Freedom from Facets

Discovering the data you really need

October 2011

A White Paper by

Dr. Barry Devlin, 9sight Consulting
barry@9sight.com

Imagine a naïve business intelligence user simply typing “Sales of discounted women’s footwear in Manchester this month” into a search box and receiving back a ranked list of 35 valid, current query results that satisfy this search. Imagine her eyeballing the graphical representations of these results, noticing an unusual spike in sales returns and searching for “Returns by stores where bought in the northern region”. Then diving into the credit cards associated with the sales with returns in different stores. And the Eureka moment...

The example briefly portrays a process of innovative exploration that many of us are used to doing on the Internet using simple keyword searches of Web content. Today, we call it googling. It is a process that has proven elusive for users of enterprise business intelligence systems for years.

“And now for something completely different...”[1] In this paper, we describe freeform search—a method that enables business users with no knowledge of data structures and relationships to explore for themselves their structured business data in a wholly unstructured and adaptive manner. We explain how an information context can be created automatically that supports reliable parsing and interpretation of such freeform searches and also enables agile and speedy delivery of this functionality to end users with minimal IT involvement. And we introduce neutrinoBI, a new product that implements freeform search against data in a combination of data warehouse, data mart and personal data stores and delivers valuable and innovative business insights to decision makers with limited data skills and minimal IT involvement.

Contents

Sponsored by:
Neutrino BI
www.neutrinobi.com

Why do we never get an answer…?

Why do we never get an answer / When we're knocking at the door?
With a thousand million questions / About hate and death and war.[2]

Business Intelligence. If you work for a large corporation or government organisation, the phrase may sound like an oxymoron. If you're a BI software vendor, it may mean something very specific—like your product. And if you're a consultant or analyst, well, who knows…? Howard Dresner brought the term *business intelligence* from DEC when he joined the Gartner Group in the early 1990s. It gradually replaced *data warehousing*, *decision support systems (DSS)* and other terms in common use then. Today, many vendors and analysts are promoting *analytics* of various flavours as the next stage in the evolution of business intelligence.

But, still, we never get an answer—a long-term, widely-accepted answer—to the simple question of how to support decision making in business[3]? Dresner envisaged business intelligence bringing together all aspects of the solution to this problem. In marketing terms, this succeeded. However, many business users would argue that it's still very difficult to get the answer they need, accurately and quickly. And IT groups supporting them would likely point to their ongoing—and often poorly received—efforts to provide users with useful information and tools. Given that it's a quarter century since the initial introduction[4] of the data warehousing architecture that aimed to provide the answers, we might justifiably ask: *what on earth is going on here*?

There are three related issues:

1. **What is the question:** How a business user approaches and formulates a question concerning some business issue is a very personal choice. Many approaches have been tried over the years with varying degrees of success. We explore the options in the next section.

2. **What does it mean:** In the posing of any question, the context in which it is asked is key; and this is particularly so in the business environment. This leads us to consider the thorny problem of metadata, and especially business metadata in the following section.

3. **How do we decide:** The process by which business users make decisions is often far from linear; it is cyclical and iterative in the information used. It often involves peers and superiors. And it is far from pre-defined. We explore this adaptive cycle in a subsequent section.

Before diving in, it's worth posing one more question: why should we be able to solve these issues now, given our inability to do so over the past decades? There are a number of answers. First, developments in exploration of content / soft (often called *unstructured*[5]) information are being applied to hard data to provide a new understanding of query context. Second, the experience gained in 25 years of data warehousing and advances in quality of traditional data sources are beginning to address previously intractable metadata problems. Third, recent advances in hardware power and speed, as well as database structures, are enabling us to apply far more intelligence to anticipate users' needs and guide them towards the answers they require.

What is the question?

The Ultimate Question of Life, The Universe, and Everything?
"What do you get if you multiply six by nine"?[6]

Depending on their levels of skill and subject matter expertise, business users who are making decisions find and explore the information they need through one or more of the quadrants shown in figure 1. The simplest method, and one that been used since business began, is through ***standard reports***, shown in the bottom left quadrant. Such reports are simple to use and often adequate for basic and repeating decision making. However, they are fixed in their content and severely limit exploration or innovative thinking by the user. Despite this, such predefined reports are in widespread use for repetitive decision making and may be a good place for inexperienced users to learn what information is available.

Many users begin to explore by moving to ***spreadsheets*** in the top-left quadrant, an approach much maligned by IT because of its often-damaging effect on data quality. The "approved" approach to increased exploration of business information is through BI tools that provide some form of ***structured query***, most commonly SQL. In either case, exploration involves information to which users already have an understanding of the content and context of their query.

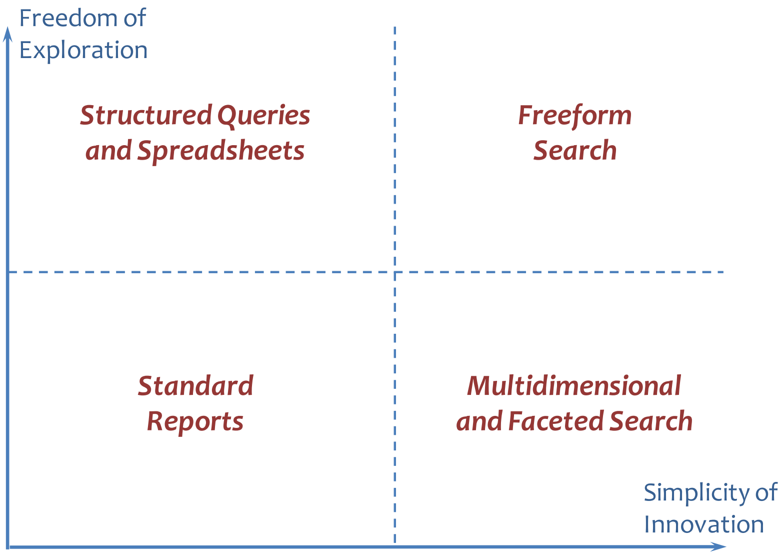


Figure 1: The Innovation-Exploration Question Space

Structured query approaches have been prevalent in data warehousing since its initial development in the 1980s. However, the difficulties that many (or, indeed, most) business users experience in constructing such queries and the widespread performance issues that result even from well-formed queries in traditional relational databases led vendors as far back as the early 1990s to propose OLAP (online analytical processing) cubes[7] as a solution. These cubes, irrespective of their implementation, offer users easy ***multidimensional*** exploration of information, leading us to the lower-right quadrant. While restricted to predefined exploration paths through the data, this approach allows users to be more innovative in their use of information and is particularly useful in the type of drill-down and slice-and-dice types of analysis typical of much standard business usage.

Faceted search[8] is largely a formalisation and web-based application of this approach, extending its use beyond numerical data. Faceted search enables users to navigate a heterogeneous information space by combining text search with a progressive narrowing of choices along multiple dimensions (akin to filtering prepared lists). Every dimension is divided into multiple subsets, each defined by an additional restriction on a property. These properties are called the facets. By selecting multiple facets, the user drills down into the different aspects of the overall question.

Multidimensional and faceted search have proven successful and widely popular in BI and e-commerce applications respectively. However, there are limitations. In addition to the somewhat restrictive, pre-defined drill-downs that users can perform, the biggest issue is in BI data consisting of thousands of columns (for example, in some SAP tables) where hundreds of drill-down paths can be envisaged but cannot be easily defined or presented to the user. Furthermore, in terms of free exploration of the data, these approaches guide users down pre-defined paths of analysis, from which point it can be unclear how best to expand the view again to find another interesting drill-down.

Nonetheless, many of the techniques and tools in these three quadrants, which developed over the past quarter century, remain in widespread and relatively successful use today. They are likely to continue to provide much of the day-to-day business intelligence for a significant subset of users. However, the limitations of these approaches are becoming more apparent in the face of exploding information volumes and varieties. This leads us to the upper-right quadrant.

Freeform search

Way back in late 1998, a small start-up pioneered the minimalist user interface, shown in figure 2, which has set the expectation for search among end-users since. This expectation is that a user can type in a string of words and be presented with a highly relevant list of hits. Google's success sprang fundamentally from the patented algorithms that consistently placed the items of most interest to the user at the top of the list. But, perhaps most significant from a user viewpoint was the stark simplicity of the interface—a single input box and two buttons—through which we perform a ***keyword search***. We have been well and truly Googled.

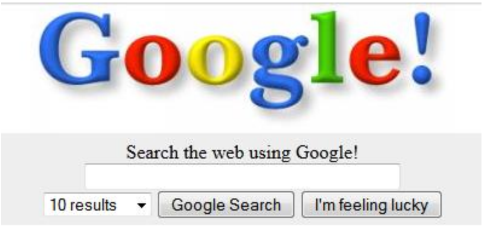


Figure 2: The first Google search page, circa Feb. 1999

With Google's rapid rise in visibility and popularity (to "google" was recognised as a common verb in the Oxford English Dictionary in 2006), it became both the bane and the Holy Grail of enterprise software developers as business users demanded the elegant simplicity of the Google search interface and its almost uncanny ability to deliver the results expected by the searcher.

The challenge (or opportunity to an optimist!) is to translate the approach to the enterprise, an entirely different environment to Google's original design point. The issues include:

1. **Size:** While the sheer size of the Internet poses problems for any search engine, it also enhances the statistical significance of even very small percentages of relevant results

2. **Scope:** The variety and variability in even the largest enterprise information resource is far less than that found within Internet information, reducing relevance of result sets in enterprise search

3. **Sentiment:** The numbers and variety of Internet users similarly provides a much larger range of sentiment contributing to result weighting due to popularity

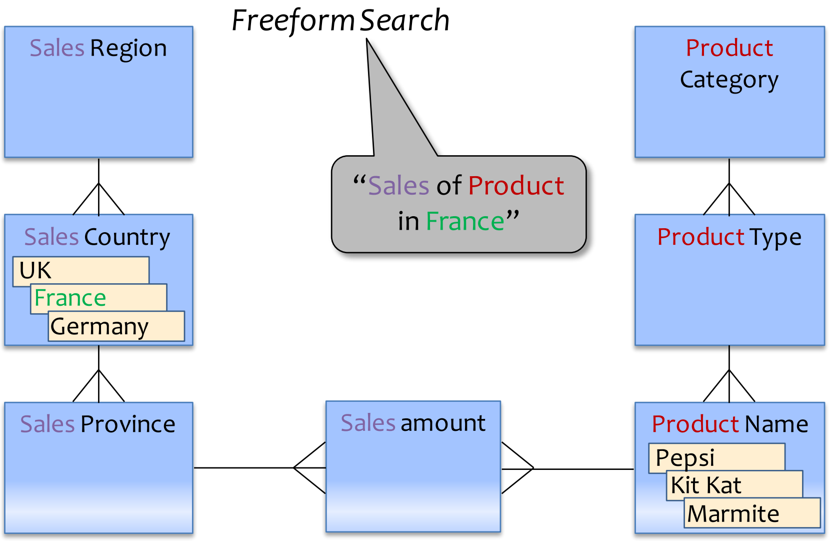
4. **Structure:** Particularly in the case of BI, the data structure and relationships as expressed in the relational database model contributes significantly to information understanding in a way that is uncommon on the Internet

In the case of business intelligence, a key goal is to enable Google-like search of hard information. We call this ***freeform search***. The secret of moving from keyword search of soft information to freeform search of hard information lies in understanding and representing the structure of relational data in a way that supports intelligent parsing of free text searches. This structure originates in data modelling and is carried into relational database design and associated metadata—which columns occur in which tables, identification of primary and secondary keys and foreign-key relationships between tables. Structuring, whether into normalised or multidimensional schemata, also constrains allowable values in certain columns and relationships between them.

The result is a conceptual hierarchical structure that is hardwired in the database or application in the multidimensional approach. For example, geographical location is often expressed in the form of regions, which are comprised of countries, which are further comprised of states / provinces, which break down to counties, each of which contains a known and limited set of values. In the case of freeform search, a process known as ***hierarchical value decomposition*** is used to automatically analyse database structures and create the indexes and metadata—the ***information context***, unique and specific to the structure and content of the underlying data—needed to extract meaning from the keywords entered by the user at search time. Figure 3 provides a simplified view of how the words in a freeform search match to the hierarchy and allow the search to be interpreted and run. The context conceptually shows a "fact table"—sales amount—and two "dimension tables"—geography and product. The colour coding shows where words in the search match into the context: sales and product match the names of the tables at a structure / metadata level, whereas "France" matches at the content level. With many searches, both types of term appear, and the information context must thus span both metadata and appropriate content. The result of the search shown will return a ranked list of possible results with all allowable combinations of data from these different tables.

Of course, the search techniques that work well on the Internet—word matching algorithms based on nearness, relevance and occurrence, for example—also play a role in interpreting the search string. Synonyms from business dictionaries, for example, pseudonyms in common parlance and temporal concepts, such as "last month", are also used to understand the user's needs.

Figure 3: The information context for freeform search



This ***business context***, within which the search is occurring, is a function of both the information itself and its model (information context) as well as the role, identity and even history of the user making the search (personal context). The extent to which all of this contextual information is factored into the search interpretation determines how relevant the prioritised results will be. These are all aspects of business metadata and the subject of the section "What does it mean?"

The techniques above have been described solely in relation to hard data, because this is the aspect that is most relevant to traditional BI. However, soft information is of rapidly growing importance in business and users have a growing need to search hard and soft information in combination with one another as discussed elsewhere[9]. Many of the above techniques apply equally to this search space.

What does it mean?

"If particulars are to have meaning, there must be universals" – Plato

As we've seen in the previous section, understanding the business context of information and the intention of its users is vital to freeform search and, to a lesser extent, the other three quadrants of the query space. Business-oriented metadata is the foundation of this understanding, but gathering and managing such metadata has proven challenging over the entire history of BI. Understanding the nature of the challenge and how to overcome it is vital to successfully achieving true innovation in information use by business people.

The ultimate source of business metadata is fairly obvious: business people themselves are ultimately responsible for all business definitions, and every change in them. The traditional means of documenting business metadata is through requirements gathering and subsequent information modelling based on those requirements. Data quality and data management projects further add to the store of business metadata.

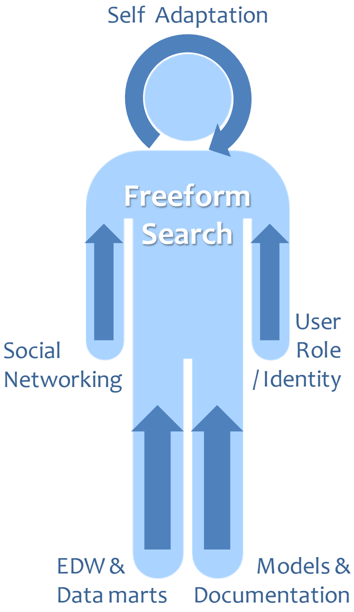


Figure 4: Business metadata sources for freeform serach

As seen in figure 4, sources of such information range from Word documents and Excel spreadsheets, to modelling tools to database comments fields. These sources should be utilised to the fullest degree to populate the contextual information needed for freeform search. The best and most obvious source is the enterprise data warehouse and the modelling tools or design documentation used in its construction. This metadata (where it exists) probably represents the closest approach to common, "universally agreed" business definitions. Such agreed definitions are an excellent starting point that ensures that the initial context incorporates prior data reconciliation and cleansing work. However, we need to keep in mind that different departments and user groups may have different definitions, and that we need to document relationships between these variations. For example, "profit" may include or exclude a number of different factors, depending whether the view is from finance or sales. Such information may be available from data marts or their models or design documentation.

The user's identity and, particularly, role thus indicate which business metadata is applicable in a particular situation. The CFO and CMO mean different things when they search for "profit". Even within the marketing organisation, a regional manager's use of the search term "profit" without any further clarification may most likely mean the profit level within her region, and search should (initially) prioritise results differently for the CMO and the regional manager.

Business definitions, like language, are constantly evolving. In the past, capturing such changes proved very difficult. With the ever-increasing pace of change in business, business metadata morphs more and more quickly. These developments must be captured directly from the users themselves and as automatically as possible. Social networking approaches are key. They must actively encourage users to discuss and document emerging definitions through collaborative tools such as wikis, for example. Changes in user role may also become visible here.

Finally, freeform search tools must capture and interpret search input and adapt to user behaviour and uses of terminology. As all this metadata is captured and made directly and instantly available to users, business users benefit immediately through better understanding of the data and reduced search times and are thus proactively encouraged to contribute to the process.

How do we decide?

"Informed decision-making comes from a long tradition of guessing and then blaming others for inadequate results" – Scott Adams, author of Dilbert

Given that business intelligence is, by definition, valuable only insofar as it supports decision making, we must finally touch on the process through which decisions are reached. Decision making is traditionally divided into operational, tactical and strategic. Here we focus on tactical and strategic decision making, where freeform search is most relevant. In these environments, the process by which we decide is typically experimental, iterative and often times collaborative. We discuss the process in the context of the Adaptive Information Cycle (AIC) which was introduced in a previous white paper[10], driven by these precise needs.

As shown in figure 5, the AIC begins with an event requiring analysis and a decision. This event may be singular, external and localised (for example, a drop in sales of widgets in the southern region) or internal and pervasive (such as the CEO questioning whether to outsource IT), or any combination in between. As a result, an initial set of information is ***recorded*** or gathered and put into a form (***conditioned***) that is suitable for analysis. It is then ***utilized*** in any appropriate statistical or visual tools and, in the simplest case, a decision follows. However, most situations are more complex, and the user recognises and ***assimilates*** the need for more information, more input from peers or deeper analysis. This closes the loop, often leading back to the gathering of more information. The point to note here is that the need for additional data or input is unknown at first—it emerges as part of the analysis itself. Similarly, the direction an enquiry will take is highly adaptive—an anomaly noticed in a result set the third time we loop through the utilize step may shift us into a whole new set of searches or queries on the succeeding cycles.

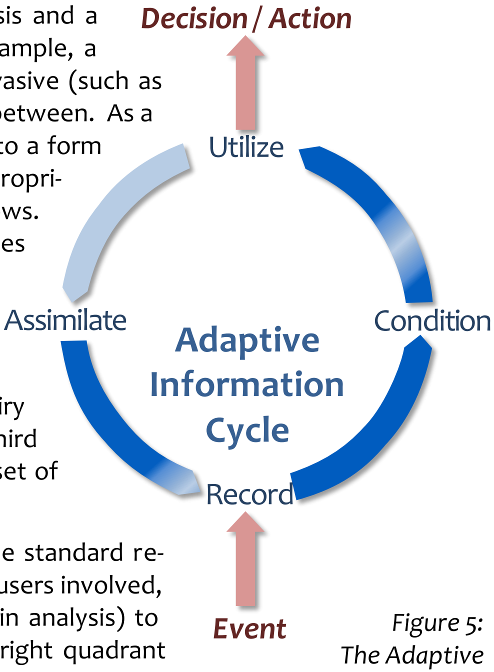


Figure 5: The Adaptive Information Cycle

It is this need for adaptive behaviour that first drives users out of the standard reports quadrant of figure 1. Depending on the analytic skill level of the users involved, they tend either towards the top-left quadrant (those highly skilled in analysis) to use spreadsheets or structured language queries, or to the bottom-right quadrant which demands lower analytic skills to explore the data through drill-down or faceted search. Clearly, freedom of exploration combined with simplicity of innovation, as offered in the top-right quadrant of freeform search, provides an ideal environment for implementation of the AIC.

So far, we've dealt with two aspects of freeform search: (1) keyword-based contextual search that leads the user to relevant information or existing results, and (2) the business metadata that forms the basis for defining the context. The third aspect of freeform search relates to the presentation of results and the linkage across a series of queries and results that provides users with an intuitively clear path to follow in search of the insights they seek.

In the case of a Google search, the ranked results are presented in the form of a list with a couple of lines of text from each result page. Recently, cached thumbnails of the pages have also been made available. In the case of freeform search in BI, the business needs and technical situation are different. While there may be existing reports of interest that satisfy the search, many of the results will need to be calculated on the fly. As the demand for ever more current information grows, on-demand calculation will become increasingly important. Furthermore, the results must be shown graphically, so that the user can judge relevance at the earliest moment.

The opportunity and, indeed, the challenge here is that the user is presented with multiple, ranked results to the search entered. The opportunity is ease of exploration and innovation; the challenge is to maintain focus. (How many times have you searched for a simple answer in Google and emerged a fascinating hour later having explored the farthest reaches of the known universe?) In freeform search, the user must be enabled to easily and visually map the path being taken and easily retrace it as needed. This is a very different approach to faceted search, for example, which progressively narrows focus along a largely predefined path.

Introducing NeutrinoBI

"Accelerate your discoveries"

NeutrinoBI is a new tool that provides freeform search for hard data in the business intelligence environment. The objective is a user experience that is similar to Internet search, but which focuses on BI users, allows for finding and unifying data from a wide range of data sources and enables faster, more agile decision making.

The user experience begins with typing a freeform keyword expression into a search box and clicking search. A set of results are returned in the form of a ranked list and a carousel of graphical thumbnails of the results of running the queries in real-time, with the highest-ranked result on top. Figure 6 shows the carousel. The user can scroll through the results, eyeballing for items of interest. From there, any result of interest can be dragged on to the canvas (workspace) and manipulated in a variety of ways familiar to any BI tool user.

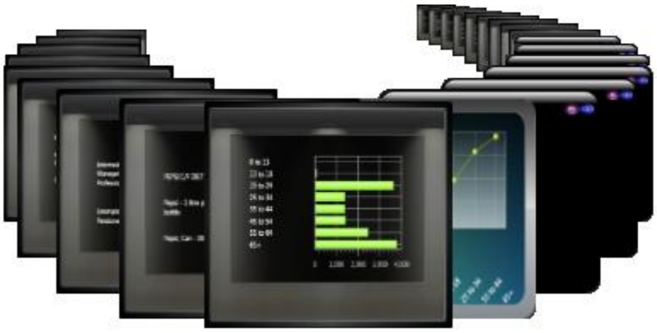


Figure 6: NeutrinoBI search results carousel

Continuing to focus on the exploration and innovation aspects of the tool, the search paradigm is extended in two ways that go beyond content keyword search. First, the user can progressively narrow a search by popping up child search boxes from any search; the results of a child search are a filtered subset of the parent. Second, the user can create any number of primary search boxes in order to explore different ways of looking at the information. All these search boxes and their result sets are maintained throughout the session, so the user can easily flip back and forth between them. Furthermore, results from different searches can all be dragged into the same workspace and graphically combined in a variety of ways. Results in the form of reports and dashboards can be saved and published in standard office tools, and shared with colleagues.

In this first release of the product, parsing of the freeform search input is driven entirely by the information context built during setup. The business vocabulary thus emerges directly from the data sources; a separate ontology (business metadata) is not required, although its use will enhance search term recognition and matching. In the next release, parsing will be enhanced with automated suggestions to complete keywords, support for pseudonyms and temporal concepts. Future plans look to enhancing the search context with personal information about the user role and discovered preferences and behaviours.

NeutrinoBI can access corporate data in data warehouses based on Teradata and Oracle, data marts on these same platforms, SQLServer and Sybase, as well as standard ODBC access to other platforms. Access to personal data in Excel, Access and delimited files is also provided. Users can combine data from all sources (subject to security restrictions), with the search engine visually indicating the level of trust that the user should place on the data depending on its sourcing. Searches / queries are run in real-time, with processing split between the neutrinoBI server and back-end databases. Typically, a small percentage (less than 1%) of the data resides locally in aggregates and indexes that neutrinoBI determines it needs during the initial setup phase.

Rapid, agile setup of the data environment is a very important consideration for most IT shops today, and neutrinoBI scores well in this area as a result of its approach. Traditional approaches require an application designer or database administrator to define a data structure into which data must be loaded before it can be made available to users. NeutrinoBI, in contrast, analyses and maps the data structures of the underlying sources, and builds, in a semi-automated process, the metadata and indexes it needs to describe the information domain. Beyond this initial mapping phase, no IT involvement is needed; changes and additions are auto-generated. Where table and column names are business-friendly, they immediately become the keywords used in search. Manual enhancement with business metadata is required to create synonyms for IT-created table or column names or to support specific business terminology. NeutrinoBI will support import of various types of business metadata and ontology in the next release.

Conclusions

The Answer to the Ultimate Question of Life, The Universe, and Everything?[11] *"42"*

The emergence of the Internet in the 1990s and 2000s drove a great surge of research and development into how to make the content of the Web accessible to non-specialists in information retrieval. Following Google, the most successful and long-lived approach to innovative exploration and location of content by "amateurs" has been "simple" keyword search. That success has been based on statistically leveraging the enormous volume and variety of soft information on the Web. For users with specific objectives for their search and among more homogenous and limited sets of content, faceted search has surfaced as the tool of choice.

This success has whetted business users' appetites for similarly powerful yet simple tools for exploring hard data in the business intelligence environment. However, until now, the very structure that characterises hard data has limited the ability of non-experts in data to achieve the type of innovative exploration found on the Web. Deep exploration was possible as long as you understood data structures. Simplicity of innovation was provided through the multidimensional analysis approach that mirrors faceted search on content. But the magic of Google-like search eluded us. Until now.

The development of freeform search for hard information marks a potential turning point for innovative exploration of data in business intelligence. End users can explore their business information in their own terminology without needing to understand the structure of the data or the rules by which data can be logically combined. The information context, constructed through hierarchical value decomposition when the data is initially mapped and loaded, makes business specific terminology and data relationships available directly to the search parser. This same largely automated initial process also provides end users with early and agile access to data, without the need for a lengthy prior development cycle.

In its initial implementation of the concept of freeform search, neutrinoBI offers a large, useful and usable subset of the potential full functional scope. The ability to simply enter business questions in common parlance and receive back ranked, real-time, valid answers should certainly tempt jaded users to experiment with the approach. And sceptical BI teams should be impressed with the agility with which the project can go live. NeutrinoBI also have some interesting plans for additional metadata, intelligent search term completion and true collaborative function in the next release.

The time has finally come for business users to google enterprise hard information!

Dr. Barry Devlin is among the foremost authorities on business insight and one of the founders of data warehousing. He is a widely respected consultant, lecturer and author of the seminal book, "Data Warehouse—from Architecture to Implementation." Barry's current interest extends to a fully-integrated business, covering informational, operational and collaborative environments to offer an holistic experience of the business through IT. He is founder and principal of 9sight Consulting, specializing in the human, organizational, and IT implications, and design of deep business insight solutions.

About NeutrinoBI

NeutrinoBI offers a breakthrough in self-service BI search with the first freeform keyword search-tool that works in a similar way to an internet search engine, finding and unifying data from a wide range of data sources to enable faster decision-making for agile businesses.

With high performance in-memory analytics for blisteringly fast search, neutrinoBI generates fresh results in minutes. It detects and records real-world relationships, rather than row and column based associations, and is an adaptive search-tool, enabling users to make the most of data exploration. And it does all of this without the need for programming and development.

Architected to enhance existing BI investments, neutrinoBI technology sits on top of existing BI infrastructures where it's configured to harness multiple structured data sources to the power of three - enterprise data-warehouses, local databases, and personal or web-based data including excel files: it's BI at its agile best..

Visit www.neutrinobi.com
Neutrino Concepts Ltd,
1 Devon Way, Longbridge Technology Park,
Birmingham, B31 2TS, United Kingdom

[1] The catch-phrase of Monty Python's Flying Circus, BBC TV, 1969-74

[2] Hayward, J., "*Question*", The Moody Blues, 1970

[3] I'm using the word *business* throughout in the broadest sense to include all organizations with objectives and systems to plan and review progress, including governmental institutions, non-profit bodies and more.

[4] Devlin, B. A. and Murphy, P. T., "*An architecture for a business and information system,*" IBM Systems Journal, Volume 27, Number 1, Page 60 (1988), http://bit.ly/EBIS88SJ

[5] In common IT parlance, the terms "structured" and "unstructured information" are used. The latter is an oxymoron, because information, by definition, has structure and without it would simply be noise!

[6] As computed (incorrectly) by Arthur Dent, last survivor of the biggest supercomputer ever built, Earth. Adams, D., "*The Restaurant at the End of the Universe*", Pan Macmillan, 1980

[7] Codd, E. F. et al, "*Providing OLAP to User-Analysts: An IT Mandate*" E. F. Codd & Associates, 1993

[8] Hearst, M. A., "*Design Recommendations for Hierarchical Faceted Search Interfaces*", ACM SIGIR Workshop on Faceted Search, 2006

[9] Devlin, B., "*Beyond the Data Warehouse: A Unified Information Store for Data and Content*", May 2010, http://bit.ly/uis_white_paper

[10] Devlin, B., "*Collaborative Analytics—Sharing and Harvesting Analytic Insights across the Business*", June 2009, http://bit.ly/otVWP

[11] As computed by the supercomputer "Deep Thought" after 7½ million years of processing.
Adams, D., "*The Hitchhiker's Guide to the Galaxy*", Pan Books, 1979