



Data, Information and Process Integration  
with Semantic Web Services

**DIP**

*Data, Information and Process Integration with Semantic Web Services*

**FP6 - 507483**

Deliverable

**WP 12: Market Observation**

**D12.4**

**White Paper on Roadmaps and Potential Use Cases for  
SWS (Search with SWS)**

Joachim Quantz (Berlecon Research)

Thorsten Wichmann (Berlecon Research)

July 6<sup>th</sup>, 2005



---

## EXECUTIVE SUMMARY

The application area search has its root in information retrieval. With the advent of the World Wide Web, search has become a standard everyday functionality for almost everyone. Google has been dominating the market of Internet search engines for some years now. But recently, the market, which has been static for some time, has become dynamic again and several new and established search engines are offering promising alternatives to Google and Yahoo.

Search is not only limited to Internet search and the importance of desktop search and enterprise is constantly increasing. Internet search engines such as Google and Yahoo are trying to push their solutions into these markets, but there are already many established vendors offering sophisticated solutions, especially for enterprise search.

Techniques for content extraction and categorization have matured also and are increasingly integrated into search technology, e.g. to categorize search results on the fly and allow the clustering of results.

Finally, many Internet search engines combine search with location information or offer specialized search functionality for specific formats or contents (news, images, blogs, TV).

There are basically two application scenarios for using SWS in search. On the one hand, special search engines emerge for searching Semantic Web content. On the other hand, Semantic Web Services could be used to provide functionality such as information extraction or categorization to search engines. It would also be possible to provide an intelligent search service as an SWS, performing pre-processing of queries and post-processing of results in order to increase the search quality.

In the larger sense the deliverable contributes to the DIP goal of exploitable tools, as it provides information on application areas in which tools could be successfully exploited.

The deliverable should be read by anyone involved in exploitation of DIP results and in strategic planning of future work to be performed in DIP.

Disclaimer: The DIP Consortium is proprietary. There is no warranty for the accuracy or completeness of the information, text, graphics, links or other items contained within this material. This document represents the common view of the consortium and does not necessarily reflect the view of the individual partners.

## Document Information

<b>IST Project Number</b>	FP6 – 507483	<b>Acronym</b>	DIP
<b>Full title</b>	Data, Information, and Process Integration with Semantic Web Services		
<b>Project URL</b>	<a href="http://dip.semanticweb.org">http://dip.semanticweb.org</a>		
<b>Document URL</b>			
<b>EU Project officer</b>	Kai Tullius		



<b>Deliverable</b>	<b>Number</b>	12.4	<b>Title</b>	White Paper on Roadmaps and Potential Use Cases for SWS (Search with SWS)
<b>Work package</b>	<b>Number</b>	12	<b>Title</b>	Market Observation

<b>Date of delivery</b>	<b>Contractual</b>	M 18	<b>Actual</b>	July 2005
<b>Status</b>	V1.0		Final Version	
<b>Nature</b>	Prototype <input type="checkbox"/> Report <input checked="" type="checkbox"/> Dissemination <input type="checkbox"/> Ontology <input type="checkbox"/>			
<b>Dissemination Level</b>	Public <input type="checkbox"/> Consortium <input checked="" type="checkbox"/>			




<b>Authors (Partner)</b>	Joachim Quantz (Berlecon), Thorsten Wichmann Berlecon			
<b>Responsible Author</b>	Joachim Quantz		<b>Email</b>	jq@berlecon.de
	<b>Partner</b>	Berlecon	<b>Phone</b>	+49 30 28 52 96 0

<b>Abstract (for dissemination)</b>	This deliverable analyzes the potential of Semantic Web Services in the application area of search. It contains brief descriptions of the main aspects of search. Then, existing solutions and current trends are presented in some detail, focusing on Internet search engines, RSS/blog search, desktop search, enterprise search, and taxonomies, categorization, and information discovery. Finally, the potential of SWS in the context of search is discussed. The deliverable ends with recommendations for the DIP project.	
<b>Keywords</b>	Search, Semantic Web Services	

## Project Consortium Information

Partner	Acronym	Contact
National University of Ireland Galway	NUIG  National University of Ireland, Galway <i>Ollscoil na hÉirann, Gaillimh</i>	Prof. Dr. Christoph Bussler Digital Enterprise Research Institute (DERI) National University of Ireland, Galway Galway Ireland Email: <a href="mailto:chris.bussler@deri.org">chris.bussler@deri.org</a> Tel: +353 91 512460
Fundacion De La Innovacion.Bankinter	Bankinter  .com	Monica Martinez Montes Fundacion de la Innovacion. BankInter Paseo Castellana, 29 28046 Madrid, Spain Email: <a href="mailto:mmtnez@bankinter.es">mmtnez@bankinter.es</a> Tel: 916234238
Berlecon Research GmbH	Berlecon  information technology economics	Dr. Thorsten Wichmann Berlecon Research GmbH Oranienburger Str. 32 10117 Berlin, Germany Email: <a href="mailto:tw@berlecon.de">tw@berlecon.de</a> Tel: +49 30 2852960
British Telecommunications Plc.	BT 	Dr John Davies BT Exact (Orion Floor 5 pp12) Adastral Park Martlesham Ipswich IP5 3RE, United Kingdom Email: <a href="mailto:john.nj.davies@bt.com">john.nj.davies@bt.com</a> Tel: +44 1473 609583
Swiss Federal Institute of Technology, Lausanne	EPFL 	Prof. Karl Aberer Distributed Information Systems Laboratory École Polytechnique Fédérale de Lausanne Bât. PSE-A 1015 Lausanne, Switzerland Email : <a href="mailto:Karl.Aberer@epfl.ch">Karl.Aberer@epfl.ch</a> Tel: +41 21 693 4679
Essex County Council	Essex  Essex County Council	Mary Rowlett, Essex County Council PO Box 11, County Hall, Duke Street Chelmsford, Essex, CM1 1LX United Kingdom. Email: <a href="mailto:maryr@essexcc.gov.uk">maryr@essexcc.gov.uk</a> Tel: +44 (0)1245 436524
Forschungszentrum Informatik	FZI 	Andreas Abecker Forschungszentrum Informatik Haid-und-Neu Strasse 10-14 76131 Karlsruhe Germany Email: <a href="mailto:abecker@fzi.de">abecker@fzi.de</a> Tel: +49 721 9654 0

Partner	Acronym	Contact
Institut für Informatik, Leopold-Franzens Universität Innsbruck	UIBK 	Prof. Dieter Fensel Institute of computer science University of Innsbruck Technikerstr. 25 A-6020 Innsbruck, Austria Email: <a href="mailto:dieter.fensel@deri.org">dieter.fensel@deri.org</a> Tel: +43 512 5076485
ILOG SA	ILOG 	Christian de Sainte Marie 9 Rue de Verdun, 94253 Gentilly, France Email: <a href="mailto:csma@ilog.fr">csma@ilog.fr</a> Tel: +33 1 49082981
inubit AG	Inubit 	Torsten Schmale inubit AG Lützowstraße 105-106 D-10785 Berlin Germany Email: <a href="mailto:ts@inubit.com">ts@inubit.com</a> Tel: +49 30726112 0
Intelligent Software Components, S.A.	iSOCO 	Dr. V. Richard Benjamins, Director R&D Intelligent Software Components, S.A. Pedro de Valdivia 10 28006 Madrid, Spain Email: <a href="mailto:rbenjamins@isoco.com">rbenjamins@isoco.com</a> Tel. +34 913 349 797
NIWA WEB Solutions	NIWA 	Alexander Wahler NIWA WEB Solutions Niederacher & Wahler OEG Kirchengasse 13/1a A-1070 Wien Email: <a href="mailto:wahler@niwa.at">wahler@niwa.at</a> Tel:+43(0)1 3195843-11
The Open University	OU 	Dr. John Domingue Knowledge Media Institute The Open University, Walton Hall Milton Keynes, MK7 6AA United Kingdom Email: <a href="mailto:j.b.domingue@open.ac.uk">j.b.domingue@open.ac.uk</a> Tel.: +44 1908 655014
SAP AG	SAP 	Dr. Elmar Dorner SAP Research, CEC Karlsruhe SAP AG Vincenz-Priessnitz-Str. 1 76131 Karlsruhe, Germany Email: <a href="mailto:elmar.dorner@sap.com">elmar.dorner@sap.com</a> Tel: +49 721 6902 31

<p>Sirma AI Ltd.</p>	<p>Sirma</p>  <p><b>Ontotext</b> Knowledge and Language Engineering Lab of Sirma</p>	<p>Atanas Kiryakov, Ontotext Lab, - Sirma AI EAD Office Express IT Centre, 3rd Floor 135 Tzarigradsko Chausse Sofia 1784, Bulgaria Email: <a href="mailto:atanas.kiryakov@sirma.bg">atanas.kiryakov@sirma.bg</a> Tel.: +359 2 9768 303</p>
<p>Unicorn Solution Ltd.</p>	<p>Unicorn</p> 	<p>Jeff Eisenberg Unicorn Solutions Ltd, Malcha Technology Park 1 Jerusalem 96951 Israel Email: <a href="mailto:Jeff.Eisenberg@unicorn.com">Jeff.Eisenberg@unicorn.com</a> Tel.: +972 2 6491111</p>
<p>Vrije Universiteit Brussel</p>	<p>VUB</p>  <p>Vrije Universiteit Brussel</p>	<p>Pieter De Leenheer Starlab- VUB Vrije Universiteit Brussel Pleinlaan 2, G-10 1050 Brussel ,Belgium Email: <a href="mailto:Pieter.De.Leenheer@vub.ac.be">Pieter.De.Leenheer@vub.ac.be</a> Tel.: +32 (0) 2 629 3749</p>

## TABLE OF CONTENTS

<b>EXECUTIVE SUMMARY.....</b>	<b>1</b>
<b>TABLE OF CONTENTS.....</b>	<b>6</b>
<b>1 INTRODUCTION.....</b>	<b>8</b>
<b>2 ASPECTS OF SEARCH.....</b>	<b>8</b>
2.1 Internet Search Engines and Desktop/Enterprise Search .....	8
2.2 Taxonomies, Navigation, and Clustering .....	10
2.3 Information Retrieval, Recall and Precision.....	11
<b>3 EXISTING SOLUTIONS AND CURRENT TRENDS.....</b>	<b>11</b>
3.1 Internet Search Engines .....	12
3.1.1 Google and Yahoo .....	12
3.1.2 Alternatives to Google and Yahoo .....	13
3.1.3 Specialized Search.....	14
3.2 Blog/RSS Search .....	14
3.3 Desktop Search.....	16
3.3.1 Vendor Landscape .....	16
3.3.2 Security Issues .....	17
3.3.3 Expected Trends .....	18
3.4 Enterprise Search.....	19
3.4.1 Crawler Performance.....	19
3.4.2 Document Pre-Processing .....	20
3.4.3 Query Pre-Processing .....	20
3.4.4 Post-Processing of Results .....	20
3.4.5 Vendor Landscape .....	21
3.5 Taxonomies, Categorization, and Information Discovery.....	22
3.5.1 Using Taxonomies and Categorization in Search .....	22
3.5.2 Vendor Landscape .....	22
3.6 Conclusion and Expected Trends .....	23
<b>4 SWS POTENTIAL FOR SEARCH .....</b>	<b>24</b>
4.1 Searching Semantic Web Resources .....	24
4.1.1 Semantic Web Search Engines.....	24
4.1.2 Discovering Semantic Web Services.....	25
4.2 Using SWS in Search Engines .....	26
4.2.1 Information Extraction and Categorization .....	26
4.2.2 Intelligent Search as a Semantic Web Service .....	26
<b>5 RECOMMENDATIONS FOR DIP.....</b>	<b>27</b>

**6 REFERENCES ..... 28**

## 1 INTRODUCTION

Work Package 12 provides the DIP consortium and in particular the technology providers in DIP with real-world background information on potential application areas for Semantic Web Services (SWS). This work was started with Section 3 in Deliverable D12.1 [1], which briefly evaluated the potential of Semantic Web Services in the application areas of business process management, content syndication, contextual ads, enterprise application integration, enterprise collaboration, product information management, single European electronic market, search/mining, and social software. Based on these initial evaluations, Section 4 of D12.1 and D12.3 [2] contained an in-depth analysis of the potential of SWS in the application areas Business Process Management and Enterprise Collaboration, respectively.

This deliverable analyzes the potential of Semantic Web Services in the application area of search. It contains brief descriptions of the main aspects of search. Then, existing solutions and current trends are presented in some detail, focusing on Internet search engines, RSS/blog search, desktop search, enterprise search, and taxonomies, categorization, and information discovery. Finally, the potential of SWS in the context of search is discussed. The deliverable ends with recommendations for the DIP project.

The approach taken in this deliverable and in WP12 in general is market or application driven rather than technology driven. The analysis thus starts with an assessment of the currently available commercial solutions in the application area. Based on these results, the potential of SWS technology is evaluated in a second step, focussing on how SWS could be integrated into existing solutions in order to enhance these solutions, make them more efficient, or provide significant value-add with respect to costs or quality.

D12.1 distinguishes two variants of Semantic Web Services: “on the one hand, Web Services and their descriptions can be semantically enriched to enhance the potential for discovering and combining services. On the other hand, Web Services can be used to provide interfaces to existing Semantic Web technology, e.g. ontologies or logic-based reasoners” [1, p. 1]. The analysis in this deliverable takes into account both variants of SWS, although DIP focuses only on the former variant.

## 2 ASPECTS OF SEARCH

This section briefly introduces the main aspects of search starting with an overview over Internet search technology in Section 1.1. Section 1.2 then presents taxonomies, navigation and clustering and Section 1.3 discusses the notions of precision and recall as used in information retrieval.

### 2.1 Internet Search Engines and Desktop/Enterprise Search

With the advent of the World Wide Web, search has become a standard everyday functionality for almost everyone. Google has been dominating the market of Internet search engines for some years now. But recently, the market, which has been static for some time, has become dynamic again (see Section 2 for details on current market trends).

SearchEngineWatch<sup>1</sup>, the definitive resource for information related to Internet search engines, offers brief résumés of the most important search engines and their fate over the years.<sup>2</sup> Section 2 will present these search engines in more detail. Here we will focus on the general technology and functionality provided by such search engines.

Basically, all Internet search engines are based on the same approach:

- The user enters a query (depending on the search engine, a query can consist of simple keywords, complex query expressions, or questions in natural language).
- The search engine looks up the references matching the query (depending on the search engine, such references can be HTML pages, pdf documents, database entries, discussion threads in forums, images, etc.)
- The search engine presents the references in a result list (depending on the search engine, the results are accompanied by relevance values, additional information on the reference, category information, etc.).

In order to perform the second step in real-time, search engines build up an index of references by using so-called spiders, crawlers, or robots. These tools roam the Internet for new information sources (e.g. by following hyper links) and then add them to the index. The index basically maps keywords to all the resources in which they occur. Most search engines perform a pre-processing of resources before indexing is actually performed.

Obviously, this technology is not restricted to the World Wide Web, but can also be applied to search information sources on a single PC or in an enterprise network. Such functionality is usually called desktop search and enterprise search, respectively. Whereas desktop search is usually restricted to standard file types, e.g. Word documents, enterprise search is often integrated with content management solutions, may also use database content, and often has to deal with issues related to security. Section 2 will present existing solutions for desktop and enterprise search in some detail.

The presentation so far has not looked at the context in which a search occurs. It is, however, important to keep in mind that search can occur in very different scenarios. At least the following scenarios should be distinguished:

- Looking for a specific answer, e.g. searching for a specific document, whose title and directory location one cannot remember, searching for the current exchange rate between Yen and Euro, searching for the home page of a hotel or a company, or searching for someone's email address.
- Looking for information on a specific topic. In contrast to the first scenario not a single answer is expected but rather a potentially large list of results. This scenario can further be divided into two sub cases:
  - Looking for information in a field where one already has some expertise. Based on such expertise, search queries can be made more accurate and the relevance of results can be assessed more or less straightforwardly.

---

<sup>1</sup> [www.searchenginewatch.org](http://www.searchenginewatch.org).

<sup>2</sup> See [3] and [4], a slightly nostalgic retrospective.

- Looking for information in a field where one has no or little expertise. In such a scenario tools for grouping and interpreting search results become much more important.

Taxonomies and clustering are two increasingly popular methods to make searching easier and more efficient, especially in the context of the third search scenario. The next section will describe these methods in more detail.

## 2.2 Taxonomies, Navigation, and Clustering

Taxonomies can be used to hierarchically structure information. From the beginning, Yahoo! distinguished itself from other search engines such as AltaVista by offering a hierarchy of categories, which helped users to focus their search.

On the one hand, a taxonomy can be used to navigate or browse for information instead of simply searching information by keywords. Thus, given a taxonomy it is usually possible to restrict search to certain categories in the taxonomy. On the other hand, taxonomies can be used to classify the results of a search, e.g. by indicating the category to which a specific search result belongs, or by grouping all results belonging to the same category together. With respect to the latter application, there are also solutions performing clustering of results, i.e. dynamic classification on the fly, without a pre-defined taxonomy.

It should be obvious, that the benefit provided by taxonomies does not come for free. First of all, building a taxonomy is a complex task. In particular, it involves difficult decisions on how to

- structure a domain, e.g. on whether to prefer a flat structure with few hierarchical levels or a deep structure with more than five levels;
- name the concepts or categories in the taxonomy.

Once a taxonomy is in place, resources have to be categorized accordingly. Doing this manually is time consuming but in general more accurate than automatic categorization. On the other hand, manual categorization has often been blamed as a source of inconsistency, especially if performed by more than one person. Different people tend to categorize information sources differently. Automatic categorization tends to be more consistent. And even if it assigns information sources incorrectly it will probably do this consistently wrong. This might make retrieval easier than if categorization is done inconsistently.

In general, a taxonomy also has to be maintained. Depending on the scope of the taxonomy, new categories may have to be included from time to time or categories have to be merged, due to experiences gained in using the taxonomy.

There are also solutions for automatically clustering search results without the need of a pre-defined taxonomy. These solutions just compute similarities between documents and then group similar documents in so-called clusters. Usually, these clusters are then tagged with a prominent keyword. These approaches have the advantage that no efforts have to be assigned to taxonomy building and maintenance. On the other hand, the quality of the clustering can vary considerably depending on the information available in the resources to be clustered.

## 2.3 Information Retrieval, Recall and Precision

In principle, Internet search engine are just a variant of the traditional field of information retrieval. According to the wikipedia, “information retrieval (IR) is the art and science of searching for information in documents, searching for documents themselves, searching for metadata which describes documents, or searching within databases, whether relational stand alone databases or hypertext networked databases such as the Internet or intranets, for text, sound, images or data.”<sup>3</sup>

In particular, there are two established criteria in IR for measuring the quality of an IR tool, namely precision and recall. Informally, the higher the precision, the lower the number of irrelevant documents retrieved, and the higher the recall, the lower the number of relevant document not retrieved.

Formally, precision is the ratio of correctly found resources to the total number of resources returned. Recall is the ratio of correctly found resources to the total number of resources matching the query. Suppose a query returns 9 documents, one of which is not really matching the query, and does not return 2 documents which would have matched the query, then precision would be 0.89 (8/9) and recall would be 0.8 (8/10).

There usually is a trade-off between precision and recall, i.e. if a retrieval method is too strict, it will achieve a high precision but a low recall, and if it is too lax, it will achieve a high recall but a low precision.

Obviously, it is difficult to actually measure precision and recall, as it presupposes a clearly identified set of the correct resources. In the case of Internet search engines this is a completely unrealistic assumption. Nevertheless, precision and recall are useful criteria when talking about the quality of a search engine. And Google’s success is at least partly due to its high precision, especially with respect to the first 10 results presented.<sup>4</sup>

The field of information research also offers standard methodologies for calculating similarities or distances between resources and between resources and queries. Mathematically, this is usually achieved by formally representing resources and queries as vectors, which can then be compared with various methods.

## 3 EXISTING SOLUTIONS AND CURRENT TRENDS

This section contains an overview of currently, i.e. 2005, existing solutions for search. It covers the following categories of search solutions:

- Internet search engines
- RSS/Blog search
- Desktop search
- Enterprise search
- Taxonomies, categorization, and information discovery

---

<sup>3</sup> en.wikipedia.org/wiki/Information\_retrieval.

<sup>4</sup> A good discussion on how to evaluate Internet search engines can be found in [5].

### 3.1 Internet Search Engines

This section gives a brief overview over the current state and recent trends in Internet search engines, covering three major areas:

- Google and Yahoo
- Alternatives to Google and Yahoo
- Specialized search engines

#### 3.1.1 Google and Yahoo

Google and Yahoo are clearly dominating the sector of Internet search engines at the moment. Quantitative numbers on search engine usage are available but are not always easy to interpret, as there are several criteria, which can be used for measurement.<sup>5</sup>

The situation is further complicated by the facts that several search engines use the same underlying technology and that they change the underlying technology from time to time: “Some search engines get their results by turning to third-party ‘search providers’ to ‘power’ their listings. To make matters more confusing, these search providers may run their own search engines sites, as well” [8].

*Google*<sup>6</sup> is without any doubt the star among today’s Internet search engines. Google’s history begun at Stanford University, where graduate students Larry Page and Sergey Brin developed a search engine called BackRub in 1996.<sup>7</sup> They mainly applied two innovative concepts:<sup>8</sup>

- Computation of a document’s relevance took into account the number and relevance of links pointing to this document (PageRank).
- High performance and scalability was achieved by using a cluster of PCs instead of a single expensive server.

In 1998 Google.com went online as a beta version and by 2000 it had become the largest Internet search engine. The climax, at least for the time being, of this development was reached in August 2004, with Google’s IPO having a value of US \$ 1.7 billion and the subsequent rise of stock value.<sup>9</sup>

Though Google’s main focus is on extensive web coverage and the fast retrieval of relevant web pages, it also offers additional functionality, such as search for images and product information or search in usenet newsgroups. Furthermore, Google Labs offers prototypical implementations of innovative functionality, some of which will be discussed below in the subsections on localized search and clustering. Finally, Google

---

<sup>5</sup> See, for example [6] or [7].

<sup>6</sup> [www.google.com](http://www.google.com)

<sup>7</sup> See [www.google.com/intl/en/corporate/history.html](http://www.google.com/intl/en/corporate/history.html) for a complete history of Google – from the Google perspective.

<sup>8</sup> In addition to these technological innovations, Google introduced adwords, which contributed significantly to its economic success.

<sup>9</sup> The initial stock value was 85\$ a share, which rose to 200\$ in January 2005 leading to a market capitalization of over 50 billion dollars.

also provides solutions for desktop and enterprise research (see the respective subsections below for details).

*Yahoo*<sup>10</sup> was launched in 1994 and offered a web directory containing web sites classified into categories by human editors.<sup>11</sup> This approach ensured a high precision but usually also entailed a low recall (see Section 1.3 for a definition of these terms). In 2002, Yahoo also switched to crawler-based approach and initially used Google as a provider of search results. In February 2004, Yahoo replaced Google as underlying search engine by its own Yahoo Search Technology, which is based on technology from Yahoo subsidiaries Inktomi and Overture Services.<sup>12</sup>

### 3.1.2 Alternatives to Google and Yahoo

For some time it looked as if Google had succeeded in completely controlling Internet search, with Yahoo being the only serious contender. Recently, however, the search market has seen considerable activity and alternatives to Google and Yahoo are continuously improving their functionality and offering new features.

On the one hand, established search sites such as MSN search, continue to attract a significant share of visitors. On the other hand, new search sites are being launched, e.g. Amazon's A9 or Seekport. These sites offer new, innovative features or promise an improvement regarding search quality and usability in order to compete with Google and Yahoo.

In February 2005, Microsoft replaced search technology from Yahoo on its MSN Search site<sup>13</sup> with its own Internet search engine, developed over roughly two years. The new search site aims to provide users not only with tons of links but also with quick answers to their questions. The new MSN search also retrieves content from Microsoft's Encarta encyclopaedia and MSN Music.<sup>14</sup>

A similar combination of general web content and additional content is offered by Amazon's A9 web site<sup>15</sup>: "A9.com is a powerful search engine, using web search and image search results enhanced by Google, Search Inside the Book® results from Amazon.com, reference results from GuruNet, movies results from IMDb, and more."<sup>16</sup> A9 also offers rich functionality for storing past searches and uses the search history for recommending new sites and alerting users to new search results.

Clusty.com<sup>17</sup> is a new search engine developed by Vivismo focusing on clustering of search results, i.e. it organizes results into folders containing similar items. A search for "semantic web", for example, leads to result folders such as World Wide Web, Ontology, Tim Berners-Lee, or Knowledge.

---

<sup>10</sup> [www.yahoo.com](http://www.yahoo.com)

<sup>11</sup> [docs.yahoo.com/info/misc/history.html](http://docs.yahoo.com/info/misc/history.html)

<sup>12</sup> See [9] for details and a discussion of the impact of this development.

<sup>13</sup> [www.msnsearch.com](http://www.msnsearch.com).

<sup>14</sup> See, for example, [10].

<sup>15</sup> [www.a9.com](http://www.a9.com).

<sup>16</sup> [a9.com/-/company/whatsCool.jsp](http://a9.com/-/company/whatsCool.jsp).

<sup>17</sup> [www.clusty.com](http://www.clusty.com).

### 3.1.3 Specialized Search

In addition to general Internet search engines, there is a wide range of special purpose search engines. These include search engines for images, news, newsgroups, answers, kids, shopping, or specific domains, e.g. medical, legal, or financial. Due to their focus on a subset of Web content such engines often produce more accurate results than Google.

Steve Bass discusses some of these engines such as Chowhound for restaurants in certain cities, Craigslist for classified ads, or TVEyes for monitoring television shows [11]. These examples highlight a couple of trends that are already apparent and will probably become even more dominant in the near-term future:

- The combination of search with location information.
- The inclusion of multimedia information in search.

In the last year, Google, Yahoo, MSN Search, and Ask Jeeves<sup>18</sup> have all added localized search to their search functionality. The basic idea is to combine search results with location information, e.g. to search for restaurants or service providers in a certain city or geographic area.<sup>19</sup>

Most search engines already support search for images but recently Google and Yahoo have also added functionality for searching TV programmes. Such a specialized search service is also offered by blinkx.tv.<sup>20</sup>

Another approach to focus on special aspects of search is exemplified by Answers.com<sup>21</sup>, powered by GuruNet. This site provides definitions and explanations for over 1 million topics, i.e. it is answering queries of the type “who/what is ...”.

## 3.2 Blog/RSS Search

During the last two years, RSS (Rich Site Syndication) has become the widely accepted standard for content syndication. Most news sites and blogs make their content available via RSS.<sup>22</sup>

The main advantage of RSS is that users can quickly review a wide range of information sources through a unified feed reader. Usually, the feed reader shows a short summary of the news article or the blog entry. Only if an entry really is of interest to the user, does she actually have to visit the original site on which the entry has been published.

There are mainly three approaches to combine RSS and search:

- Several sites offer a search interface allowing searching for RSS feeds.

---

<sup>18</sup> [www.askjeeves.com](http://www.askjeeves.com).

<sup>19</sup> See [maps.google.com/](http://maps.google.com/), [local.yahoo.com/](http://local.yahoo.com/), [local.msn.com](http://local.msn.com), and [local.ask.com/local](http://local.ask.com/local).

<sup>20</sup> See, for example, [12].

<sup>21</sup> [www.answers.com](http://www.answers.com).

<sup>22</sup> There also is an alternative standard called Atom. In the following, the presentation will focus on RSS, but the arguments presented here are also valid for Atom.

- Some sites offer a search interface allowing searching for individual entries in RSS feeds.
- Some sites provide search results as RSS feeds.

Although RSS feed readers considerably simplify the access to a wide variety of information sources, the increasing number of RSS feeds available leads to the information overload already known from the Web in general.

There are several sites offering search functionality specifically for RSS or blogs, such as technorati.com, Syndic8.com, PubSub.com, Bloglines.com, Deli.cio.us, feedster.com, Blogdigger.com, and Completerss.com.

Most of these sites allow searching for feeds by specifying keywords. Some also support category systems, e.g. based on DMOZ/Open Directory or NewsIsFree, which allow browsing for feeds belonging to a specific category.<sup>23</sup> Some sites also provide web services interfaces to access their functionality.<sup>24</sup>

In addition to searching for entire feeds, some RSS search engines also allow the creation of custom feeds based on a keyword search.<sup>25</sup> These filtered feeds are then again provided as RSS feeds. There also is an experimental alpha feature under development at MSNSearch, returning search results as RSS feeds.<sup>26</sup>

Blogdigger.com<sup>27</sup> offers a search interface for finding individual blog entries based on RSS/Atom technology. It also makes the search results available in RSS or Atom so that users can subscribe to their individual search feeds and easily become aware of new content available. Similarly, Technorati.com currently, i.e. Q2 2005, tracks almost five million weblogs and makes their content available for search.<sup>28</sup>

The main challenge for blog search engines are the real-time requirements: “A new weblog is created every 7.4 seconds, which means there are about 12,000 new blogs a day. Bloggers — people who write weblogs — update their weblogs regularly; there are about 275,000 posts daily, or about 10,800 blog updates an hour.”<sup>29</sup>

Although RSS search engines help to cope with the information overload generated by blogs and RSS feeds, additional technology will probably be needed to further improve efficiency. Bill Burnham argues that Meta-feeds will be needed to address the problem of feed overloads: “Meta-feeds are RSS feeds comprised solely of metadata about other feeds. Combining meta-feeds with the original source feeds enables RSS readers to display consistently categorized posts within rich and logically consistent taxonomies” [13].

---

<sup>23</sup> See, for example, [www.syndic8.com/feedcat.php](http://www.syndic8.com/feedcat.php).

<sup>24</sup> See, for example, [www.syndic8.com/web\\_services/](http://www.syndic8.com/web_services/).

<sup>25</sup> See, for example, [www.completerss.com/tools/customfeed.aspx](http://www.completerss.com/tools/customfeed.aspx).

<sup>26</sup> See [blogs.msdn.com/mnsnsearch/archive/2005/01/11/35106.aspx](http://blogs.msdn.com/mnsnsearch/archive/2005/01/11/35106.aspx) for details.

<sup>27</sup> [www.blogdigger.com](http://www.blogdigger.com).

<sup>28</sup> [www.technorati.com/about/](http://www.technorati.com/about/).

<sup>29</sup> [www.technorati.com/about/](http://www.technorati.com/about/).

However, Burnham also points out that there still is a long way to go before Meta-feeds will be widely used and that “in order for such feeds to become truly accepted, standards will have to be developed that incorporate meta-feeds into readers and allow for interoperability between meta-feeds”.

### 3.3 Desktop Search

Desktop search allows the user to search for data available on the hard disk of a computer, e.g. Word documents, emails, PowerPoint slides or Excel sheets. Desktop search tools can be used both in a consumer or a business context and most tools are not strictly restricted to a single desktop but also support search in a network.

#### 3.3.1 Vendor Landscape

As indicated above, most Internet search engines have started to make their technology available for desktop search. In addition, there are also companies offering dedicated desktop search tools, some of them already in business for several years. Finally, some vendors of enterprise search solutions are currently, i.e. 2005, developing desktop search versions of their products.

A desktop search matrix<sup>30</sup> by Goebel Group Inc., currently, i.e. Q1 2005, lists 17 desktop search tools, including desktop versions from AOL<sup>31</sup>, Ask Jeeves, Google, Lycos<sup>32</sup>, MSN, and Yahoo. Most of these tools are still in a beta version and available for free, but there are also stable product versions ranging from 0\$ (e.g. Blinkx, Copernic) over 75\$ (X1) to 500\$ (ISYS:desktop).

The most obvious distinctions between the various tools available, concern the languages (primarily English) and operating systems (primarily Windows) they support and, most importantly, the searchable file types. Whereas Internet search is primarily working on HTML pages, desktop search has to deal with a wide variety of formats, e.g. for documents, emails, or database content, which often depend on the particular application software used. Thus, in order to search ones email with a desktop search tool, the tool has to support the particular software used for reading and storing email, e.g. Outlook, Netscape, Lotus, AOL.

The desktop versions of Internet search engines are still rather limited with respect to the file types they are supporting and are mostly restricted to English. Established desktop search vendors, on the other hand, usually support a wide range of file types and languages.

*Blinkx*<sup>33</sup> is offering desktop search for Windows PC and Apple Mac and has recently also released a TV search engine. A special feature of blinkx are so-called smart folders, i.e. “intelligent folders that automatically and persistently update their content as new information becomes available.”<sup>34</sup> Based on documents initially put into such a smart folder by the user, blinkx finds similar documents and puts them into the folder.

---

<sup>30</sup> [www.goebelgroup.com/desktopmatrix.htm](http://www.goebelgroup.com/desktopmatrix.htm).

<sup>31</sup> [www.aol.com](http://www.aol.com).

<sup>32</sup> [www.lycos.com](http://www.lycos.com).

<sup>33</sup> [www.blinkx.com](http://www.blinkx.com).

<sup>34</sup> [www.blinkx.com/content/backgrounder.php#smartfolders](http://www.blinkx.com/content/backgrounder.php#smartfolders).

*Copernic*<sup>35</sup> offers a desktop search tool, which also supports summarization and change tracking. In October 2004, Copernic has created a new separate company, Coveo Solutions Inc., which focuses on secure enterprise search software.

*ISYS*<sup>36</sup> was founded in 1988 in Australia and thus has a comparatively long history of experience in the search market. “ISYS:desktop is a network-based search application that can be used on a single PC, searching email and personal documents, right up to thousands of PCs across enterprise-wide LANs and WANs.”<sup>37</sup> ISYS also offer products supporting search in a web-based environment.

*X1*<sup>38</sup> was founded in 2001 by Bill Gross, who “reassembled many of the team members who in 1987 created Lotus Magellan – the original PC-based file search product.”<sup>39</sup> An important feature of X1 is its speed – search results are presented while typing the search query.

Other vendors of desktop search include DT Search<sup>40</sup>, Enfish<sup>41</sup>, JetBrains<sup>42</sup>, Svizzer<sup>43</sup>, The Sleuthhound<sup>44</sup> and Wizetech Software<sup>45</sup>. Recently, *Aduna*<sup>46</sup> has released AutoFocus, a desktop search solution based on RDF and Sesame, an open source RDF implementation.

### 3.3.2 Security Issues

Following the release of Google’s desktop search beta version in October there has been an intense debate on potential security threads. Three main aspects can be distinguished:

- Google indexes files that might not be suitable for retrieval through search, e.g. cached web pages or decrypted versions of encrypted files.
- Google mixes results from Internet search with desktop search results and there have been concerns about Google being able to spy out desktop contents.
- The initial release contained a security flaw allowing “hackers to gain access to the user’s computer and trick the desktop search feature to retrieve information from another website instead of the local hard disk drive”.<sup>47</sup> This flaw has been fixed in subsequent releases.

---

<sup>35</sup> [www.copernic.com](http://www.copernic.com).

<sup>36</sup> [www.isysusa.com](http://www.isysusa.com).

<sup>37</sup> [www.isysusa.com/products/index.html](http://www.isysusa.com/products/index.html).

<sup>38</sup> [www.x1.com](http://www.x1.com).

<sup>39</sup> [www.x1.com/about\\_us/](http://www.x1.com/about_us/).

<sup>40</sup> [www.dtsearch.com](http://www.dtsearch.com).

<sup>41</sup> [www.enfish.com](http://www.enfish.com).

<sup>42</sup> [www.jetbrains.com](http://www.jetbrains.com).

<sup>43</sup> [www.svizzer.com](http://www.svizzer.com).

<sup>44</sup> [www.isleuthhound.com](http://www.isleuthhound.com).

<sup>45</sup> [www.wizetech.com](http://www.wizetech.com).

<sup>46</sup> [www.aduna.biz](http://www.aduna.biz).

<sup>47</sup> See, for example, [14].

With respect to the first concern, it has been pointed out that Google does not make any information available, which has not been available before.<sup>48</sup> The real problem is the storage of decrypted documents and critical web page content in the respective caches, not the indexing of these caches by Google. All users having administrator rights on a desktop can view such cache files anyway – regardless to which user they belong. To run Google desktop search one also has to have administrator rights, i.e. the privacy problem is not created by the desktop search. However, finding sensitive information will become much easier via desktop search. And one might thus argue that less criminal energy is required to violate privacy issues on a machine on which Google desktop search is installed.

### 3.3.3 Expected Trends

It is sometimes argued that nobody needs desktop search, since a clearly structured directory system is all you need for efficiently locating information on a desktop. However, it is very likely that more and more users will install a free desktop search tool on their computers and use it more or less regularly for accessing information. After all, installation is simple and search interfaces are well known from Internet search. And everyone can probably remember situations in which information could not easily be located via the directory structure of the file system – even if these are not situations occurring every day.

Although there will thus be some demand for desktop search, it is less clear whether dedicated desktop search tools will be the solution of choice. On the one hand, most applications offer specific search tools, on the other hand generic search functionality will become a part of operating systems in the near future.

Application-specific search interfaces can take into account the particular structure of the information used in the application. Thus, tools for searching email usually offer means to search for keywords in mail specific fields such as sender, recipient or subject. Search in music tools such as itunes allows searching for titles, artists, albums, etc.

Both Apple and Microsoft are currently, i.e. 2005, developing search functionality to be included in their upcoming file system releases. Microsoft<sup>49</sup> originally planned a new, revolutionary file system called WinFS as a central component of Longhorn. However, it seems that WinFS will not be included into Longhorn after all, but will be made available as a test version sometime in 2006 (the release of Longhorn is also foreseen for 2006).<sup>50</sup>

Apple<sup>51</sup>, on the other hand, is developing its Spotlight technology, which will be part of Mac OS X version 10.4 Tiger.<sup>52</sup> Spotlight will also support the use of meta data and smart folders, to which files are added based on specified search rule. Spotlight is thus offering the same functionality currently provided by most desktop search engines and could make the installation of such a tool redundant.

---

<sup>48</sup> See, for example, [15] for a good and concise discussion of the security issues at hand.

<sup>49</sup> [www.microsoft.com](http://www.microsoft.com).

<sup>50</sup> See, for example, [16].

<sup>51</sup> [www.apple.com](http://www.apple.com).

<sup>52</sup> [www.apple.com/ca/macosx/tiger/spotlight.html](http://www.apple.com/ca/macosx/tiger/spotlight.html).

### 3.4 Enterprise Search

Enterprise search is a highly dynamic market, which has originally developed more or less independently of Internet search engines. Autonomy and Verity have been the dominant players in this area, but recently Internet search engines have begun to offer products addressing this market as well.

Enterprise search shares some of the characteristics of Internet search and desktop search but there are also several issues that are particular to this field. As in desktop search, the environment is usually controlled and closed in contrast to the open environment of the Internet. But as in Internet search, enterprise search has to deal with a distributed environment in which document location is potentially more open than in desktop search.

One major advantage of a controlled environment is that it allows the effective use of metadata. Internet search engines face the problem of “metadata spamming”, i.e. the inclusion of misleading metadata by webmasters in order to improve their page ranking and drive extra traffic to their sites. As a consequence, most Internet search engines tend to ignore metadata. In an enterprise environment, on the other hand, metadata spamming will hardly become an issue – all parties involved will cooperate to make the enterprise search solution as good as possible.

In general, including human intervention, i.e. manual steps in addition to purely automatic information processing is both highly useful and feasible in the context of enterprise search. This includes the provision of metadata by document authors or editors, the explicit ranking of documents as best answers for certain queries, or the creation of new content based on the evaluation of user behaviour.

Finally, there is one area where enterprise search is facing a challenge that is usually not relevant for Internet search or desktop search: security. An enterprise search solution has to make sure that content is only made available to those employees entitled to access the content. This usually involves complex user and rights management.

The following subsections address in more detail how enterprise search solutions provide added value to the standard search functionality provided by Internet search engines.<sup>53</sup> This concerns the following topics:

- Crawler performance
- Document pre-processing
- Query pre-processing
- Post-processing of results

The section then ends with a brief overview over the vendor landscape in enterprise search.

#### 3.4.1 Crawler Performance

The crawler is responsible for locating information resources and storing the relevant parts of their content into a database (the index). One challenge for any crawler is to find all information resources based on a few starting points. In a file system this can be

---

<sup>53</sup> See, for example, [17] for a good overview on enterprise search technology.

achieved by looking into all sub directories of a specified start directory; on a web site the crawler begins with a start page and then recursively follows all the links included in the web pages.

There are some intranet technologies which pose major challenges for crawlers, such as password-protected sites, sites with cookie-based content, sites generated by content management systems or databases, or sites using non-standard coding techniques. In order to deal with these challenges, enterprise search solutions usually offer integration with content management systems. Another major challenge for crawlers is to detect duplicate pages or documents and to keep such duplicated information out of the index.

Incremental crawls can be used to optimize crawler performance. Instead of crawling an entire web site, only the dynamic parts of a site are crawled on a regular basis, whereas static parts are only revisited in larger intervals. Information on content changes can also be obtained by integrating the enterprise search system with a content management system.

Finally, bandwidth restrictions and time range settings can be used to make sure that crawler activity does not have any negative impact on other network traffic.

### **3.4.2 Document Pre-Processing**

A key challenge in enterprise search is the extraction of useful metadata. Human intervention is sometimes feasible in the context of enterprise information but convincing document authors to provide metadata is often a futile endeavour.<sup>54</sup>

Automatic processing of documents and content extraction is gaining importance (see the section on information discovery below) and domain-specific solutions are beginning to produce qualitatively interesting results.

### **3.4.3 Query Pre-Processing**

The standard approach to querying a search engine is the input of one or more keywords. Most search engines do also support so-called advanced searches, e.g. by allowing Boolean expressions. However, the vast majority of end users does not use these advanced search features at all and some experts argue that they create more damage than benefit.<sup>55</sup>

Instead of relying on advanced search features, some search engines therefore perform an implicit query pre-processing to improve the quality of their search results. Examples for such a pre-processing are the use of synonyms or of grammatical variants of the keywords entered and automatic insertion of Boolean operations.

Although these techniques can help to improve the search quality, the most promising improvements seem to be achievable by a post-processing of search results.

### **3.4.4 Post-Processing of Results**

How search results are presented has a tremendous effect on the usability of the results. A very interesting experience report on the optimizing of search results on the BBC web site is given by Martin Belam [19]. In particular, he describes the effect of “banner

---

<sup>54</sup> See [18] for a profound discussion on the benefits and pitfalls of metadata.

<sup>55</sup> A useful variant of “advanced” search is the use of special purpose fields for using certain content types, e.g. “director”, “actor” for movies or “band”, “title” for songs.

blindness”, which lead users to ignore prominently displayed links on the ad-free intranet because they were presented in boxes resembling banner ads.

Key aspects of result post-processing concern the display of informative titles, creation dates, and the highlighting of relevant passages containing the keywords. Some search engines also perform on the fly categorization of search results and arrange them into clusters. This is especially useful if the search terms used are ambiguous or have different meanings in different contexts.

A particular challenge is posed by pages generated automatically and dynamically on the basis of cookies or frames. It is only possible to reconstruct such pages if frame and/or cookie information is stored in the index.

### 3.4.5 Vendor Landscape<sup>56</sup>

Autonomy and Verity are usually identified as the major vendors in the enterprise search market. *Autonomy*<sup>57</sup> is often perceived as providing highly priced, complex systems. It positions itself not as a search engine but “as an intelligent operating system, sitting on top of the actual operating system”.<sup>58</sup> Their Intelligent Data Operating Layer (IDOL) is based on automated pattern recognition and probabilistic technology, leveraging information theory work by Bayes and Shannon. It computes the relevancy of unstructured information in a given context by analyzing and understanding its content.

*Verity*<sup>59</sup> has started as a search vendor but is currently, i.e. 2005, orienting itself towards business process management. It has acquired search technology provider Ultraseek, forms vendor Cardiff Software, and NativeMind, a company offering natural language functionality. Verity currently has a market capitalization of 451 million dollars.

Convera and FAST are perceived as the current runners up behind Autonomy and Verity. *Fast Search and Transfer (FAST)*<sup>60</sup> has been gaining significant market share in the last two years and is now a serious competition for Autonomy and Verity. It is calling itself the world leader in enterprise search. It moved into the enterprise market after selling its All-the-Web Internet search business to Overture. In return, it acquired Alta Vista’s enterprise business from Overture. One of FAST’s key advantage is its modular, modern architecture, which allows easy integration and management on the basis of XML and other Web standards.

*Convera*<sup>61</sup> has a strong position in the eGovernment market and also targets industries with sophisticated classification challenges, such as pharmaceuticals, life sciences, or medical research. It combines statistical algorithms with an ontological approach to identify and extracts terms, phrases, and concepts.

---

<sup>56</sup> The content of this subsection is largely based on [20].

<sup>57</sup> [www.autonomy.com](http://www.autonomy.com).

<sup>58</sup> [www.autonomy.com/content/Autonomy/FAQ.html](http://www.autonomy.com/content/Autonomy/FAQ.html).

<sup>59</sup> [www.verity.com](http://www.verity.com).

<sup>60</sup> [www.fastsearch.com](http://www.fastsearch.com).

<sup>61</sup> [www.convera.com](http://www.convera.com).

Other vendors in the area of enterprise search include Coveo, a company created by desktop search vendor Copernic, Endeca<sup>62</sup>, Google, Hummingbird<sup>63</sup>, Mondosoft<sup>64</sup>, NorthernLight<sup>65</sup>, Open Text<sup>66</sup>, and Thunderstone<sup>67</sup>.

### 3.5 Taxonomies, Categorization, and Information Discovery

Taxonomies and categorization are intimately related to search in general and to enterprise search in particular. Autonomy, Verity, and Convera all offer functionality in this area and FAST has formed a partnership with Stratify to cover this area.

#### 3.5.1 Using Taxonomies and Categorization in Search

There are several ways in which taxonomies and categorization can be used in the context of search. The most obvious is the browsing or navigation of content organized in a taxonomy. This involves the subtasks of first modelling a hierarchy of categories and then classifying documents into the appropriate categories. The latter can either be performed manually or automatically by using (a combination of) statistical algorithms, linguistic methods, or hand-coded business rules.<sup>68</sup>

A second area of usage concerns the clustering of search results into categories. This is widely acknowledged to significantly enhance usability of search results, especially in scenarios where precision is low, large numbers of documents are returned, or users have little expertise in the subject matter they are searching for. Such a clustering can be based both on a pre-defined taxonomy as on dynamic categorization. In the latter case, categories are computed on the fly, based only on the information contained in the search results and without taking into account any pre-defined taxonomy.

A wide variety of software is currently, i.e. Q1 2005, available for “auto-categorization”, usually also allowing human intervention to improve the quality of the automatic approach.<sup>69</sup> Some solutions also offer related technology, such as automatic summarization, metadata generation, or information extraction, or provide functionality for not only categorizing entire documents but also individual paragraphs in a document.

#### 3.5.2 Vendor Landscape<sup>70</sup>

In addition to vendors already covered in the section on enterprise search, such as Autonomy, Convera, Google, and Verity, there are several technology providers focussing specifically on taxonomies, categorization and information discovery.

---

<sup>62</sup> [www.endeca.com](http://www.endeca.com).

<sup>63</sup> [www.hummingbird.com](http://www.hummingbird.com).

<sup>64</sup> [www.mondosoft.com](http://www.mondosoft.com).

<sup>65</sup> [www.northernlight.com](http://www.northernlight.com).

<sup>66</sup> [www.opentext.com](http://www.opentext.com).

<sup>67</sup> [www.tunderstone.com](http://www.tunderstone.com).

<sup>68</sup> See, for example, [21].

<sup>69</sup> See [22] for a good introduction and overview.

<sup>70</sup> This section is based on [23].

*Inxight*<sup>71</sup> was founded in 1997 and positions itself as “one of the world’s leading providers of information discovery solutions that enable customers to discover, retrieve, and connect with precise information contained in unstructured data sources in all major languages”.<sup>72</sup> Products provide advanced text analysis tools including search, entity extraction, event extraction, categorization and visualization.

*Stratify*<sup>73</sup> offers solutions for the management of unstructured data with a focus on legal eDiscovery. Target users are attorneys, litigation teams, and Legal IT. *Entopia*<sup>74</sup> was founded in 1999 and develops innovative solutions for information discovery, including enterprise search and content visualization. *Mohomine* technology from Kofax<sup>75</sup> offers solutions for classification and for information extraction from candidate resumes. *Entrieva*<sup>76</sup> offers products for advanced text analytics, allowing improvement of search engine performance and optimization. *Wherewithal*<sup>77</sup>, founded in 1998 by former Netscape visionary Steve Thomas, provides software allowing the collaborative creation of central category structures and search engines in an enterprise.

### 3.6 Conclusion and Expected Trends

The sections above have shown the widespread use of search technology today. Given the development in the last two years, it is very likely that the importance of search will increase and that search will become a ubiquitous feature. This is, for example, illustrated by recent functionality for searching television programmes.

Google and Yahoo will try to dominate the search market and extend their portfolios, e.g. by offering products for desktop search and enterprise search or by providing additional specialized search services. But in spite of the dominance of Google and Yahoo, other search engines have still not given up the race. And newcomers like Amazon’s A9 might become serious alternatives due to the additional functionality they are providing.

And then there are Microsoft and Apple, which are currently, i.e. 2005, working on including advanced search technology into their operating systems. Apple’s spotlight seems to be already rather advanced, whereas Microsoft’s plans for including a new revolutionary file system (WinFS) into Longhorn seem to be seriously delayed.

Finally, techniques for information extraction and categorization have reached a quality, which makes their inclusion in search technology highly beneficial. It is very likely that most search engines will offer categorization functionality in the near future to help structuring their search results.

---

<sup>71</sup> [www.inxight.com](http://www.inxight.com).

<sup>72</sup> [www.inxight.com/about/](http://www.inxight.com/about/).

<sup>73</sup> [www.stratify.com](http://www.stratify.com).

<sup>74</sup> [www.entopia.com](http://www.entopia.com).

<sup>75</sup> [www.kofax.com](http://www.kofax.com).

<sup>76</sup> [www.entrieva.com](http://www.entrieva.com).

<sup>77</sup> [www.wherewithal.com](http://www.wherewithal.com).

## 4 SWS POTENTIAL FOR SEARCH

There are basically two main scenarios for using Semantic Web Services in Search:

- On the one hand, search functionality can be provided for Semantic Web resources or for Semantic Web Services itself, e.g. supporting the discovery of available services;
- On the other hand, Semantic Web Services could be used to enhance the functionality currently provided by “traditional” search technology, e.g. regarding categorization or pre-/post-processing of documents.

Note that the first approach requires information to be available in a special format, e.g. RDF, OWL, OWL-S, or WSMO. It thus differs from current approaches to search, which take unstructured information sources as input, and solutions would therefore complement existing search solutions. The second approach, on the other hand, would be much more integrated with existing solutions, as it would offer additional functionality to be combined with functionality already provided by current search engine technology.

This section will discuss four application scenarios in more detail:

- Searching Semantic Web resources
- Discovering Semantic Web Services
- Information extraction and categorization
- Search as a Semantic Web Service

### 4.1 Searching Semantic Web Resources

#### 4.1.1 Semantic Web Search Engines

Swoogle<sup>78</sup> “is a crawler-based indexing and retrieval system for the Semantic Web -- RDF and OWL documents encoded in XML or N3. (...) Swoogle's database currently, i.e. Q1 2005, has information on 337,241 semantic web documents which contain 47,566,836 triples and define 96,572 classes, 54,356 properties and 7,278,890 individuals. (...) Currently, the most popular kinds of documents are FOAF files and RSS files.”<sup>79</sup>

Swoogle can search for entire documents as well as for specific terms. Document searches can be constrained to specific file types, such as FOAF or RSS or to specific encodings, such as xml/rdf or n3. Term search can be used to search for specific classes, properties, or instances. Swoogle also computes metadata for the Semantic Web documents it indexes, such as the number of defined classes, properties, or instances.<sup>80</sup>

A similar solution is offered by [www.semanticwebsearch.com](http://www.semanticwebsearch.com), a site allowing “both people and computers to precisely locate and gather information published on the

---

<sup>78</sup> [www.swoogle.org](http://www.swoogle.org).

<sup>79</sup> [pear.cs.umbc.edu/swoogle/about.php](http://pear.cs.umbc.edu/swoogle/about.php).

<sup>80</sup> See [swoogle.umbc.edu/modules.php?name=Documents&file=manual](http://swoogle.umbc.edu/modules.php?name=Documents&file=manual) for a Swoogle manual.

Semantic Web.”<sup>81</sup> In contrast to Swoogle, SemanticWebSearch allows the use of specific properties in queries, e.g. (foaf:surname), (rss:item), or (dc:creator). In addition to a GUI for human users, there also is a web service interface allowing the use of SemanticWebSearch by computer programs, e.g. by intelligent software agents.

The MuseumFinland uses OntoViews, a tool for creating semantic web portals, for its web sites and provides end users with a semantic-based search engine and a recommendation system.<sup>82</sup> The content stems from several heterogeneous Finish museum databases and has been consolidated via RDF and ontologies.<sup>83</sup>

As already mentioned above, it should be noted that these approaches presuppose information to be available in a Semantic Web format, e.g. RDF, OWL, or FOAF. They should therefore not be seen as alternatives to existing solutions for generic Internet search, desktop search, or enterprise search. However, the demand for these solutions will continuously increase as more and more Semantic Web information will become available.

#### 4.1.2 Discovering Semantic Web Services

In addition to searching the Semantic Web it will also become increasingly important to search for Semantic Web Services. In the context of Web Services this is usually called discovery, a task supported by UDDI in the case of traditional Web Services.

For Semantic Web Services, complex reasoning can be used to facilitate service discovery. This involves matchmaking between the properties of available services and properties required by the user, including capabilities, input and output arguments, as well as pre-conditions and post conditions. Technically, matchmaking is realized via subsumption checking, e.g. by checking whether the post conditions guaranteed by a service are subsumed by the post conditions required.<sup>84</sup>

In the context of WSMO, discovery is mainly realized via the notion of goals. In order to find Semantic Web Services, users specify goals they want to achieve. Discovery then provides a service appropriate for the specified goal, potentially by composing a complex orchestrated service from available basic services.<sup>85</sup> More precisely, discovery can be divided into goal definition, web service discovery, and service selection (see [26, 1.1.2].).

During Goal Definition the user specifies the goal she wants to achieve. This step will be supported by tools for browsing goal repositories or for refining an initial abstract goal. The outcome of this phase is a formally specified goal that can be used as input for the next phase, service discovery. In this phase the specified goal is compared with the capabilities of WSMO web services. This phase will return a set of services which provide capabilities sufficient to achieve the goal specified by the user. The third phase then involves the actual selection of a Web Service from the set of services returned in

---

<sup>81</sup> [www.semanticwebsearch.com/faq.rsp](http://www.semanticwebsearch.com/faq.rsp).

<sup>82</sup> [museosuomi.cs.helsinki.fi/](http://museosuomi.cs.helsinki.fi/).

<sup>83</sup> For details, see [24].

<sup>84</sup> A concept A is subsumed by concept B if A is more special than or equal to B.

<sup>85</sup> Other SWS language such as OWL-S and METEOR offer similar functionality for SWS discovery. See, for example, [25] for a good overview.

the second phase. This can involve the evaluation of information not contained in the capability specification and even negotiation.

## 4.2 Using SWS in Search Engines

### 4.2.1 Information Extraction and Categorization

The search applications of Semantic Web Services discussed so far were based on information available in structured formats, such as RDF, OWL, FOAF, WSMO, or OWL-S. But it would also be possible to use Semantic Web Services in solutions dealing with unstructured text, e.g. for information extraction and categorization.<sup>86</sup>

The general idea would thus be to produce structured information from unstructured texts. This could be, for example, general meta data, as well as data in Semantic Web formats such as OWL or RDF. Such extraction services are probably most realistic for specific domains or document types, e.g. News or RSS Feeds.

Furthermore, given the current state of the art, such services could probably only be used during the pre-processing of documents. On the fly categorization of result documents is already performed in real time and it is unlikely that a Semantic Web Service would be able to meet these time constraints while producing qualitatively better results.

However, time constraints are less demanding during the pre-processing of documents. Thus, offering a Semantic Web Service taking an HTML document (and a taxonomy) as input and returning lists of extracted information items and categories as a result would be highly useful. Search engines could experiment with such a service in order to evaluate their potential for improving the quality of search results.

Such a service would not only be useful for Internet search engines but could also be used in the context of desktop search, RSS search, and enterprise search. It is very probable, that the importance of functionality for information extraction and categorization will increase significantly in the near-term future. Semantic Web Services are clearly the most straightforward means to making such functionality widely available and easy to integrate. The main challenge consists in the development of robust, accurate, efficient, and scalable algorithms implementing such functionality, however.

### 4.2.2 Intelligent Search as a Semantic Web Service

Finally, it would be possible to realize arbitrary search functionality as a Semantic Web Service. This is often the implicit assumption underlying scenarios describing future search applications in the context of the Semantic Web.<sup>87</sup> In these scenarios, intelligent agents roam the web to extract and combine useful information.

Such a service would probably not be able to operate in real-time. This, however, might not necessarily be a show stopper, as there are clearly contexts in which users would not require search results to be available within seconds. In particular, a considerable

---

<sup>86</sup> See, for example, [www.nitle.org/tools/semantic/search.htm](http://www.nitle.org/tools/semantic/search.htm).

<sup>87</sup> See, for example, [www.arisem.com/en/products/semanticsearchengine.html](http://www.arisem.com/en/products/semanticsearchengine.html) or [tap.stanford.edu/tap/ss.html](http://tap.stanford.edu/tap/ss.html).

increase in the quality of the search results would outweigh the delay in response in many scenarios.

The added value of a semantic search service would be highest in a scenario where the user provides a search query and information about the search context, e.g. what kind of results are expected (compiled answer, lists of relevant sites, research papers, etc.). This corresponds to a complex goal in the WSMO terminology used in DIP. The search service could then provide functionality for pre-processing the query, selecting information sources, and post-processing of results.

The most straightforward example of such query pre-processing is the use of hierarchical information or instance relations from ontologies. Thus, when the user enters a keyword like “computer” the pre-processing could generate search queries for “laptop”, “PC” in addition to the original search query “computer”.

In a sense, such a service would act like a meta search engine, however one which would not only distribute queries and collect results, but would also perform intelligent processing during query distribution and result compilation. Thus the selection of appropriate information sources could be based on the specified search query and the search context. Moreover, ontological information could be used to automatically derive the appropriate keywords from the terms used in the search query.

Finally, post-processing of search results could be automated by an intelligent search service. This includes refining search queries based on the results initially obtained, clustering of search results, as well as extraction and combination of relevant information contained in the search results.

## **5 RECOMMENDATIONS FOR DIP**

Although search is an application area with high potential for Semantic Web Services, there are not too many options for using DIP results immediately in a search-based application. With respect to the four application scenarios outlined in Section 4, discovery of services is the scenario closest to current research activities in DIP.

In the specification and implementation of discovery (D4.8, D4.14, D4.17, D4.18, D4.19), DIP could reuse experiences from existing search solutions. This concerns the refinement of queries as well as the presentation of search results in general, and the clustering of results in particular. As pointed out in Section 3.4.4, the information provided in the presentation of search results has considerable impact on the perceived quality. It would thus be highly beneficial to include service attributes relevant for service selection when displaying all services whose capabilities match the goal specified in discovery.

Information extraction and categorization will become increasingly important and providing these functionalities as Semantic Web Services would make a lot of sense. However, DIP does not address text processing at all and does not develop any functionality in this area.

What could be realistically achieved within DIP would be the implementation of a simple SWS on top of existing search engines. Such a service could, for example, query Semantic Web resources via Swoogle or could focus on a specific format (RSS) or domain (TV, News). One straightforward option would be a search function on the DIP web site using a small “DIP ontology” for pre-processing and post-processing of search

queries. The main benefit of such a service would be that it would be highly visible and demonstrate the potential of SWS in a lightweight manner.

However, there is currently no deliverable for such a use case and there are thus no resources for its development. It would thus be necessary to include an explicit deliverable in the revised annex for the third year of DIP in order to make its realization possible.

With respect to exploitation of DIP tools, industrial partners of DIP should consider search as one of the application areas with high potential. However, an exploitation of technological components alone will probably be difficult in this area. More promising is the exploitation in the form of specific Semantic Web Services, e.g. for information extraction and categorization. To achieve this, the tools developed in DIP would have to be coupled with tools for text processing.

## 6 REFERENCES

- [1] DIP Deliverable D12.1, Joachim Quantz, Thorsten Wichmann, Report on current usage of Web Services and Semantic Web, June 2004
- [2] DIP Deliverable D12.3, Joachim Quantz, Thorsten Wichmann, Report on Key Technology Issues in Current EAI, E-Business and Knowledge Management (Enterprise Collaboration with Semantic Web Services), January 2005.
- [3] Danny Sullivan, “Major Search Engines and Directories”, SearchEngineWatch, April 2004, [searchenginewatch.com/links/article.php/2156221](http://searchenginewatch.com/links/article.php/2156221)
- [4] Danny Sullivan, “Where Are They Now? Search Engines We’ve Known and Loved”, SearchEngineWatch, March 2003, [searchenginewatch.com/sereport/article.php/2175241](http://searchenginewatch.com/sereport/article.php/2175241)
- [5] Danny Sullivan, “Inktomi, Google Win In Recent Relevancy Test”, SearchEngineWatch, April 2003, [searchenginewatch.com/searchday/article.php/2192401](http://searchenginewatch.com/searchday/article.php/2192401)
- [6] Danny Sullivan, Major Search Engines and Directories, SearchEngineWatch, April 2004, [searchenginewatch.com/links/article.php/2156221](http://searchenginewatch.com/links/article.php/2156221)
- [7] Chris Sherman, Yahoo & MSN Closing the Google Gap, SearchEngineWatch, January 2005, [searchenginewatch.com/searchday/article.php/3458351](http://searchenginewatch.com/searchday/article.php/3458351)
- [8] Danny Sullivan, Who Powers Whom? Search Providers Chart, SearchEngineWatch, July 2004, [searchenginewatch.com/reports/article.php/2156401](http://searchenginewatch.com/reports/article.php/2156401)
- [9] Mitko Gerensky-Greene, Yahoo!’s New Search Engine, SEO Chat, March 2004
- [10] Juan Carlos Perez, Microsoft Spotlights Its Search Engine, PCWorld, February 2005, [www.pcworld.com/news/article/0,aid,119512,00.asp](http://www.pcworld.com/news/article/0,aid,119512,00.asp)
- [11] Steve Bass, Smart Searches, Without Google, PCWorld.com, February 2005, [www.pcworld.com/howto/article/0,aid,119623,00.asp](http://www.pcworld.com/howto/article/0,aid,119623,00.asp)
- [12] Google launches TV search service, BBC News, January 2005, [news.bbc.co.uk/1/hi/technology/4205267.stm](http://news.bbc.co.uk/1/hi/technology/4205267.stm)
- [13] Bill Burnham, Saving RSS: Why Meta-feeds will triumph over Tags, Burnham’s Beat, January 2005, [billburnham.blogs.com/burnhamsbeat/2005/01/saving\\_rss\\_why\\_.html](http://billburnham.blogs.com/burnhamsbeat/2005/01/saving_rss_why_.html)
- [14] Ravdeep Hora, Google fixes desktop search security flaw, CoolTechZone, December 2004, [www.cooltechzone.com/index.php?option=content&task=view&id=876&Itemid=0](http://www.cooltechzone.com/index.php?option=content&task=view&id=876&Itemid=0)
- [15] Bruce Schneier, Desktop Google Finds Holes, eWeek, November 2004, [www.eweek.com/article2/0,1759,1730748,00.asp](http://www.eweek.com/article2/0,1759,1730748,00.asp)

- 
- [16] Mike Ricciuti, The long march to Longhorn, News.com, September 2004, [news.com.com/The+long+march+to+Longhorn/20101016\\_35360695.html?part=rss&tag=5360695& subj=news.1016.20](http://news.com.com/The+long+march+to+Longhorn/20101016_35360695.html?part=rss&tag=5360695&subj=news.1016.20)
  - [17] Using Technology to Make Search Productive, Mondosoft Whitepaper, October 2004
  - [18] Tom Reamy, To Metadata or Not To Metadata, EContentMag.com, October 2004, [www.econtentmag.com/Articles/ArticleReader.aspx?ArticleID=7118](http://www.econtentmag.com/Articles/ArticleReader.aspx?ArticleID=7118)
  - [19] Martin Belam, Fine Tuning Your Enterprise Search – How To Get The Best Results To Your Users, 2004, <http://www.currybet.net/articles/finetune/index.shtml>
  - [20] Laura Ramos, Enterprise Search Vendors: Incumbents, Giga Research, September 2003 and Whit Andrews, Magic Quadrant for Enterprise Search, 2004, Gartner, May 2004, ID Number: M-22-7894
  - [21] Classification, Taxonomies and You, Verity White Paper
  - [22] Auto-Categorization: Coming to a Library or Intranet Near You!, Econtent, November 2002
  - [23] Taxonomy & content Classification, Market Milestone Report, Delphi Group White Paper, 2000
  - [24] E. Mäkelä et al., OntoViews – A Tool for Creating Semantic Web Portals, in S.A. McIlraith et al. (Eds.), International Semantic Web Conference 2004, Springer LNCS 3298, pp. 797-811, 2004
  - [25] Ruben Lara, Semantic Web Services discovery, DERI, [www.deri.at/teaching/seminars/internal/slides/SWSdiscovery.pdf](http://www.deri.at/teaching/seminars/internal/slides/SWSdiscovery.pdf)
  - [26] DIP Deliverable D4.8/4.17 SWS Discovery Module Specification, P2P & QoS Enabled Discovery Specification, Draft Version, Jun 9, 2005