

The Information Society Library
GETTING THE BEST OUT OF CYBERSPACE

FINDING INFORMATION IN CYBERSPACE

FROM IRRITATION TO INSPIRATION

Stefano Baldi • Eduardo Gelbstein • Jovan Kurbalija



P R E F A C E

There is no shortage of books on all matters relating to information management and information technology. This booklet adds to this large collection and attempts to do a number of things:

- To give non-technical readers an insight into the few principles that are important and reasonably stable;
- To present the material in a context relevant to the work of those involved in international relations;
- To make the reader curious enough to go beyond this booklet and investigate and experiment and thus develop knowledge which meets each reader's specific needs.

The format of these booklets and their contents evolved from courses given by the authors over the last few years in various environments and the feedback of the attendees. Readers' feedback on these booklets would be greatly appreciated by the authors so that future editions can be improved. The coordinates of the authors are given at the end of this booklet.

ISBN 99909-55-19-0

Published by DiploFoundation

Malta: c/o Mediterranean Academy of Diplomatic Studies, University of Malta
Msida, MSD 06, Malta

Switzerland: c/o Graduate Institute of International Studies,
rue de Lausanne 132 CH-1211 Genève 21, Switzerland

W-mail: diplo@diplomacy.edu

Web: <http://www.diplomacy.edu>

Cover Design by Nenad Dosen

Layout & prepress by Rudi Tusek

© Copyright 2003, Stefano Baldi, Eduardo Gelbstein and Jovan Kurbalija

C O N T E N T S

A short history of information.	5
Using information.	11
Defining information needs	18
Finding information	25
Validating information.	45
References.	49
Authors	51



SECTION



1

A short history of information

*What counts is not the technology
but what you can do with it.*

Anonymous

A SHORT HISTORY OF INFORMATION

Tempting as it is to launch straight into a discussion of Internet Search Engines, this may not be the best place to begin, and this short section discusses information in the context of humanity and its history.

Humanity has left its mark as a communicator for more than 25,000 years, as the paintings left in caves before the end of the last Ice Age prove.

By some 10,000 years ago, the domestication of animals and the cultivation of grains displaced the previous lifestyle of hunter-gatherers and soon after that humanity learned to produce food surpluses.



This turned out to lead to a new information need: that of recording the ownership of food surpluses, the rights of access to water, land and the solution, which emerged concurrently in the middle East and in China, was the invention of writing, some 5,000 years ago. This invention is one of the major landmarks in the development of the Information Age.

Writing needed to include words and numbers, and this created the need to create symbols for both tangible items as well as for abstract ideas. The book *The Story of Writing* (ref. 1) provides a very clear overview of how these concepts evolved around the world.



This idea of creating symbols led to many forms of creating written records such as the Egyptian and Mayan hieroglyphics, phonetic writing, where symbols represent sounds, and ideograms, where symbols represent concepts.

The history of information and communication is also the history of diplomacy. This very old profession emerged when it was discovered that it was better to hear the message than to eat the messenger.

In the electronic age, strings of ones and zeros are used as symbols to represent tangible and abstract “facts”. These strings are called **data**.

Data presented in a specific context becomes **information**. For example, to say “28” is just to mention a number.

When saying “the outside temperature is 28 degrees Centigrade” this number becomes information.

SHARING INFORMATION

Among the many interesting attributes of information, is the fact that it can be shared without destroying it or consuming it. In fact, this often leads to new information being created as is the case in science.

There are of course circumstances where sharing information is a bad thing, such as during examinations (then it’s called cheating) and when dealing with sensitive negotiations or personal matters where confidentiality and the ability to keep a secret adds more value to the information than sharing it.

Sharing information through symbols – regardless of whether they are written text or strings of ones and zeros – requires a special skill: that of literacy. Fortunately, this is easily acquired as all children going to school amply demonstrate.



For a major part of human history, information was a low speed item. Writing was an art form as well as a skill and even the most literate person needed time to record information on a clay tablet, a waxed surface, parchment and any other material available for this purpose.

Having recorded it, information could only be transported as fast as the physical transport that carried it – perhaps a runner, like in the battle of Marathon, or a rider as in the Pony Express, or a ship. In the early 1800s, news from the United States of America needed six weeks to reach Europe.

Human inventiveness looked for many ways to transport information faster, from smoke signals and jungle drums to the mechanical telegraph invented by the Frenchman Chappé around 1780. All of these required good clear weather and an unobstructed line of sight and their range was limited.

The next major landmark in the creation of the Information Age was the invention and practical deployment of the electric telegraph, first demonstrated by Samuel Morse in 1844. The history of the electrical telegraph and its impact are discussed in a short book entitled *The Victorian Internet* (ref. 2).

My God, this is the end of diplomacy!

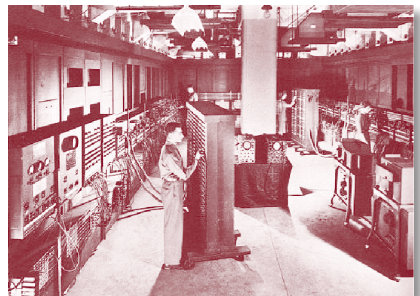
*Statement attributed to Lord Palmerston in the 1840s
when the first telegram was delivered to his desk.*

The dots and dashes of Morse's telegraph anticipated the ones and zeros of the digital age (binary digits or bits) and the telegraph was followed by an avalanche of invention which ranged from early mechanical computing machines (Babbage's Analytical Engine in the late 19th Century) to the punched cards used by Jacquard's machines for the manufacture of textiles and starting in the early 1940s, the development of electronic programmable computers.

In their earlier days, computers were singularly user-hostile and required a large amount of specialised knowledge to be able to do anything at all with them.

In the late 1960s, there was a substantial community of computer people researching science and working on defence projects in the United States and it was decided that they should be able to exchange information through some kind of network and that this network should be designed to survive a major war, including one where nuclear weapons were used.

This became known as the Defence Advanced Research Projects Network and by the early 1970s, its designers had formulated a set of definitions of how this



network should work called the Internet Protocol and the Transport Control Protocol, now known as TCP/IP. This network became the seed from which the Internet grew.

In the mid 1970s, this early Internet already had an effective e-mail service and many other features that are still in use, such as the File Transfer Protocol.

Personal computers first emerged in the late 1970s and were adopted by enthusiasts and, at first like many other inventions not taken seriously (Ken Olsen, the Chief Executive Officer of Digital Corporation, one of the most successful manufacturers of minicomputers since the 1960s, said in 1977 that he “could not see why anybody would want to have a computer at home”).

In 1981, IBM introduced a new design for a personal computer (PC), aimed at the corporate market and this move gave the PC respectability. It took some time for these PCs to be connected together into a network, a Local Area Network, but in general these local area networks were limited to a part of a building or, at best, a whole building and not beyond.

In 1989, Tim Berners Lee, then working at CERN, in Geneva (the European Centre for Nuclear Research) put forward a workable mechanism to electronically link documents which he called the “World Wide Web”.

It was only in 1993 when Marc Andriessen produced a piece of software that could be obtained free of charge called the MOSAIC Browser that the World Wide Web became near-intuitive to use and since then the number of websites has grown from a handful to a population of tens of millions.

The last ten years have also seen a major shift towards encoding all kinds of information in digital formats: not just text as word processors have done for years but also images, photographs, audio and video.

These digital formats make it easier for information to be labelled and catalogued (and therefore found) and therefore shared. It also makes it easier to modify, copy, forge and do things to information that their creators may not like. The issues of intellectual property and digital rights management are an important part of the cycle of using, creating and finding information.



SECTION



2

Using information

*Water, water, everywhere,
Nor any drop to drink.*

*“The Rime of the Ancient Mariner”
Samuel Taylor Coleridge*

USING INFORMATION

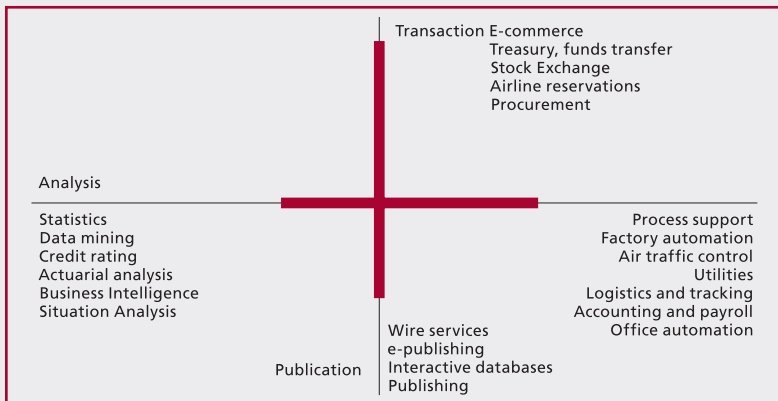
Many people now feel they suffer from an information overload – television, newspapers and magazines, reports, e-mail, voice-mail, cellphone SMS, and the knowledge that there are tens of millions of web-sites out there... and yet, there is this feeling that there is so much more information that would be useful to us if only..., hence the quotation from the Ancient Mariner’s poem.

Finnish sociologist Jaako Lehtonen refers to the permanent deficit between the growing availability of information and our limited capacity to process it as “Information Discrepancy”.

Anyone can add information. Therefore, the result of this information discrepancy is that more ends up being less.

Information in all its forms – documents, books, newsletters, e-mail, databases, reports, images and many more, is a valuable resource in all societies. So much so that it is now said that “quality information is to an organisation what healthy blood is to the body”.

Information is used in every human activity. This use can be grouped in four main categories: Transaction, Process support, Publication and Analysis:



The left part of the picture above, Analysis, will be the focus of this booklet, which will also include brief discussions on publication, in particular through websites, and transactions in the context of preparing analysis reports that are customised for a specific event.

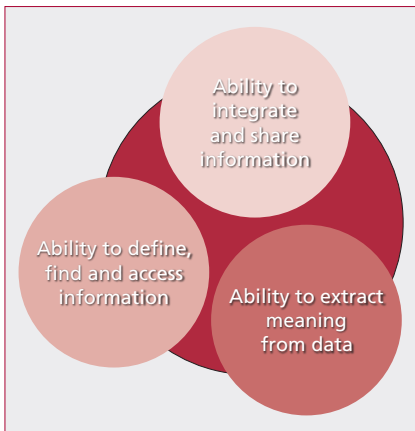
The right part (process support) and essentially the right hand corner of the diagram represents the world of automation: here technology is used to replace human effort and is particularly effective at dealing with well structured, repetitive tasks ranging from an airline seat reservation to a robotic manufacturing plant.

This field of computer-based automation represents the largest investments in information technologies and it turned out to have two important side effects that impact society:

- the loss of intellectual stimulation resulting from looking at a screen and processing transactions without having to think, which in extreme cases leads to the alienation of the workforce;
- the dependency of society on automated processes in every sphere of activity, including the operation of critical infrastructures and services. This has created an area of risk as a result of the vulnerability of widely networked computer systems to attack (information insecurity).

The activities listed under Analysis in the picture are just a few examples of what is now often called “knowledge work” and typical of the activities of people active in international affairs.

The individual skills and literacy of each person define how effectively information will be put to use – this is precisely the meaning of the statement at the beginning of Section 1, “what counts is not the technology but what you can do with it”.

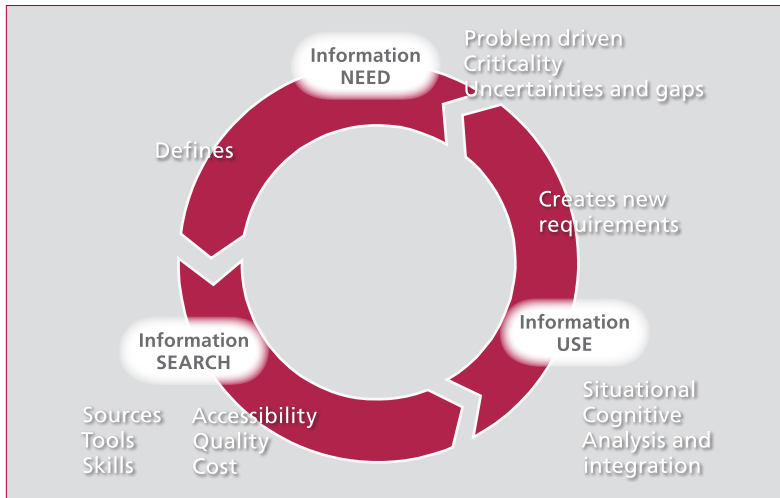


The concept of an Information Intelligence Quotient (Info IQ) first appeared under the name of “Corporate IQ” in an article published in the *Harvard Business Review* (ref. 3) and is highly appropriate to this discussion, which examines only one of the three abilities shown in the diagram, that of defining an information need, finding the information and accessing it.

The subjects of extracting meaning from data and integrating and sharing information are part of the science of Information Management and fall outside the scope of this booklet.

The *Harvard Business Review* article also introduces the concept of a complexity index, defined as the combination of the number of sources that need to be accessed to compile the information, the degree to which it is possible to determine the precise meaning of each component and the relationships that may exist between the various components.

Returning to the subject of using information, this needs to be seen as part of a cycle of activities with three distinct stages:



Knowledge workers deal with unstructured situations where last month's research and analysis usually do not apply to the problem in hand. Such challenges require a wide range of skills, such as for example, statistical analysis, risk analysis, the ability to organise thoughts to create a model of the situation under study as well as an appropriate level of background knowledge.

In using information in this way to deal with complex situations, there is an implicit assumption that there are many uncertainties concerning both the situation under study and as to the completeness, accuracy and appropriateness of the information available.

Information is a major weapon in the fight against uncertainty. Reduced uncertainty leads to better understanding and decisions. These in turn

lead to more effective action and therefore a better outcome than if things were left to chance.



Reduced uncertainty implies that the information collected is of appropriate quality for the intended use. The parameters that define the quality of information are discussed in the next section.



SECTION



3

Defining information needs

*The information you have is not the information you want.
The information you want is not the information you get.
The information you get is not the information you need.
The information you need is not available.*

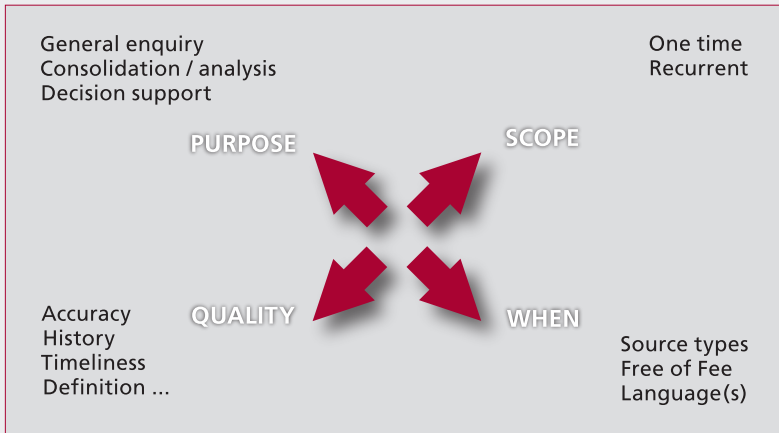
Anonymous

DEFINING INFORMATION NEEDS

The depth of knowledge that an individual has of the domain for which information is required will have a great influence in defining what specifically is being looked for.

For example, a casual need to look up an entry in an encyclopaedia is easily dealt with. The need to find the complete bibliography of a particular author is a more complex undertaking and compiling a comprehensive inventory of national legislation on cyber-crime is certainly not a trivial task – even for a knowledgeable lawyer practising in this field.

The figure below summarises the four areas of questions that assist in formulating an information need:



Purpose

What motivates the search for information?

Is it just curiosity or is this information going to support some kind of knowledge work – be it statistical analysis, a situation report, risk analysis or any other project of this kind?

Is this information going to support research that will lead to a report or publication?

Is this information going to play a critical role in making a decision? (And how critical is this decision going to be?)

The answer to these questions will give a first indication of the complexity index associated with the information need. A casual enquiry in a reputable encyclopaedia to satisfy one's curiosity will not require any further consideration and is quickly done.

Information to support research and an eventual publication needs to meet certain quality attributes (discussed below) and may be subject to restrictions about its further use (copyright for example). If the subject is important, it may be necessary to consult several sources to ensure consistency and cross-validation.

The information needed to support critical decisions must fulfil the strictest requirements concerning its validity. Anything else could prove to be a major embarrassment to the researcher, if not a Career Limiting Move.

Scope

It is essential to define the scope of an information search simply because humanity has accumulated an enormous amount since the creation of libraries and other repositories. The amount of information in electronic form is growing so fast that it is projected that in the next two years, the amount of information created since the invention of writing will double.

The following questions should be explored to define the scope of a search:

Is the information required likely to be unstructured (text with or without images) or structured (databases that can be queried for specific results)?

Is the information required in the private domain (minutes of meetings, auditors' reports within an organisation) or in the public domain (published annual accounts, reference books, newspapers and so on)?

If the information is in the public domain, is it available free of charge (by request or from a website), against a fee (for example the archives of the New York Times) or does it require one to be a subscriber to a particular service (for example Oxford Analytica).

What is the desired format of the information (printed, online, CD-ROM, ...)?

What are the languages in which this information may be presented?

Traditional Library culture	Hyperlinked culture
Based on classification <ol style="list-style-type: none"> stable hierarchically organised follows specific interests 	Based on diversification <ol style="list-style-type: none"> flexible single level allows all possible associations
Careful selection <ol style="list-style-type: none"> quality of editions authenticity of the text elimination of old versions 	Access to everything <ol style="list-style-type: none"> inclusiveness of editions availability of texts save "everything"
Permanent collections <ol style="list-style-type: none"> preservation of a fixed text browsing follows interest 	Dynamic collections <ol style="list-style-type: none"> intertextual evolution playful browsing

Quality

At this point, the information need has been formulated in enough detail to consider the next set of questions: how good does this information need to be? This definition may be a useful one:

“A quality information resource is one that meets the need of its end user.”

Easier said than done because the needs of the end user are many and not simple.

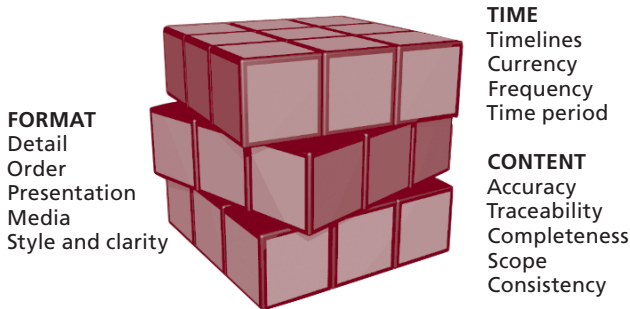
Today, there are two kinds of publications:

- Those that follow the traditional pattern of serious publishers, which include a process of formal acceptance of an abstract and a manuscript, peer review, approval by an editor, detailed editorial review and modifications to ensure clarity and consistency of style, assignment of copyrights and several other processes of this kind.

- Those brought forward by the internet culture where anyone can be a publisher. This freedom has resulted in a large volume of information which might have never been made available following the traditional publishing process used for books and paper publications, which is labour intensive and time consuming. It also operates as a filter that often discards much valuable work from ever being published.

The first approach does not constitute a guarantee of quality but provides a certain level of confidence in a review and editorial process, and it is usually not too difficult to ascertain the credibility and authority of both publisher and author.

The second approach puts greater demands on the end user to take steps to ensure that the information found is of appropriate quality for its intended use. The paragraphs that follow provide pointers of what these steps should include:



FORMAT

Quality indicators in this category include, in addition to those shown in the diagram:

- The degree to which the use of language and grammar is free of errors;
- The way in which complex issues are structured and articulated – badly composed information may raise more questions than it answers;
- The absence of mistakes in the content;

- The absence of incongruities – consistency is vital (it was once said that “the man with one watch knows the time. The man with two watches can never be sure.”)

TIME

The time elapsed between completing a manuscript and its final publication in the traditional world is measured in months, but on the internet is usually in days. Quality indicators regarding time include:

- The date of publication and the latest date covered by the information;
- The frequency with which this information is updated (if ever);
- The relevance of the time period covered by the information to the intended application.

CONTENT

Now that anyone can be a publisher on the internet, it is important to apply a number of tests to the information found to validate its content:

- Does the author/publisher have the credibility and authority to publish this material?
- Can this information be validated against other sources?
- Does the author/publisher provide a list of references?

When

When defining an information need, the consideration of whether it constitutes a one time need or a recurrent need influences the mechanisms and procedures used for finding the information.

Recurrent needs may require the subscription to one or more services or information sources. In addition, there are advantages in having a structured approach to such searches, for example documenting the sources used, the search methodology applied and the processes for transferring the findings to one’s desk or computer.

There are advantages in organising the access to frequently visited websites by using the web browser Bookmarks (in Netscape) or Favorites (in Internet Explorer) rather than relying on search engines time and time again.



SECTION



4

Finding information

*The seven most expensive words in life are:
“I didn’t know you could do that.”*

Anonymous

FINDING INFORMATION

There are just too many sources of information and not all information is available in electronic form. An attempt to conduct a search to identify all of these sources is likely to demand more time and effort than is practical to devote to this task.

This section briefly discusses the scope of information sources and concentrates on the tools and techniques most appropriate to search for information in the World Wide Web.

Some of the tools described, particularly search engines, are increasingly being adopted for internal information sources such as Intranets.

Information sources

INTERNAL TO AN ORGANISATION

Organisation-specific material consists of public domain information and, most likely the majority of it, in private domains with defined rules as to who may be entitled to access what information.

Internal sources in electronic form would typically include:

- Computer systems and databases;
- Workflow systems – including electronic mail and its archives;
- An Intranet – perhaps providing access to an electronic library and subscription services;
- Archives that increasingly provide access to documents in electronic form.

Internal sources in paper form:

- Circulars, reports, press releases;
- Minutes of meetings;
- Library services;
- Press reviews.

EXTERNAL

The following presents a sample of the sources of external information sources most likely to be used – it is by no means comprehensive:

- Internet portals;
- Internet sites (destinations);
 - Government departments
 - International and Regional Organisations
 - Non-Governmental organisations
 - Newspapers and journals
 - Reference sites
 - Services from commercial information providers
 - Vendors, industry associations, special interest groups.

In addition there are less structured information sources such as News-groups, Chatrooms and Blogs.

Delivery modes

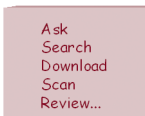
Our ancestors were hunter-gatherers – they spent much of their time looking for food. Once societies became settled roles became more specialised: some were producers and others were consumers.

The same is true today in the internet. It is possible to operate as a hunter-gatherer by navigating through dozens of websites, looking for pieces of information from multiple sources and collect them to construct something new out of them.

The techniques for pulling information will be discussed in the rest of this section.

There are benefits of having information delivered – in effect having it pushed to your computer as and when it is produced to meet specific requirements.

However, the push services have also been adopted by other commercial interests and our attention is diverted by pop-up windows, spam and junk mail.



Unsolicited advice
 Smart Agents
 E-mail alerts
 Service providers
 TV, radio, e-mail,
 Advertising...

The booklet on Computer and Data Hygiene will discuss techniques to deal with these unwanted sources – as far as it is possible to do so.

Pulling information

There are various types of structure that can be applied to pulling information from the World Wide Web.

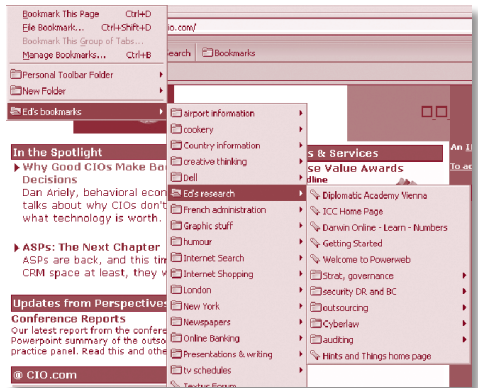
“Surfing” is found at the simplest level as it is virtually unstructured. It’s the equivalent of going out for a walk and instead of following a map, one follows whatever road appears interesting hoping that serendipity will lead to something useful.

While it may not be the most effective possible use of time, surfing has the attraction of potential discoveries and should not be ruled out as a method of understanding what is out there in the World Wide Web.

Bookmarks and Favorites have been briefly mentioned as means to quickly access websites which are regularly used.

The two most commonly used web browsers, Internet Explorer and Netscape include the facility of recording the address of any web page (“Bookmark This Page” in Netscape and “Add to Favorites” in Explorer) as well as the facilities to organise them in thematic folders. In Netscape this is done through the “Manage Bookmarks” option and in Explorer through “Organise Favorites”. Mastering this tool is easy and highly beneficial.

In reality, the websites of most interest to an individual are found either by the recommendation of someone else, by surfing or, more often, by structured searching.



Before entering the subject of searching the web, a word of warning: the World Wide Web needs to be thought of as consisting of two distinct levels:

- the surface web which is the one visible to the search engines commonly used;

- a deep web buried far down on dynamically generated sites that search engines cannot find and therefore require special tools.

The surface web contains several billion documents, growing at a rate of seven to ten million documents per day. However the search engines generally acknowledged to be most effective, for example Google, Fast and Northern Light, have only been able to index less than half these documents.

The deep web is thought to be at least 500 times larger than the surface web and is the fastest growing category of new information on the internet. It is estimated that 95 percent of its content is publicly accessible information without fees or subscriptions.

A brief section on the deep web follows the discussion of how to best use search engines.

Search Engines

Search engines create listings of websites in two ways: owners and authors of websites can submit their web pages for inclusion and, in addition, search engines use special software, (called “spiders”, “robots” or “bots”), to gather and index information.

Spiders go through the internet and collect web pages by following hypertext links to gather all the other pages that can be accessed through these hypertext links on each site.

The gathered pages are stored in large databases and indexed. Because of the growing integration of intranets and internet, it is possible to find pages not intended for public use which ended up being indexed because they were on a hypertext link or stored on the server or site. Where intranets have good security features you can expect to get a message that the page is not accessible.

A well written and more detailed description of how search engines for the World Wide Web work can be found at the following website:

<http://computer.howstuffworks.com/search-engine.htm/printable>

Is there a best search engine? Regrettably, this is not a question with a simple answer, as “it all depends” on the nature of the search and the expertise of the searcher.

The fallacy of abundance:

Don R. Swanson, an information retrieval pioneer, stated that “on a sufficiently large system...almost any query would retrieve some useful documents. The mistake is to think that just because you got some useful documents the information system is performing well.

What you don’t know is how many better documents the system missed.”

Quoted from *On the Internet* by Hubert Dreyfus, 2002 (Ref. 6).

The choice of a search engine will be influenced by the searcher’s expertise and preferences as well as by the nature of the search. A word of warning however: even the best search engines can only find between 20 to 30 percent of the relevant information. Search engines do not have “common sense” and, like all computers so far, have no capability to put information in context.

A search can be started and conducted with one or more of the following criteria:

Text and document searches by:

- subject matter;
- key word (title, author, publisher);
- word string (see tips and hints below);
- using Boolean operators (And, Or, Not);
- languages.

The importance of languages should not be underestimated. Search engines are mindless tools that do a particular job efficiently but have no common sense. They cannot recognise that the searcher may be interested in documents in various languages and therefore a set of searches – for example using key words in various languages - will lead to a much greater selection of documents than a search conducted in one language only.

Structured data searches:

- databases supporting SQL (structured query language or “sequel”);
- data mining and pattern recognition.

These are advanced subjects that require specialised skills on the part of the searcher and are therefore beyond the scope of this booklet.

Types of search engines

While there may not be universal agreement on how many types of search facilities there are, five categories can be readily identified. These are:

- Directories;
- Key word search engines;
- Metacrawlers;
- Portals;
- Other search possibilities.

Although computer experts are not known for expressing their spiritual preferences, the Observation Service for Internet (<http://www.ua-ambit.org/soi/soi.htm>), an initiative inspired by the Pontifical Council for Social Communications, has carried out research in different realms of the world of computers and cyberspace to discover the saint who best reflects the concerns and ideals of the experts.


The patron chosen by the pioneers of the new frontier of technology is Saint Isidore, who was born in Seville, Spain in 556 and wrote the "Etymologies", a type of dictionary. He is described as having been ahead of his time and constituted a cultural bridge between the Ancient and Medieval Ages.

DIRECTORIES

These are the World Wide Web equivalent of the traditional library catalogue system. A group of people and/or computer systems have created a hierarchical classification system which is easy to follow by almost anyone with a reasonably clear idea what they are looking for. Good examples of such directories are Yahoo and Northern Light.



Many of these directories are available in several languages – for example Yahoo Description searches.

Local Yahoo!s 

Europe	Asia Pacific	Americas
<ul style="list-style-type: none"> • Catalan • Denmark • France • Germany • Italy 	<ul style="list-style-type: none"> • Norway • Spain • Sweden • UK & Ireland 	<ul style="list-style-type: none"> • Japan • Korea • Singapore • Taiwan
	<ul style="list-style-type: none"> • Asia • Australia & NZ • China • Hong Kong • India 	<ul style="list-style-type: none"> • Argentina • Brazil • Canada
		<ul style="list-style-type: none"> • Mexico • U.S. in Chinese • U.S. in Spanish

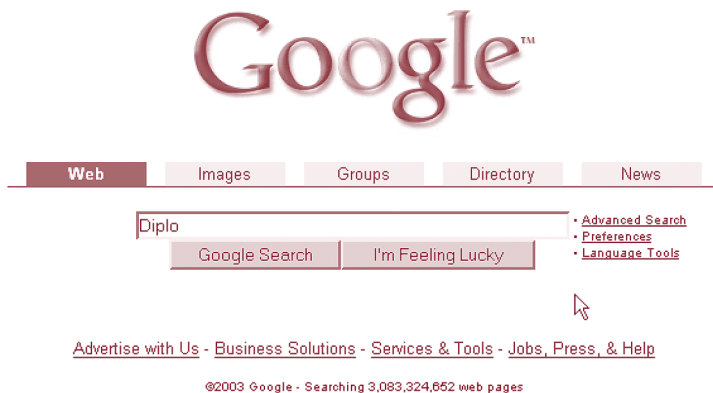
U.S. Cities: [Atlanta](#) - [Boston](#) - [Chicago](#) - [Houston](#) - [LA](#) - [NYC](#) - [SF Bay](#) - [Seattle](#) - [more...](#)

KEY WORD SEARCH ENGINES

These operate along the lines of telephone directory enquiries: the searcher defines the search and the engine provides answers. Alas, not one telephone number but often hundreds of possible sites – and this from the surface web.

The best way to reduce the number of answers to the search is to formulate the search in the tightest, least ambiguous manner. Some search engines, but not all, offer the facility to conduct a more detailed search on the results found in previous searches.

One of the best known and highly rated search engines of this kind is Google (<http://www.google.com>).



Google™

[Web](#) [Images](#) [Groups](#) [Directory](#) [News](#)

[Advanced Search](#)
[Preferences](#)
[Language Tools](#)

[Advertise with Us](#) - [Business Solutions](#) - [Services & Tools](#) - [Jobs, Press, & Help](#)

©2003 Google - Searching 3,083,324,652 web pages

Google is available in many countries around the world and will determine to which server to direct the searcher on the basis of his or her Internet Protocol (IP) address. Google has evolved considerably since it was first introduced and now offers additional features such as the ability to search for images, the inclusion of a Directory comparable to that of Yahoo and a structure to categorise discussion groups.

There are many search engines of this kind, many of them specialised by subject, such as:

<http://www.webwombat.com.au> (a search engine dedicated to all things Australian);

<http://www.searchedu.com> (a search engine for education domain web-sites);

<http://www.searchmil.com> (a search engine for pages dealing with military material);

and many more. A good index for search engines specialising in various topics, countries and languages can be found at <http://www.allsearchengines.com>.

AllSearchEngines.com
A very handy site — lists 'em all!

Search 

or click on these Popular Search Terms:

[Online Casino](#) [Hair Loss](#) [Weight Loss](#) [Viagra](#) [Business](#) [Gambling](#) [Employment](#) [Education](#)
[Home Business](#) [Management](#) [Web Hosting](#) [Diet](#) [Real Estate](#) [Cars](#) [Ebusiness](#) [Books](#)
[Womens Health](#) [Marketing](#) [Search Engine](#) [Golf](#) [Home](#) [Travel](#) [Make Money](#) [Computers](#)

or Search Engines: **Topic Search Engines:** **Sponsors**

METACRAWLERS

A metacrawler is a search engine that is linked to a number of other search engines, which aggregates the results and presents the consolidated results. This is effective when it is particularly important to cover as wide a range of surface web documents as possible.

An internet search using the word “metacrawler” will identify several such tools. One of the well established metacrawlers is Copernic.

copernic Discover. Compete. Profit. Home | Shopping Cart | Support | Contact Us | Language

Welcome to Copernic.
DEVELOPERS OF AWARD-WINNING SOFTWARE TO SEARCH, FIND, AND MANAGE INFORMATION.

Products:
 Copernic Agent
 Copernic Summarizer
 Copernic Enterprise Search

Online Store
Downloads
Support
Contact Us
Company Info

Information for:
Business Professionals

PRODUCTS

copernic AGENT VERSION 6 DESKTOP SOFTWARE

Copernic Agent takes Internet searching to a whole new level, and is available in three versions: *Basic*, *Personal* and *Professional*.

- > [Download Copernic Agent Basic FREE](#)
- > [Purchase Personal or Professional](#)
- > [What's New in Version 6.1](#)
- > [Take the Copernic Agent Product Tour](#)

[Product Home Page](#)

COMPANY NEWS

Copernic Agent v6.1 Update Released: Copernic released a new version of Copernic Agent including many improvements and corrections based on user feedback.
April 29, 2003

Copernic wins the Best Meta Search Engine Award for 2002.
February 4, 2003

OUR CLIENTS

Altria Group, Altria

Although the very basic version of Copernic can be downloaded from their website (<http://www.copernic.com>) free of charge, the full personal version costs around 30 US dollars, and there is a professional version costing 80 US dollars. Their designers also offer a product called the Summariser (60 US dollars) that produces brief summaries of any desired web page.

PORTALS

A portals is usually defined as a site featuring commonly used services and serving as a starting point and gateway to the World Wide Web or a specialised topic (vertical portal or “vortal”).

These are increasing in number and now cover a wide range of topics and levels of sophistication. As an example of a portal serving the international community, the website at <http://www.unsystem.org> provides a

comprehensive listing of all the official websites of organisations in the U.N. System.

English

Alphabetic Index
Thematic Index
UN System chart
UN System Highlights
About this Site
UN News Service
Missions to the UN
UN Information Centers
UNCAPS
DEPOLIB
Careers at the UN

Official WEB Site Locator for the
UNITED NATIONS
System of Organizations

Alphabetic Index of Websites of the United Nations System of Organizations

[Submit URL]
[Text Navigation]

Aa Bb Cc Dd Ee Ff Gg Hh Ii
Jj Kk Ll Mm Nn Oo Pp Qq Rr
Ss Tt Uu Vv Ww Xx Yy Zz

Aa Bb

- [ACC Network on Rural Development and Food Security](#) - Rome, Italy (e-mail)
- [ACC Subcommittee on Drug Control \(ACC/SDC\)](#) - Vienna, Austria (e-mail)
- [ACC Subcommittee on Nutrition \(ACC/SCN\)](#) - Geneva, Switzerland (e-mail)
- [ACC Subcommittee on Oceans and Coastal Areas \(ACC/SOCA\)](#) - Paris, France (e-mail)
- [ACC Subcommittee on Statistical Activities \(ACC/SSA\)](#) - New York, USA (e-mail)

OTHER SEARCH POSSIBILITIES

Commercial and subscriptions websites

The provision of specialised information has been a major business for many years. Financial information service providers such as Reuters and Bloomberg have held a strong position in the market well before the World Wide Web was invented.

The number of such services now available through the web is very large and the cost of their subscriptions reflects the market value of the information they provide and ranges from a few hundred dollars a year to tens of thousands per registered client.

Many online newspapers (such as the UK's Financial Times) no longer provide access to all of their material free of charge. Specialised sources such as Lexis-Nexis, the Economist Intelligence Unit, Oxford Analytica and many other, also operate in this mode.

Netcraft

If you would like to find out which are the government sites in a particular country, or how many sites a particular company or organisation has,

you can use tools offered at the website of Netcraft, a company specialising in security services and web statistics (<http://www.netcraft.com>).

The image shows two sections of the Netcraft website. The left section, titled "Webserver Search", features a search box with the placeholder text "What's that site running?...", a "Search" button, and examples: "Example: microsoft.com" and "Example: www.netcraft.com". Below this is an "RSS feed" section and a "Subscribe to Netcraft News" button. The right section, titled "Search Web by Domain", displays "Explore 40,195,574 web sites" and the date "13th May 2003". It includes a "Search:" dropdown menu with options: "site contains", "site starts with", "site ends with", and "subdomain matches". A search input field contains ".microsoft" and a "lookup!" button. Below the input field, a snippet of search results is visible: "site contains .microsoft." and "whats that ssl site running? | add your site".

A Netcraft query on “Site ends with” formulated as *.gov.py (the asterisk will be explained below under tips and hints), generated a list of 36 websites registered to the government of Paraguay.

The areas of knowledge needed to exploit Netcraft are therefore a knowledge of the domain names currently available and that of country codes. A complete list of both of them can be found at many locations, for example:

<http://www.norid.no/domenenavnbasert/domreg.html>

Converting IP addresses to names and locations and vice versa

The ability to convert from a domain name to its IP address or vice versa is useful. For example, a user of a European keyboard who needs to access a website with a domain name in Cyrillic, Chinese or Arabic characters would be able to do so by using the IP number to define the web address of the target website.

The image shows a screenshot of the HCIDATA website's "Convert Host/Domain Name to IP Address and vice versa" tool. The page has a navigation menu with links for "Home", "Company Information", "Request Information", "Prices", "Gifts", and "Site Map". The main content area features a search form with a ".CO.UK" logo on the left. The form has two rows: "Domain Name" with the value "localhost" and a "Find IP Address" button; "IP Address" with the value "127.0.0.1" and a "Find Host Name" button. Below the form, a note states: "This page can be used to find the IP of a host machine or domain name or find the name of one of the hosts at an IP address." At the bottom, there is a "W3C HTML 4.01" logo and a copyright notice: "Last Updated: 23-Aug-2001. WebMaster: webmaster@hcidata.co.uk © copyright 1998, 2003 HCI Data Ltd."

<http://www.hcidata.co.uk/host2ip.htm>

This facility is also useful in trying to determine the authenticity of an e-mail message given that its routing information will contain the IP address of the sender.

Sometimes it is also necessary or desirable to identify the physical location of either a website or of their visitors. For example Google uses this information to select the appropriate server and language to display its home page.

GEOBYTES [Home] [Contact Us] [Support] [Privacy] [Buy Now!] [New Services]

IP

IP Address Locator Tool

This IP Address Map lookup service is provided for FREE by Geobytes, Inc to assist you in locating the geographical location of an IP Address. [Click here](#) to checkout our other FREE IP-MAP localization services.

Are you responsible for generating visitors to your companies web site? Geobytes' free GeoDetection service generates over 10,000 visitors per day which for 1 cent per visitor will be visiting your companies web site. If your site could use some real traffic then please drop me an email at info@geobytes.com for more details. At 1 cent per click - it is an incredibly cost effective way to instantly build your traffic.

The following results were generated using GeoSelect version 11.

IP Address to locate:

Country Code	<input type="text" value="RU"/>	Country	<input type="text" value="Russsavia"/>	Distance to Nearby Cities	
Region Code	<input type="text" value="RUSR"/>	Region	<input type="text" value="Srbija (Serbia)"/>	0 0 Belgrade, SR, RU	
City Code	<input type="text" value="RUSRBLG"/>	City	<input type="text" value="Belgrade"/>	4 5 Zemun, SR, RU	
Cityid	<input type="text" value="5785"/>	Certainty	<input type="text" value="76"/>	13 8 Pančevo, VO, RU	
Latitude	<input type="text" value="44.8330"/>	Longitude	<input type="text" value="20.5000"/>	24 8 Banja Luka, SR, RU	
Capital City	<input type="text" value="Belgrade"/>	TimeZone	<input type="text" value="+01:00"/>		
Nationality Singular	<input type="text" value="Serbian"/>	Population	<input type="text" value="10477290"/>		
Nationality Plural	<input type="text" value="Serbs"/>	Is proxy	<input type="text" value="false"/>		
CIA Map Reference	<input type="text" value="Europe"/>	Currency	<input type="text" value="Yugoslavian Dinar"/>		
MapData Remaining	<input type="text" value="Free"/>	Currency Code	<input type="text" value="RUB"/>		

Check out Geobytes other products including:
[GeoSelect](#), [GeoSelectPro](#), [GeoSelect](#), [GeoSelectPro](#), [GeoSelect](#), [GeoSelectPro](#), [GeoSelect](#), [GeoSelectPro](#), [GeoSelect](#), [GeoSelectPro](#)

Search WHOIS data at:

[Click here](#) to find out why our data can differ from the WHOIS data.

<http://www.geobytes.com/iplocator.htm>

Geobytes (<http://www.geobytes.com>) is one of several websites that provides this conversion facility. Unfortunately the complete world map of IP numbers has not yet been finalised and it sometimes fails to give the needed information.

Virtual Library

The Virtual Library (VL) (<http://www.vlib.org>) is one of the oldest catalogues on the web. It was started by Tim Berners Lee, (creator of HTML and the web). Unlike many others, it is not a commercial service and is run by volunteer experts who compile pages of key links for their particular subject areas.

It can be considered an example of a specialised Directory of links. Each volunteer is responsible for the contents of his or her own pages, as long as they follow certain guidelines. The VL is not the largest index of the web, but some of its sections are of very high quality and can be extremely helpful.

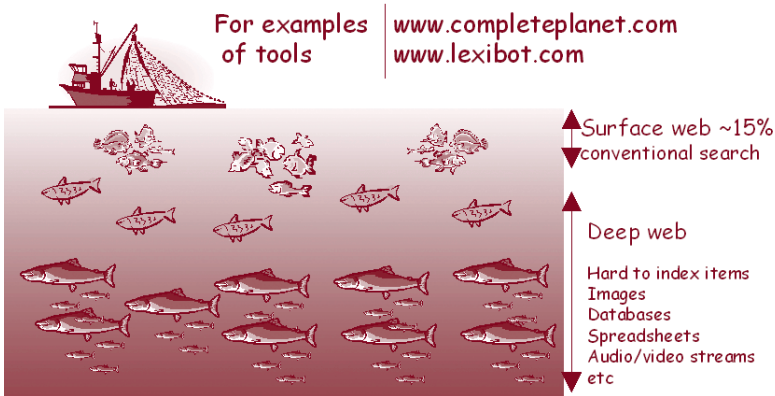
One particular example of such collections is the WWW Virtual Library on International Affairs Resources located at:

<http://users.etown.edu/s/SELCHewa/vl/>

This section of the VL offers some 2,600 annotated links on a range of international affairs topics. Sites are chosen for their long-term value, with preference given to those with cost-free, high-quality information and online analysis.

Searching the deep web

The deep web differs from the surface web in a fundamental way: content is stored in databases that are searchable and only produce results in response to a direct request. This is a complex and laborious process and special tools are used to create multiple and parallel direct queries and then to retrieve, classify and organise the findings.



A White Paper entitled “The Deep Web: Surfacing Hidden Value” (<http://www.brightplanet.com/deepcontent/tutorials/DeepWeb/index.asp>) provides a good tutorial on the deep web (ref. 4). *The Invisible Web: Uncovering Information Sources Search Engines Can’t See*, by Chris Sherman and Gary Price, is a good book on this topic (ref. 5).

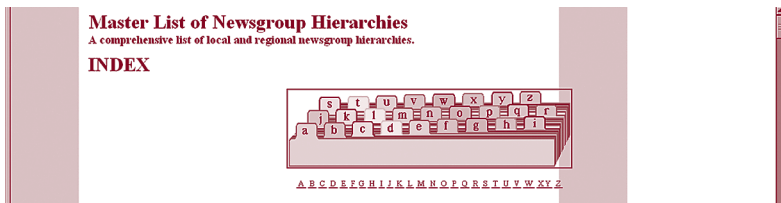
Newsgroups

A newsgroup is a discussion about a particular subject consisting of notes written to a central internet site and redistributed through Usenet, a worldwide network of news discussion groups. Usenet uses the Network News Transfer Protocol (NNTP).

Newsgroups are organised into subject hierarchies, with the first few letters of the newsgroup name indicating the major subject category and sub-categories represented by a subtopic name. Many subjects have multiple levels of subtopics.

This is an informal and anonymous forum of people with a common interest. There is no guarantee that the information posted and discussed in these newsgroups is reliable. They are however a very interesting option for exploring controversial and complex topics in a world where everyone can be an “expert”.

The website below (<http://www.magma.ca/~leisen/mlnh/mlnhtables.html>) contains a substantial listing of newsgroups, and there are several other such sites.



TIPS AND HINTS

There are a number of simple ways one can take advantage of the sophistication of search engines. The following may be useful in specific circumstances:

1. Using *inverted commas* to search for an exact phrase. For example “New York Times” will return only results with that exact group of words.

This is very effective when looking for very specific items, for example a biography of the U.N. Secretary-General:

“Kofi Annan” +biography

This search will return web pages that contain both the sequence “Kofi Annan” and the word biography.

2. Using the *asterisk* to define a word stem (also referred to as truncation). For example fem* will return results for female, females, feminine, feminist, feminists, feminism, etc., and thus greatly increase the number of results for a search. This technique should be used with care, for instance when the spelling of a word is uncertain.

3. *Boolean* functions allow the search to require, combine and exclude words or phrases in their results. They are probably the most powerful method to make searches highly focused and therefore successful.

The Boolean operators used are AND (+), NOT (-) and OR. Examples of their use:

“United Nations” + “digital divide”

would give results where both expressions between inverted commas appear in each of the returned results.

“United Nations” + “digital divide” – resolution

would give results where both the expressions between inverted commas appear in each the returned results but these will exclude any web pages that contain the word “resolution.”

The expression OR (no arithmetic symbol) in a search will result in a larger number of results returned as it will identify all the documents where either of the words appears.

“United Nations” OR peace OR operations

would give results with web pages referring to the United Nations, pages referring to peace and pages referring to operations even though these documents may not be related. This kind of search is probably the least valuable when a subject is well known.

4. *Searching by field* – is allowed by certain search engines by allowing the specification of where the terms for the search are located in the documents. This allows the search to be restricted to documents which contain the search terms, for example, in the title, thus giving a more relevant set of results.

The fields usually allowed in searching by field include: URL, Title, Domain/country, Link.

Searching by domain can be particularly valuable – for example by limiting the search to government websites, domain *.gov*. Not all search engines support this feature.

5. *Searching by date* – is also allowed by certain search engines and this limits search results to a range of dates. This is useful when looking for situations which are either new or where their history is known.

A slight problem – the “date” on a Web document may be one of many things, such as the date it was actually created, the date it was put on a Web server, the last time it was modified or the date it was added to the search engine’s database.

6. *Searching by file format* – not a universal feature in search engines, but when present it includes the following file formats: image files, such as *.jpg*, *.gif* and others, video, audio, *.mp3*, Acrobat (*.pdf*), Shockwave and Java.

7. *Search within results* - several search engines allow users to refine the results returned by means of a “Search within these results” feature. This allows a sequence of searches to be performed narrowing the results to a manageable number.

For example, a search for “United Nations” returns over one million results. Searching within these results for “Economic and Social Council” reduces the results to twenty five thousand. A further search within these results for “Sustainable development” results in two hundred results, at which point the researcher has the possibility of conducting a more detailed evaluation of these results.

8. *404 –page not found* - sometimes a search might lead to pages that have been removed from the web, giving the message “404 Not Found” instead of the wanted page.

It is possible that the page may have been transferred to another section of the same website, or that the document was renamed. It may be worthwhile to visit the parent site of the missing page by entering a truncated version of its address in the *Address* field (Internet Explorer) or *Location* (Netscape).

The summary tables that follow present a summary of the features offered by popular search engines in mid-May 2003. These tables confirm the view that there is no “best” search engine. Each engine has strengths and weaknesses and these should be taken into account before deciding which one to use.

SUMMARY TABLES OF THE MOST POPULAR SEARCH ENGINES

Search engine	URL address	Features
AltaVista	http://www.altavista.com	Easy to use and effective, with separate sections for searching images, newsgroups and news. Also has a section structured like an index (directory). It also provides the possibility of automatically translating foreign pages into English.
Excite	http://www.excite.com	Excite searches and returns results from a collection of search partners. The search engines it checks include Google, FAST, Ask Jeeves, and Inktomi
FAST	http://www.alltheweb.com	Good crawler-based search engine that provides comprehensive web coverage and outstanding relevancy.
Google	http://www.google.com	One of the best in scope, speed and effectiveness, with separate sections for searching images, newsgroups and news. Also has a section structured like an index (Directory). Cached links allow you to “resurrect” dead pages or see older versions of recently changed ones. It also provides the possibility of automatically translating foreign pages into English (not brilliant but useful).
Infoseek	http://www.infoseek.com	Allows both specific and general inquiries; the field of research may be progressively refined. Contains a section of maps.
Yahoo	http://www.yahoo.com	An excellent site that is both an index and a search engine. The main page allows one to use different search keys. The index gives a long list of available subjects.

	AltaVista	Fast	Google	Lycos	Yahoo
Boolean expressions	Yes	Yes	Yes	Yes	Yes
Phrases "..."	Yes	Yes	Yes	Yes	Yes
Word stem/truncation	Yes	No	No	No	Yes
Location in document	Yes	Yes	No	Yes	Yes
Date	Yes	No	No	No	Yes
Search within results	No	No	Yes	Yes	No
File type	Yes	Yes	No	Yes	No
Domain type	Yes	Yes	No	Yes	Yes



SECTION



5

Validating information

Caveat emptor.
(Buyer beware.)

Latin expression

VALIDATING INFORMATION

The preceding discussion shows that there is a vast amount of information available to us, much of it from the World Wide Web (and its deep web). Moreover, as anyone can become a publisher on the web, much of this material is of unknown quality and a good part of it is certainly of doubtful if not outright poor quality.

Some years ago there was a statement that poor information quality is like garbage on the information superhighway, and this proved to be correct. There are millions of websites that have been abandoned by their creators, not updated, not reviewed, with broken links and worst of all, with undated, unattributed documents which are unusable for any serious purpose.



Having found the information through the search mechanisms discussed here, how can this be validated?

Information science developed the concept of Metadata, a way to describe the attributes of data (equally applicable to information).

The 16th century map of Ireland shown in the figure illustrates how metadata is put to use in practice.

The labels in the map itself showing its title, date, publisher, orientation, etc., are essential items of information without which the map itself would not be meaningful or even recognizable.



In the case of documents retrieved through a search typical metadata that contributes to the determination of its quality would normally in-

clude: authorship, affiliation of the author, date of publication and other supporting references.

In a library or in a filing system, additional metadata would include appropriate references for the indexing, classification and physical location of the document.

Documents without such metadata should be regarded with the greatest scepticism and individuals deciding to use such documents need to recognise that they do so at their own risk.

Beware of unverified information: The Economist reported that in preparing his speech to parliament about the controversial issue of anti-AIDS drugs, the South African president asked his assistants to gather the most up-to-date information from the Web.

It appears that they relied on “snippets of negative information rather than peer-evaluated data of genuine scientific value” leading to a speech with factual errors. These errors included claims that several lawsuits had been filed in Western Europe against companies producing anti-AIDS drugs.

Copyrights

Information in digital form is easy to download, store, search, modify, copy and distribute. This may not be consistent with the copyright protection of such material and this should be validated, and the appropriate consent obtained, if this information will be used for anything else than personal use.

R E F E R E N C E S

1. Robinson, Andrew. *The Story of Writing* (Thames & Hudson, 1999).
2. Standage, Tom. *The Victorian Internet* (Berkley Pub Group, 1999).
3. Haeckel, Stephan, and Richard L. Nolan. "Managing by Wire," *Harvard Business Review* (September-October 1993).
4. Deep Content, a White Paper entitled "The Deep Web: Surfacing Hidden Value"
<http://www.brightplanet.com/deepcontent/tutorials/DeepWeb/index.asp>.
5. Sherman, Chris and Gary Price. *The Invisible Web: Uncovering Information Sources Search Engines Can't See* (Independent Publishers Group, 2001).
6. Dreyfus, Hubert. *On the Internet* (London: Routledge, 2002).

ABOUT THE AUTHORS

Ed Gelbstein

Eduardo Gelbstein is a Senior Special Fellow of the United Nations Institute for Training and Research (UNITAR) and a contributor to the United Nations Information and Telecommunications (ICT) Task Force and to the preparatory work for the World Summit on the Information Society. He is the former Director of the United Nations International Computing Centre.

In addition to his collaboration with the United Nations, he is a conference speaker and university lecturer reflecting his 40 years experience in the management of information technologies.

He has worked in Argentina, the Netherlands, the UK, Australia and after joining the United Nations in 1993, in Geneva (Switzerland) and New York (USA). He graduated as an electronics engineer from the University of Buenos Aires, Argentina in 1963 and holds a Master's degree from the Netherlands and a PhD from the UK.

ed.gelbstein@wanadoo.fr

Stefano Baldi

Stefano Baldi is a career diplomat in the Italian Ministry of Foreign Affairs, Counsellor at the Permanent Mission of Italy to the UN – New York. He has also served at the Permanent Mission of Italy to the International Organisations in Geneva, where he has developed several initiatives for the use of information technologies (IT) in the diplomatic community.

Baldi has an academic background in demography and international social issues. He also lectures on the use of internet for ministries of foreign affairs and missions at DiploFoundation's Postgraduate Diploma Course on Information Technology and Diplomacy. Baldi's most recent research focuses on the impact and future developments of information technology in international affairs.

<http://baldi.diplomacy.edu>
stefano.baldi@ties.itu.int

Jovan Kurbalija

Jovan Kurbalija is the founding director of DiploFoundation. He is a former diplomat with professional and academic background in international law, diplomacy and information technology. Since the late 1980s he has been involved in research on ICT and law. In 1992 he was in charge of establishing the first Unit for IT and Diplomacy at the Mediterranean Academy of Diplomatic Studies in Malta. After more than ten years of successful work in the field of training, research and publishing the Unit evolved in 2003 into DiploFoundation.

Jovan Kurbalija directs online learning courses on ICT and diplomacy and lectures in academic and training institutions in Switzerland, United States, Austria, United Kingdom, the Netherlands, and Malta.

The main areas of his research are: diplomacy and development of the international regime on the Internet, use of hypertext in diplomacy, online negotiations, and diplomatic law.

jovank@diplomacy.edu

NOTES