

Ballmer talk

Microsoft's Chief Executive Steve Ballmer has expressed regret over the late start in the search engine business. This statement comes after Bing managed a mere 8 per cent market share.

Long sceptical of Tim Berners-Lee's vision of the semantic web, Digit's Edward Henning met up recently in London with Gianpiero Lotito, co-founder of the Italian company Facility, and discovered that there are other ways to organise and retrieve the information out there on the web. And they're coming to a PC near you soon.

Facility:

The next evolutionary step in search?

**Edward Henning**

edward.henning@thinkdigit.com

One of the interesting, and perhaps key, points about the history of the Facility search engine is that the project grew out of a publishing background. One of the two founders of Facility, Gianpiero Lotito has worked in the publishing sector for many years, transferring the processes there over to digital systems.

He told me "We began to work with desktop publishing in 1987, with Aldus Pagemaker on the Mac, together with Adobe Illustrator. We were one of the first companies in Italy to work with DTP, in Milan, and soon, after about a year or year and a half, we started to work with big publishers. In this way we began to work with many big projects, which were being transferred from analogue processes to digital processes. To help them make this passage, we began to learn the core of the processes, we began to learn and understand the secrets of the publishing workflow and processes. In about three years we had learned to work with all the modern digital technologies: multi-media, CD-ROM, internet sites and so on.

"Through ten to fifteen years, we came to learn these processes for books, magazines, archives, photo-guides, CD-ROMs and so on, and we translated all these processes into an idea: the next generation search that is not only able to search and to find information, but is also able to organise it and to manage it. The problem for the next generation of search engine is to help people to organise information. I agree with your earlier comment that the semantic web is a chimera. Also, if it were possible to tag everything on the web (as suggested, for example, by Tim Berners-Lee), if this is not coordinated, if you have no standard, then you lose the advantage of tagging. On the other hand, if you have purely algorithmic engines, you cannot solve many search problems."

The work on developing a next-generation search engine started in 2001 – so what is wrong with existing search engines? Lotito explained to me the difference, as he sees it, between the approach of the engineer, typified, presumably by Google, and the approach of the publisher. The engineer's approach would consider the perfect map of London to be on a scale of 1:1 – it would be the same size as London, and would contain every single detail. But this is not practical – you cannot put it in your pocket. The job of the publisher is to find that balance; to make the map small enough to fit in your

WolframAlpha, anyone?

Is WolframAlpha really a search engine? It does not search the web for answers, but its own database of information, and computes answers from its data.

4,609,984

The number of web sites the open directory project currently catalogues and organises into over 5,90,000 categories.

Feature

pocket, but sufficiently detailed that you can find the places that you need.

The engineer's approach to information produces too much unnecessary data when a search is performed. Data-management experience is needed to filter the data appropriately. This is the publishing approach. Information is not normally the domain of engineers, and yet it is they that have been designing search engines up until now.

The implied point is this: does having several tens of thousands of search results which you will never look at really help you answer whatever question you are posing or problem you are trying to solve? Of course not, it is just a big number that might well be impressive, but where in that haystack is the one needle that you need?

Early search engines were blind in their search methods, and it is for this reason that users needed to develop skill in entering their search in the most effective way to narrow down the results and focus in on the information they were trying to find.

Lotito gave me the example of a warehouse. If you simply fill a warehouse with boxes, then when something is needed, you will have to go through and search in all of them to find the one that is needed. But if when filling the place up with boxes in the first place you label and tag them all properly, when one is needed, you can go straight to it. It is the same difference with the information on the internet – access to this information needs to be



Facility is more flexible in its approach while searching

any way that they want. He told me “This would be very confusing. If you do not regulate this, or you do not have technology that is able to recognise when words and phrases are semantically the same.

“When you tag information, the context cannot be constructed automatically. For example, sometimes the term environment

the semantic confusion in the Ballmer example just given. In a dictionary, you might find the words “fly”, “flying” and “flown” all grouped under the lemma “fly (v)”, the headword or keyword for that semantic group.

Ontological search is similar to some of the principles used in Wolfram |Alpha,

Does having several tens of thousands of search results which you will never look at really help you answer whatever question you are posing or problem you are trying to solve? Of course, it is just a big number that might well be impressive, but where in that haystack is the one needle that you need?

structured properly.

So, what is the approach offered by Facility? It is not a blind crawler-based search like Google, and it is not a limited computational engine such as Wolfram |Alpha. It can in fact use methods similar to both of these, and perhaps in that flexibility lies one of its strengths.

During our meeting, to stress the importance of the semantic approach in search, Lotito ran three Google searches, on “Steve Ballmer”, “Steven A. Ballmer” and “Steven Anthony Ballmer”. All came up with different results, and yet all refer to the current president of Microsoft. Lotito's point was that all three are semantically equivalent, and for the search to be truly meaningful, this should be understood by the engine, so that all three searches produce the same set of results.

He does not like the idea that we should simply leave people to tag information in

might refer to technology and the Windows environment; sometimes to ecology, and other times to a situation. Tagging words in this way has a level of precision that is very low. There are two ways to solve this problem: humans will tag information, or you will have an engine that will help humans to tag information. But I don't think it is possible to have a completely automatic system.”

The point was made that if you had software that could accurately tag all textual information, then the tagging would no longer be needed, as the software is effectively understanding the main meanings of the words and their contexts. We are probably many decades away from this possibility, and so some kind of human assisted tagging is needed.

The approach taken by Facility involves tagging, but also the use of lemmas and ontological search. Lemmas help to alleviate

where an answer, or result, can be computed from existing and well-structured data. For example, you may have two simple taxonomies: a list of cities in England, and a list of European cities that are capitals of their countries. If you ask the question “Which is the capital city of England?” the result can easily be computed from these two lists, as London is on both lists. If you ask for the capital of England, you are asking an ontological question, because the ontology is the connection between any two taxonomies. You are asking for the intersect between these two: London is a city in England; London is a capital city; therefore London is the answer to the question “Which is the capital of England?”

The Facility engine does not avoid the algorithmic approach – it can go out there and crawl websites with all the others. It combines these different methods in a flexible manner.

Semantic web

The concept of adding intelligent structure to the content of the web as first proposed by Tim Berners-Lee in the 90s. It never took off in the way he proposed.

According to Lotito, his methods can bring greater precision to the accuracy of search results. He claimed that the accuracy of results from purely algorithmic engines is of the order of 30 per cent – in other words, for every meaningful hit, you get three or four results that are just data noise. He claimed that with better automated organisation of the information that the precision could become around 70 per cent, but with some human intervention to validate and organise the data, you can come close to 100 per cent precision. Of course, this will not be possible every time.

But it will be possible when you have closed data sets of information, and if you begin to put many closed data sets with the same kind of organisation, you can get better results. So this is highly applicable to intranets, social networks – any closed set of data.

For this reason, the first uses of this system will not be a publicly accessible general search engine – although that will come – but the engine embedded in some system such as a social network, university web site, government web site, scientific research database, linked network of sites with a common theme, and so on. There are many possibilities.

Lotito agreed with my comment that Facility sounded like an advanced engine for LASEs (limited area search engines), but added that these could be combined together, eventually covering most of the web. I got the strong impression that he wants to set a standard here.

He made the point that “The problem



Gianpiero Lotito and Mariuccia Teroni, the two founders of Facility

for all future search engines is to be able to structure the information that you need to search.” It is impossible to provide advanced services in search without structured information, or, structured access to that information. And there lies the difference in his approach to that usually suggested.

It is important to point out that there is no intention here that web sites should themselves be changed in order for their information to be tagged and structured – the methods of Facility lie in creating a structured approach to information on the web. Part of that structuring process lies in assessing the reliability of sites, and this can only be done by human intervention. This also changes the ranking of sites – engines such as Google rank mainly on the basis of the number of links to various sites, but this is completely blind to issues such as reliability.

Facility employs a pre-filtering engine which reduces the number of potential sites, and then the experts judge those that are the most reliable from the

results. Their starting point is the ODP taxonomy of the internet. The Open Directory Project is an ongoing attempt at cataloguing the contents of the web, and claims to be “the most comprehensive human-edited directory of the Web, compiled by a vast global community of volunteer editors.”

Going down a level of detail, there is also built into the system an ability to catalogue structured sites automatically. This is called the Automatic Taxonomy Extractor, and would be useful for such things as blogs, databases, perhaps news sites, and so on. Sites where there is a regular format to the information. The engine will produce for you an automatic taxonomy, which you can then go in and edit – discarding mistaken headings and tags, adding or combining others, and so forth. This is an example of how the human editing function is aided by software that performs the first part of the work.

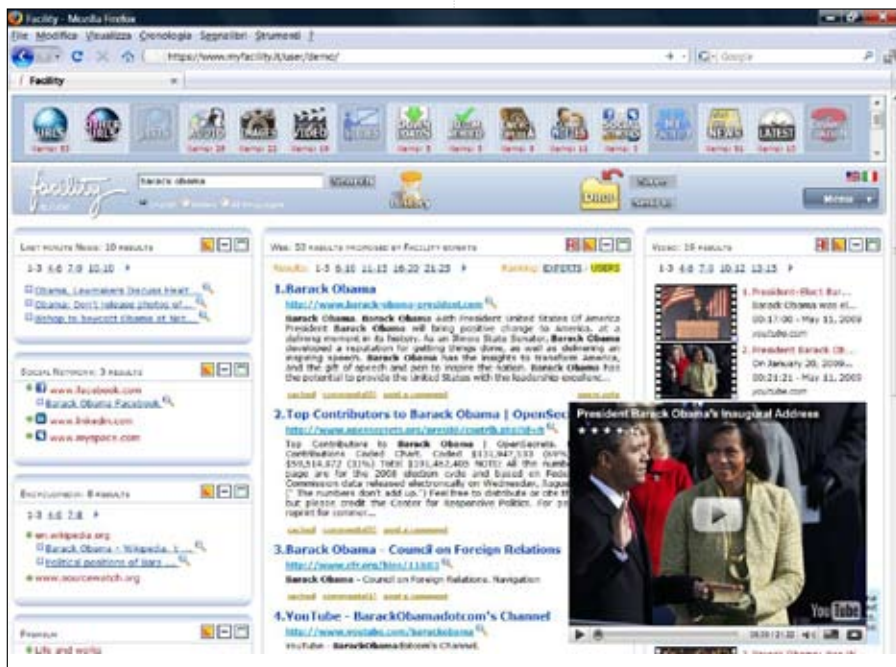
This can also work in two ways. If I am developing a search engine for a particular field, then the taxonomy that is developed in this way does not, of course, change the entries in the target site. But, if it is my own site that I am preparing, then I could very well use this tool as a way of tagging directly. This does more than just tag individual items – the overall collection, on say a blog, is structured because the taxonomy is developed and stored. Many people are currently tagging posts on such sites, but without having an overall scheme. This system can propose such a scheme for you, which you can then edit and improve. It will also read those existing tags if they are there.

So, with Facility you have a two-way connection between the engine and the contents of the web. Of course, this is not the semantic web of Tim Berners-Lee that grows from the bottom up. Instead, this is a pragmatic approach, because the tagging of all information on the web in existing web pages is simply not possible. It will never happen.

For Lotito, the display of information in a structured manner is as important as the way in which the results are obtained. The interface is complex at first sight, but easy to use.

The interface is entirely based on Web 2.0 and all the technologies used are open source – the engine can be embedded in other internet systems without licence problems.

The display screen is divided into several sections – Lotito calls them widgets, perhaps not the best term – which include the different formats of information, sources and media that have been identified by a search. These are tiled panels on the screen, in three



Facility structures, organizes access to information being searched and then displays the results as the widgets seen here

Search engine rankings

According to research firm comScore, Google is the leader in the US market with 9.3 bn searches, followed by Yahoo! at 2.9 bn and Microsoft at 1.2 bn searches.

Google reigns

With the near apparent failure of three attempts – Cuil, Trooker and Exlead as Google alternatives, Page and Brin have reasons to cheer.

Feature

columns, the top one in the middle being the most important: this main display panel is called simply “Web” and tells you the number of results that are proposed by Facility experts. In other words, the results from Facility’s editorialised map of reliable sites.

The other widgets need not necessarily have been through the same filter, and may very well be the result of “normal” algorithmic search methods; they display other categories of results. The various categories include: *Last minute News* for any relevant items on news sites; *Social Network* – possibly relevant comments found on social networking sites; *Encyclopedia* – for hits in Wikipedia and similar sites; *Downloads* – these are typically PDF files; *Presentation* – any presentations, typically education or perhaps business oriented; *Video, Images and Audio*, all self-explanatory; and, *Disambiguation* – this is to display a variety of possibilities, such as in the search shown, on Leonardo da Vinci. You may well have been looking for information on the man himself, or a hotel that uses his name, or, of course, a certain best-selling book by Dan Brown. This widget displays such possible ambiguities.

An interesting point about the display is that the widgets give you, where possible, a

preview of the contents of the site. You do not have to click and open a new window or tab; the widget itself will expand and display the movie or whatever format the information might be in. You can, if you wish, also save the content. In this case you are saving the content itself for later viewing, not just saving a bookmark or link to the item. **It is clear here where Facility’s background in publishing comes into play – this would be a suitable interface for any kind of multi-media based set of information.**

The display is also customisable. The user can move the widgets around according to their relative importance. If you were doing a search on the use of the tobacco mosaic virus in genetics, then the News widget should hardly be at the top, but it certainly should be prominent if the search were on the US president, as in the example screenshot that is with this article.

You, as an end user, can add comments to search results, so that if another user makes the same search, they will be able to see your comments. How this will work, if at all, when large numbers of people use the system remains to be seen.

So, Facility does not only structure and organise the access to the information that is being searched, but it also organises

in a customisable way the display of the results. Potentially very useful is the fact that when you save a result – a video, PDF, etc. – the search request that produced that result is also saved; this way the context is retrievable. This should help avoid problems such as seeing a saved bookmark and not remembering the particular line of thought that led to it being saved. Every item that might result from a search can be managed in this way. The exception to this is anything that is copyright protected. This is detected automatically and the item cannot be saved.

The idea is to give users maximum flexibility to manage information. They do not have to keep leaving the search system to view information – it is brought into the Facility system instead. Of course, this has benefits for advertisers, who like it if the user is not always shooting off to some other site.

Over the next few months the final patents for this system will be settled, and the engine will start appearing in certain sites such as social network sites. The frequency with which Gianpiero Lotito mentioned such sites suggests that at least one must be high on his list of prospective clients. The general purpose public search engine will follow later. **d**

WWW.ZOTAC.COM

ZOTAC
It's Time to Play

ION

PhysX NVIDIA



ZOTAC ION GTX295
THE ULTIMATE GAMING SOLUTION

- GEFORCE GTX 295
- 1792MB GDDR3 MEMORY
- DIRECTX 10, OPEN GL 3.0



ZOTAC ION GTX275
THE ULTIMATE PERFORMANCE VALUE

- GEFORCE GTX 275
- 896MB GDDR3 MEMORY
- DIRECTX 10, OPEN GL 3.0

	ZOTAC ION ITX-A SERIES <ul style="list-style-type: none">• Intel® Atom™ N330 Dual-Core CPU• NVIDIA® ION™ Graphics Processor• DVI / HDMI / VGA• Integrated WLAN• DirectX 10• External power adapter• Mini-ITX
	ZOTAC ION ITX-B SERIES <ul style="list-style-type: none">• Intel® Atom™ N230 Single-Core CPU• NVIDIA® ION™ Graphics Processor• DVI / HDMI / VGA• Gbit LAN• DirectX 10• Mini-ITX

ZOTAC Motherboards Distributed by:

ZEBRONICS
Always Ahead
ISO 9001:2000 Certified Company

Top Notch Infotronics (I) Pvt. Ltd.
Tel : 044-43936000
enquiry@zebronic.info

ZOTAC GeForce Graphics Plus Cards Distributed by:

ADITYA

Aditya Infotech Ltd.
Tel : 011-46665666
sales@adityagroup.com

NVIDIA
ION

NVIDIA
GEFORCE
CUDA