprehensive strategies are needed. They will build on the traditions of data analysis, business intelligence, and content and knowledge management. They will recognize the constraints and leverage the opportunities posed by semantics and granularity. They will provide the depth and breadth of awareness needed to pursue business opportunity. For more insight, read on.

IN THIS ISSUE

Uses of Business Information

The role of information reuse in business can be as operational documentation of what has happened, information patterns about markets and competitors, and the use of both the above as fodder for business strategies. In addition, information can support operational process optimization and localize faults. It can be analyzed to derive trends. These patterns of how things work let people explore how things might be done differently and where such a change would be beneficial. Information also can become a saleable product in its own right. Such saleable information is not just customer lists, but also information processes or information services that offer more value when shared with others than hoarded as a proprietary differentiation.

All these ways of using business information have been around for a very long time. But, in the past, the digital information from IT applications was destined to become printouts and reports – emulations of paper targeted at human consumption. Of course, applications also consumed information, transformed it, and called for more. Nonetheless, aside from structured data in data warehouses, business data was not seen as a repurposable asset by the IT environment.

Many people are so used to working with the data inside an "application stovepipe" that they see it as a natural act, not a process. When faced with a new kind of information, often people are at a loss as to what to do with it. If it does not pertain to their job responsibilities, they may dismiss it as unimportant. If you think of use and reuse of information as explicit processes, not "natural" capabilities, you can start to see how extensive use of data can improve organizational self-knowledge, and ease the pain of corporate changes.

The traditions of the databases that hold structured data, the content management and file systems that hold unstructured data, together with the new challenges of huge data sources like e-mail and RFID, beg for some high level thinking to figure out how they can and should work together. The granularities at which data is used, and the semantics that underlie the mismatches that get in the way of that use also must be considered. Only they can one construct a comprehensive plan for better use of business information.

Structured and Unstructured Data

Whether business data is structured or unstructured is not as simple as the difference between a numerical value versus text – and an image. Databases have been used to contain all sorts of unlikely things, and numerical data is scattered pervasively in all sorts of places. The difference is the nature of the structures in which the information is persisted. Structured data derives a lot of its context, and its value, from the database structures in which it resides. Text in Content Management Systems is similarly manageable by metadata – but usually only a small slice of business information is kept in content management systems. File systems give some structure by their hierarchy (and, possibly, by naming conventions), but these structures are often set up for the convenience of the user, not to clarify the nature of the content. They often obscure it. Email systems provide basic attributes, but more particulars and relationships must be derived by analysis. Because of the inconsistency or lack of structural context, there is no one strategy that can be used to prepare information for both use and reuse.

One Version of the Truth or Many? Data Cleansing vs. Data Enrichment

The structure of a database is carefully crafted for its intended use. Choices are made determining how values are derived and how relationships are structured. If data is important that do not meet those criteria, the values may not be usable, or the results meaningful. (For more, see Exhibit 1, *The Constraint of Semantics*, on page 3.) Databases work most effectively with data that has been processed, or *cleansed*, to meet the expectations of the structure that will use it.

Data Cleansing – The Mantra for Structured Information

The need to cleanse structured data comes from a variety of situations. If the data has been scanned or entered inaccurately, then typos and misplaced decimals or six digit phone numbers – any kind of unexpected format – make the information unusable. If the underlying business logic or data definitions of multiple data sources are inconsistent, federating or importing information from different sources results in more compounded mismatches. If you can only address 70% of the values in your database, and cannot count on consistency, your analysis will be relatively useless. So, data must be cleansed to make it fully usable.

Structured data is presented as views tailored to the reader. Analysis of structured data is a well-developed art. You might think of structured data

as the baseball statistics that underlie the announcer's patter. By contrast, unstructured data is more like the color commentary.

Data Enrichment: The Opportunity Found in Unstructured Content

In unstructured data, the context is given by the content. Whereas structured data comes as a value in context, the value of unstructured data often has to be derived by indexing, text analysis, and other tools. A richer, more comprehensive understanding of a situation is given by aggregating multiple sources of unstructured data. Anomalies, which in structured data might be considered something to be cleansed, in unstructured data provide the nuances that build out a more complete picture of a situation.

Textual analysis can come in the form of entity extraction – the identification and indexing of significant entities[1]. This can be done by a positive match to a taxonomy or simple list of hot topics, such as code names of products under development. It can also be done by filtering – discarding articles, pronouns, adjectives, adverbs and most verbs and focusing on what is left. De-stemming, a technique that allows variants of the same word root to be associated, can organize what is left after filtering into more discrete terms.

Once you find all the instances of an entity, one can look in the contexts of the occurrence to derive relationships. Relationships between entities, once documented, start to build a comprehensive view of a situation. This can be done with database tables as well, but only for domains where the relationships are known, and, usually, it is the changes in the values that are interesting.

The Commonality is Metadata

In databases, table structures are persisted as metadata. So are text analytics in unstructured data analysis. In the past, metadata has not been standardized, and data schema has not strived for conformity. With the broad adoption of standard XML and all its variants, greater congruence of semantics, and broader analysis of data is now possible.

The more widely data is used and reused, the more the metadata that can be produced and kept to enrich the value of the data. Over time, with active reuse, metadata can dwarf the size of the original data, and becomes a structurable asset in its own right. The metadata that gives structure to unstructured information, in bulk, can be manipulated and searched like structured data.

Conversely, when structured data is broadly aggregated into a federation of databases, extracting and reconciling the semantic and structural differences between data structures becomes somewhat like the practice of (unstructured) text analytics. Therefore, both cleansing and analytics

Exhibit 1 — The Constraint of Semantics

As enterprises have grown and become more distributed in geography and in control structure, they have developed multiple data sources that are often heterogeneous in nature. Getting them to work as a single federated system requires some data transformations – and also an awareness of semantic differences. Such semantic mismatches are particularly pervasive in organizations that have grown by merger and acquisition. In databases, semantics crop up in definitions and how values are derived. In unstructured information, the problem is more a matter of vocabulary, and where and how broadly certain words are used.

Supporting analysis across large environments of structured data often requires a lot of reconciliation. This is the basis for a lot of Master Data Management strategies, and system of record initiatives, by which heterogeneous stores of large amounts of information can be kept in synch. Now that most forms of business information are digital, these strategies are both possible and necessary.

Master Data Management is a strategy to enforce use of current information by maintaining a golden image repository of information in a secure location. Use of the data kept peripherally triggers a message to the repository to check for updates. This is a useful strategy for product information, including product numbers that, these days, must be globally consistent. It is an important strategy for about sales information, where marketing campaigns change prices for short periods of time. It aims to replace the cherished paper printouts, which have been convenient, but also the source of introduced misinformation. Obviously, Master Data Management relies on a good network, and may raise telecommunications costs, but it more than makes up for it in better business practices.

[1] For an example of unstructured information analysis, see **The Clipper Group Navigator** entitled *When Your Life Depends On It — Inxight Federal Systems Enhances Military Intelligence*, dated August 17, 2007, and available at http://www.clipper.com/research/TCG2007282.pdf.

must be a part of a comprehensive strategy to use more of your organizational data better.

Semi-Structured Information

There is a lot of information that does not fall neatly into the categories of structured information and unstructured information. E-mail is a good example, and one that no organization can ignore. When the email is with correspondents outside the organization, it represents a significant source of corporate risk (as well as the overwhelming bulk of the corporate opportunity). This risk and opportunity, combined with the privacy risk inherent in any data involving individuals, is also the primary documentation of corporate malfeasance. None of these attributes is benign. An email has standard attributes (e.g., sender, recipient, subject, data), but its use may not be directly indicated by any of these. The use and value can be derived through text analysis. The volume of email argues for a careful analysis of risk and opportunity indicators to filter out routine, low-risk correspondence whenever possible, from the bulk to be analyzed. You may have to keep it all for compliance, but analyzing it all in any significant detail will seldom be economically supportable.

Other semi-structured challenges and opportunities include images and image sequences (video). Video can capture the reality of unattended or distributed processes. Think of the surveillance cameras in convenience stores. While they help identify robbers, they also document when the shelves were stocked and what deliveries did not show up. With pervasive digitization, there are many useful reports that can be generated from video feeds.

Sensor and RFID information, in the aggregate, provides a similar insight into operations. The view has a pointillist-style, rather like the visualization analysis of many data sources. The cadence and patterns of e-mail can have a similar in-the-aggregate value that is quite separate from their value as documentation.

This is a far richer palette of information than that of key performance indicators, though KPIs can be very valuable for measuring specific goals. The more complete palette of information provides a far more detailed view of operations than that found in well-massaged internal reports that pass up and down the hierarchy of business. Analysis of both structured and unstructured information offers the opportunity to increase the transparency of organizations, so that those in it are not constrained to the outlook of their immediate environment. Such transparency can be incredibly powerful.

It can also be distracting. One must remember that democracy, in the short term, is not the most efficient way of addressing a task. To anyone looking for a particular and verifiable piece of information, the current mania for the wisdom of the crowds can be, at times, frustrating as well as enlightening. Take a look at the kinds of information your organization generates and figure out when you need one version of the truth and where you want rich corroborative enhancement of a complex situation. The results will guide your strategy for getting value out of all you business information.

Leveraging Granularities

This aggregating into chunks or dissecting into piece parts is a basic way humans learn about things. When working with information, we have described how, with structured information, we often drill down to find out more, while with unstructured information, we often find more by aggregating sources of information about a particular event. Both these techniques inform the strategy that can be used to use and reuse business information well.

Drill Down Insights

The details of a situation may contradict expected assumptions. Unexpected nuances may enrich your view of what is happening, and let you move beyond the knee-jerk reactions based on preconceptions. This is why you do analysis.

Drill down can also expose unexpected areas of congruence between ostensibly dissimilar situations. Such unanticipated congruencies may offer opportunities for optimization – be this in sharing processes or in scheduling asynchronous processes around each other.

High Level Realizations

Amassing information and looking at it at a higher level of abstraction also has benefits. Instances become part of larger patterns. You may find that operational difficulties are not peculiar to a particular situation but may be part of a larger pattern that is amenable to change at different action points than the local situation would indicate. An example of this is tracking information. It can be used at one granularity to tell a customer where a package is, and at a higher granularity to optimize deployment of delivery vehicles.

Mash-together Revelations

Recently, mash-ups of multiple sources of

information have been shown to reveal new insights. Now, matching business data with maps is very popular. This is a *data enhancement* – it adds standard information to expose information in a way that is easier to use, or show new patterns. Geographic locations do not change. There is some granularity to them (counties, states, and provinces, countries). These coarser granularities are subject to change, but not usually in a time frame that will affect business tactics. The most valuable mash-ups come from aggregation of real-time, rapidly changing data. Think of trading desks or logistics, where feeds of information drive time-sensitive decision-making. Of course, aggregating RSS feeds from public sources do not involve the security ramifications that might come from mashing up sensitive business data.

The Domain of Relevance

Most of business information is specific to a line of business or a process. There is a need for confidentiality that constraints its use beyond its originating domain, and a need for its destruction at the end of its mandated retention period. For the most part, reuse of this data is limited to its use as documentation of a business activity. Subsequently, analysis of it in process-wide chunks can determine how to hone existing processes, or refocus them on new markets.

Two Valuable Tools: Classification and Indexing

Classification and *indexing* are two complementary tools that give you the knowledge about your information that will help you use it appropriately. Indexing is a data–centric data tool. Databases have long used indexing to support analysis of structured data. Now, indexing engines can analyze the file system path and even the text of unstructured data to derive the appropriate index terms that will make the appropriate information more findable.

Classification is a usage/user–focused tool. Classification assigns categories into which things are sorted. The categories will depend on the goal of the exercise. If it is to determine retention periods, the categories will be different than if it is to determine corporate risk. Classification is a way to determine what media information should be stored on, based on how quickly it must be retrieved. We use classification-style skills every time we plan a project or a vacation.

Indexing tells you what you have. *Classification* then lets you determine what should be done with it. Of course, there are many stakeholders that want to do different things with information. No one party knows what is best for all parties. This may be why adding classification and indexing can seem to produce more confusion than clarity.

How to Get Started

When considering data strategies for chronic use of business data, it is tempting to start small. However, the long run is better served by surveying all of the kinds of information you have to work with. Some sources of information have self-evident and often limited value, like that derived from instruments. For more motley sources like e-mail, classification and indexing can give an organization a general handle of the character and extent of its business information. Most business information will be of limited or transitory interest.

Start with Pain and ROI

Look for a problem begging to be solved–one that demands broader use of business information. This may be a specific customer–facing problem that spans business units or a potential merger that would have complex effects on both organizations. In all cases, look for *measurability of success*, for business value beyond ROI, and some *corporate champion* to whom that success is important – someone who will push for completion when the project gets hard. Where will resolving the problem give the most bang for the buck? Get your early experience here, for the benefits of solving the problem, and the experience of how you did it, are good ambassadors to help expand the process.

Think about organizational semantics as well as data semantics

Classify the relevant stakeholders by their information needs, and the relevant business information sources by its role (operational effectiveness or s vendable asset?) and its risk (what kinds of controls must be imposed?). Model how business access for reuse process will work.

Sales folk have different needs for information than product development or oversight committees. Each will have a good idea of what they need to do their job better, and also what they need to anticipate how their job may change. Modeling the requirements of your users can help determine the elements of your solution.

Create opt-in communities to test new uses for business information from power users, and key players in a process that is customer facing or otherwise has high potential. Use them to define

what enhancements to metadata are needed to accomplish the scope of what they want to do. They, as business folk, should develop the justification that IT will need to fund the project.

Think about Granularity

Differentiate between looking for particular data, looking for expertise, and looking for understanding. They are three different things. Looking for an answer, particularly in a business context, is a very granular activity. Your concern is that the information is current and has not been tampered with or transformed in any way. Looking for expertise maps more to a query kind of list. Looking for understanding is a still broader approach, often using visualization of data as much as analysis.

What needs to be done will also depend on how business users want information delivered. As a push? As a dashboard? As alerts? Or, as a sandbox of resources with which to play? Many people will want all of the above.

Prune the bulk of data that must be addressed as is possible

In all this talk about reusing business information in various ways, it is important to remember is that reuse is easiest of information that is not relevant is identified and excluded. This should be an iterative process.

Most information initiatives involve a place (a geography) and a time (data range). This is true of legal discovery, of due diligence, and of initiatives to address new markets. Determination of the relevant time and place can help exclude what is not relevant. Teach users to qualify their desires for information.

Model

Any solution involving disparate groups of users using disparate data sources is most efficiently addressed by a modular or service-oriented in nature. Users will want some kind of portal-based dashboard where they can specify requirements and preferences – for these will change too often to be run through IT. Data sources will have needs for transformations and staging that are best met by some kind of front-side information server appliance. A middle tier – the realm of operational data stores and content management systems, each optimized for a subset of data – is often useful. Such subsets could be product information, educational and training materials – all cases where completeness on as small a scale as possible is desirable.

There is a data side to the need for modeling as well. Both relational databases and content management systems are presentation-oriented. They are usually on the receiving end of process, not in a participative mode. The most effective projects will not expect more out of the data structures than that for which they were designed. They will develop processes to facilitate the data use that seems, by turns, second nature or impossible inappropriate, to many shareholders.

The emphasis of this middle tier capability is also on composability and decomposability. In the structured world, this is a matter of data sets and drill downs. In unstructured, it is visualizations and the various kinds of entity extraction and analysis by which more information can be derived. In the middle, there is a need for some kind of arbitration to broker the data services that are consumed by users – and to track usage and charge back, when appropriate.

One then must think about the system architectures and data structures that will best support what you want to do. System architectures will be a matter of response time and security. Data structures will be a matter of controls and self-sufficiency. The latter is particularly important for information that will be used by many organizations or kept for a very long time. Of course, what can be done depends on budget.

Conclusion

The information a business generates is one of its greatest assets. This asset includes, not just the information in its databases, or its emails, intellectual property and contracts, but also the operational evidence that is the hard evidence of how it conducts business. To exclude any source of information as messy or mundane can be a competitive mistake. Think more broadly of how you use your organizational information. Explore which sources of information can tell you what you need to know. Implement the strategies that will present the right information to the right decision makers at the right time. Plan carefully to assure early success, and your new business intelligence will repay your efforts handsomely.

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 Data Strategies and Information 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**, **Clipper Notes**, 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.