



**SolcaraSolSearch**  
Connecting information  
with integrated search

---

# The semantic web and federated searching

## Whitepaper



---

**Andrew Maisey**  
Product Development Manager

**Ian Pallen**  
Senior Developer



**SolcaraSolSearch**  
Connecting information  
with integrated search

---

## Contents

<b>Executive Summary</b>	<b>3</b>
<b>The Semantic Web</b>	<b>3</b>
<b>Traditional Approaches to Enterprise Search</b>	<b>4</b>
<b>One Stop Search requires Federated Searching</b>	<b>6</b>
<b>Business Benefits of Solcara SolSearch v3.0</b>	<b>7</b>
<b>Conclusion</b>	<b>8</b>
<b>Glossary</b>	<b>8</b>
<b>Contact us</b>	<b>9</b>



## Executive Summary

The semantic web provides early adopters with the potential for huge improvements in techniques for knowledge discovery and retrieval. Traditional approaches to information retrieval inherently restrict users to what are perceived to be an organisation's high-value content sources only and require a high level of domain level expertise and experience of using the chosen search vendor's product. Couple this with the inherent limitations of traditional indexing techniques explained in this white paper, users wishing to exploit semantic searching must look elsewhere for solutions.

Solcara SolSearch is a federated search product that overcomes the shortfalls of traditional enterprise search while providing an open platform that facilitates knowledge discovery through semantic techniques.

---

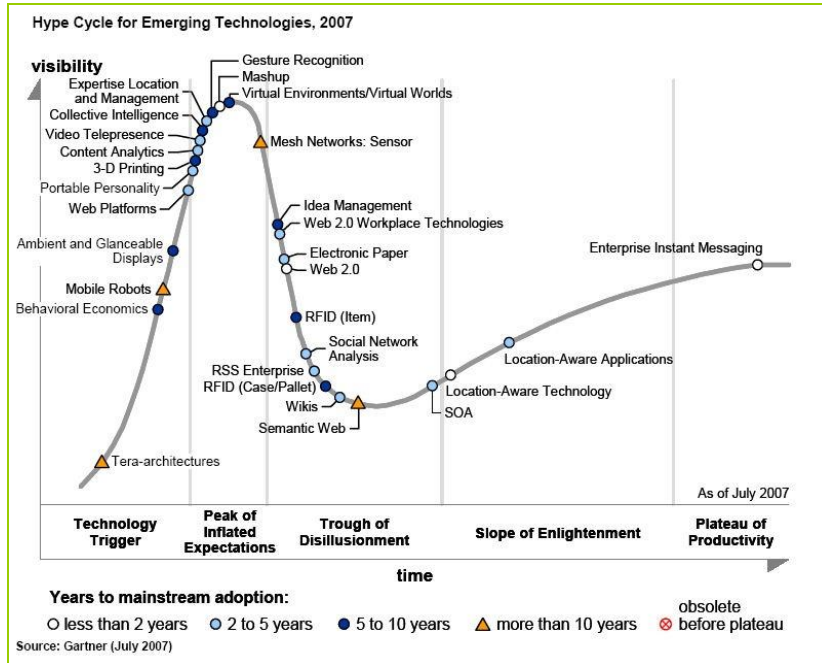
## The Semantic Web

The Semantic Web is an extension of today's World Wide Web in which the meaning of information and services is defined, making it possible for the web to understand and satisfy the requests of people and machines to use that content. Today there is little opportunity for information exchange because by its very nature the web is uncontrolled and the content it serves is created using a plethora of applications adhering to a wide range of standards, or none. The vision for the Semantic Web comes from Tim Berners-Lee's view of the web as a universal medium for data, information, and knowledge exchange.

At its core, the semantic web comprises a set of design principles, collaborative working groups, and a variety of enabling technologies. Some elements of the semantic web are expressed as prospective future possibilities that have yet to be implemented or realised. Other elements of the semantic web are expressed in formal specifications. Some of these include Resource Description Framework (RDF), a variety of data interchange formats including triples, and notations such as the Web Ontology Language (OWL), all of which are intended to provide a formal description of concepts, terms, and relationships within a given knowledge domain.

The possibilities of this vision are limitless. If the meaning of web content can be expressed in a consistent manner, new relationships between previously isolated content can be derived. The exploration of these relationships will be performed automatically by services and applications that reason and execute on behalf of the initiator of the request.

The semantic web is evolving, but the benefits have a way to go before they reach end users as shown in figure 1; Gartner's emerging technologies hype cycle.



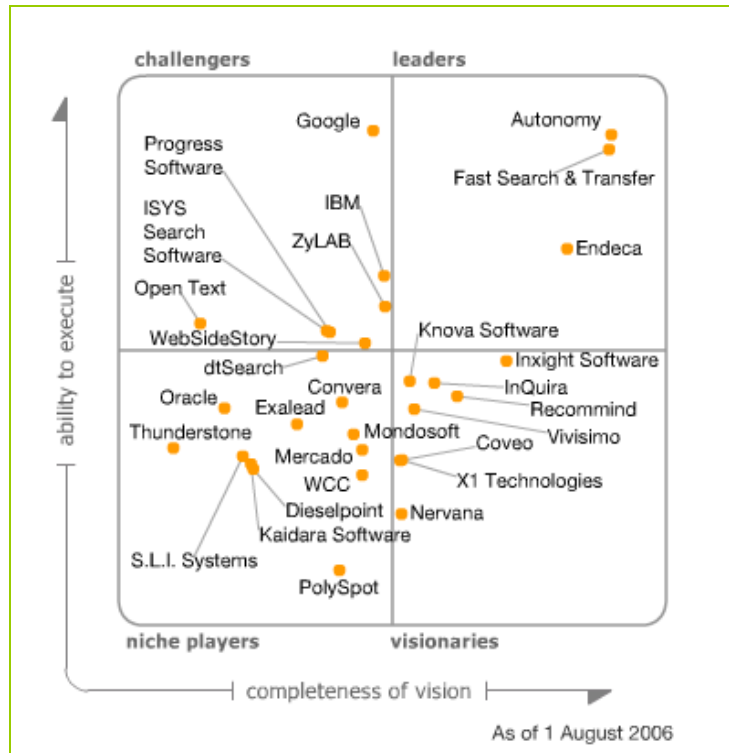
**Figure 1: Gartner 2007 Emerging Technologies Hype Cycle**

This paints a realistic view of the current state of the semantic web, suggesting that the amount of research work needed to bring this technology to the mainstream was underestimated; nevertheless the importance of the semantic web is still high, even if today it has failed to meet media expectations.

## Traditional Approaches to Enterprise Search

The importance of information retrieval has been appreciated for many years and with the increasing capacity and decreasing cost of online storage media, this technology has reached the point of being an inherent part of most software applications. Search in its many guises exists within a user's e-mail client, file-store or intranet, Document Management System (DMS), subscription web site and relational database application used for Customer Relationship Management (CRM), financial management and a host of other applications. Users expect to be able to retrieve information from these products in isolation and the provision of search is taken as a basic function.

For the enterprise there is an overall requirement to be able to search all of these disparate content sources in a single search; what is often referred to as *one stop search*. Figure 2 shows the number of software vendors who provide solutions to enterprise search.



**Figure 2: Gartner Magic Quadrant for Information Access Technology (Source: Gartner 2006)**

The traditional approach adopted by these vendors to providing one stop search is to index the content from predefined content sources that are perceived to be of high value to the enterprise. These typically include corporate e-mail, intranet and file servers and one or more critical content stores or DMS. Individual software components are developed, supported and sold by each vendor that interfaces to each of the content sources required.

Each component is written to understand the data structure and Application Programming Interface (API) of the chosen content source and provide the indexing product with access to the data to be indexed. In this way indexing *spiders* can be configured to revisit each content source at a predefined indexing interval and re-index the content in a central index. It is this index that users search when conducting a one stop search.

Google represents an Internet equivalent of this enterprise search functionality. Huge numbers of software spiders traverse the Internet's HTML sites and extract the words from their pages into a centralised index. Users search Google's index and retrieve links to the content's location. When a user selects a link to view they are shown the content in situ on the original web site. Google provides for many the one stop search of the Internet.

However, there are some serious limitations to this traditional approach to providing a one stop search that include:

- ▶ Indexing latency – the content in the centralised index is only as current as the last time the content was traversed by the indexing spider;



**SolcaraSolSearch**  
Connecting information  
with integrated search

- ▶ Hardware and software expenditure – storage may be cheap but the cost of implementing enterprise search includes primary and failover hardware, network bandwidth to support document parsing and display, support and systems' administration overhead and ongoing licence fees for the indexing technology and its associated spiders.
- ▶ Inability to index content – any content that cannot be accessed by the spiders will forever remain hidden from a user's one stop search. This includes content linked to through JavaScript and all database driven content found as a result of a search.

As traditional indexing technologies have to fix their connections to enterprise's chosen content sources, their ability to harness the potential of the semantic web's interoperability is inherently compromised.

What is needed is a new approach to one stop search.

---

## One Stop Search Requires Federated Searching

Federated searching differs from traditional enterprise searching in a number of key areas:

- ▶ Federated searching uses the content source's existing index to satisfy the search request – no separate, out of date, expensive index is required. Search results are completely up to date and are delivered in real-time
- ▶ Federated searching can be deployed on minimal hardware and requires no enhancements to current networking infrastructure
- ▶ Federated search inherently connects users to database driven content as the search is conducted in real-time

Federated searches can be conducted against any indexed content source that is able to receive a browser-based search request – as opposed to only working against a predefined list of content sources or software applications.

By offering a one stop search solution that is faster to deploy, is more cost effective and is open and extensible to new content sources, federated searching is the best solution both within the enterprise and for all Internet based content sources.

Furthermore federated searching's open architecture is best placed to exploit the potential of the semantic web.

Solcara SolSearch is a federated search product developed and maintained by Solcara, a UK software company with an impressive list of corporate customers.

Solcara SolSearch has been delivering the business benefits of federated searching to Solcara customers for over five years and over that period has been developed and enhanced to its current high performance and scalability .NET architecture.

The next version of SolSearch is being built to deliver semantic search capability to Solcara's customers.



---

## Business Benefits of Solcara SolSearch v3.0

The development of Solcara SolSearch v3.0 will bring the benefits of the semantic web to the enterprise.

One of the key issues for any organisation in adopting semantic technologies will be how to map their bespoke content to the formal description of concepts, terms, and relationships within their given knowledge domain. Very few enterprises would ever have the time or money to invest in a global data mark-up project; still less would ever be able to maintain this work going forwards. Providing semantic web interoperability through an analysis and cleansing of all an organisation's knowledge assets is therefore clearly unfeasible. Faced with the sheer scale of Internet based content that the organisation relies on but over which it has no control, the possibility of bringing semantic web technologies to the enterprise appear impossible; this is reflected in the market analysis shown in Figure 1.

Simple word or phrase searches are the foundation of traditional enterprise and Internet based retrieval products. They deliver high precision results; though often at the expense of recall. Advanced search functionality including synonym expansion, thesaurus look-up, wildcarding and natural language expansion all help to broaden the user's initial query and improve recall but at the expense of precision. The issue with current search technologies is that the user must understand what they are looking for and where to find it or be prepared to navigate a large list of potentially poor results.

Federated searching delivered by Solcara SolSearch enables customers to simultaneously search any number of indexed content sources, irrespective of location, passing the user's search term(s) to those sources for processing in their own query language.

Solcara SolSearch v3.0 will build on the business benefits of federated searching by delivering an additional semantic search capability.

Solcara SolSearch uses a *connector* to provide the link between the product's multithreaded search engine and the content source. This connector architecture already understands the meaning of the content that can be accessed. In customer engagements Solcara work with knowledge workers to target the connectors to access knowledge that is well defined and able to be published in a semantically reusable fashion.

Solcara SolSearch will enhance each content asset found as a result of the initial search with this connector based knowledge – on the fly; negating the need for complicated and expensive data cleansing or mark-up, or the need to change working practices. As the results set will be marked up in RDF format, semantically aware services and products will be able to work with the knowledge and enhance the user experience.

Solcara SolSearch v3.0 therefore enables an organisation to be able to share its legacy content with semantically aware products.

Within Solcara SolSearch this RDF enhanced search metadata will be evaluated against existing domain expertise and new data relationships discovered so that secondary searches can be conducted on behalf of the requesting user. Solcara SolSearch's asynchronous multithreaded architecture facilitates the kind of recurring queries knowledge mining requires. This discovery process goes far beyond the traditional enterprise search expansion tools mentioned above. SolSearch's semantic search capability will allow an initial user search submitted against one set of content sources to generate searches for new types of related knowledge against another set of automatically selected content sources.

This approach will dramatically improve the user's ability to find relevant information without the need for a thorough understanding of the domain they are researching. The user will experience huge increases in recall without the traditionally associated decrease in precision.



**SolcaraSolSearch**  
Connecting information  
with integrated search

Users of Solcara SolSearch v3.0 will be able to enjoy the enhanced benefits of finding information regardless of format or location or knowledge area. High quality search results will no longer require the user to understand the underlying content sources or how they are structured; users will be able to concentrate on what to search, not where.

---

## Conclusion

Organisations that wish to provide a true one stop search to all their internal and Internet based indexed content sources can today deploy Solcara SolSearch in a fraction of the time and at a fraction of the cost of traditional enterprise search solutions. At the same time they will have overcome the pitfalls of this resource hungry and inherently out of date technology.

With Solcara SolSearch v3.0 organisations will additionally be able to enjoy the benefits of sharing their current content with the growing range of semantically aware software products and services and enjoy an enhanced knowledge discovery experience that will supersede all previous expectations of what a simple search on a corporate intranet will retrieve.

---

## Glossary

### Ontology

In both computer science and information science, an ontology is a representation of a set of concepts within a domain and the relationships between those concepts. It is used to reason about the properties of that domain, and may be used to define the domain.

Ontologies are used in artificial intelligence, the Semantic Web, software engineering, biomedical informatics, library science, and information architecture as a form of knowledge representation about the world or some part of it.

### OWL

The Web Ontology Language (OWL) is a family of knowledge representation languages for authoring ontologies, and is endorsed by the World Wide Web Consortium. OWL is considered one of the fundamental technologies underpinning the Semantic Web, and has attracted both academic and commercial interest.

### RDF

Resource Description Framework (RDF) is a family of World Wide Web Consortium (W3C) specifications originally designed as a metadata model but which has come to be used as a general method of modelling information, through a variety of syntax formats.

The RDF metadata model is based upon the idea of making statements about resources in the form of subject-predicate-object expressions, called triples in RDF terminology. The subject denotes the resource, and the predicate denotes traits or aspects of the resource and expresses a relationship between the subject and the object. For example, one way to represent the notion "The sky has the colour blue" in RDF is as the triple: a subject denoting "the sky", a predicate denoting "has the colour", and an object denoting "blue".



**SolcaraSolSearch**  
Connecting information  
with integrated search

---

## Contact Us

### Solcara

The Long Room  
Coppermill Lock  
Harefield  
Middlesex UB9 6JA  
United Kingdom

### Telephone

+44 (0) 1895 820 950

### Web

[www.solcarasolsearch.com](http://www.solcarasolsearch.com)

### Email

[info@solcara.com](mailto:info@solcara.com)