Introduction to General Internet Research: Search Strategies

Introduction

In previous units, you have been presented with a showcase of printed, dlocal digital and online resources. In the case of the online resources, you have been given the Internet addresses or **Uniform Resource Locators** (**URLs**). But what happens if you do not know the URL, or even what you are looking for?

The simple answer is to use of **search engines** and **subject directories**. Yet it is important to realise that, excellent though these may be when properly used, only a small portion of the World Wide Web is accessible through conventional search engines. What has been indexed is known as the **surface web** (also called the **visible web** or **indexable web**. General-purpose search engines such as Google do not have access to anything like the entire contents of the web. In fact, the **deep web** (also called the **invisible** or **hidden web**) is said to be several magnitudes larger than the surface web.

In fact, the best way to access those hidden pages and resources is to go directly to the site in question and using its own search engine. International organisations like the UN, World Bank or European Union, and multinational corporations like UBS, BP, General Motors or Microsoft have sites containing a wealth of information and linguistic resources for language professionals, much of it unindexed by general-purpose search engines.

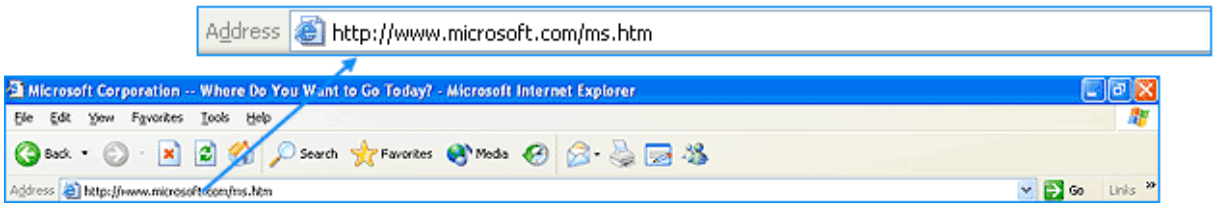
Just to try this out, go to the World Bank's site (http://www.worldbank.org/) and see what results you get when you enter the query "glossary" in its own search engine. Clearly, therefore, it is always worth consulting specific sites covering the particular field you are researching in addition to searching the web using Google and other common search engines.

In this unit, we will begin by considering how the structure of URLs can help us in our research. We will then look at the efficient and effective use of search engines and subject directories. Last but by no means least, we shall examine how to **evaluate** the documents, pages and sites which are found, a key component of instrumental competence that is central to developing effective research skills.

Uniform Resource Locators (Web Addresses)

Uniform Resource Locators, or URLs for short, are the addresses used to find one's way around the World Wide Web. If you do not know the URL of a particular organisation, you can always try to work it out or guess it. In fact, deducing addresses from the name of an organisation or institution can be a very efficient search strategy.

To do this, you have to know how URLs are structured.



In the simple example above,

- **http://** designates the communications **protocol** used for data transfer, in this case HyperText Transfer Protocol (*ftp://*, for instance, designates another protocol known as File Transfer Protocol)
- **www** is a widespread **convention** indicating World Wide Web addresses, but note that not all web addresses use this convention
- **microsoft.com** is the **domain name**: **microsoft.com** is the element sometimes known as the **second-level domain**, **com** alone is the component referred to as the **top-level domain** or **TLD** (**com** itself being a type of TLD called a **generic TLD** or **GTLD**)
- **ms.htm** is the name of a **document** at the domain **microsoft.com**, in this case a webpage (indicated by the suffix **htm** or **html**)

Now look at the following address:

http://www.zhaw.ch/~mssy/ToolsCourse/Site/Start.html

Here the URL is longer, because there has to be a **path** to the document being accessed - **Start.html**. That document is to be found in a directory with the name **ToolsCourse** - which is itself in another directory – **~mssy** – on the computer or server with the address **www.zhaw.ch**. In the domain name **zhaw.ch**, **ch** is a type of TLD known as a *country top-level domain* (as opposed to a GTLD like **com**). In all addresses, forward slashes - **/** - separate directories and documents from the domain name and one other, while dots - **.** - separate the components of a domain name.

What does this knowledge enable us to do? Top-level domains are fixed. Because of this, they can tell us about the activities of an organisation and where it originates or is based. Thus **ch** stands for Switzerland, **de** for Germany, **at** for Austria, **uk** for the United Kingdom, **au** for Australia, **ie** for Ireland, **ca** for Canada.

Various other domain extensions are sometimes placed immediately before country TLDs to indicate the kind of organisation that owns the site. In the UK, for instance, the TLD consists of two parts: companies in the UK use **co** - as in www.oup.**co.uk** -, universities in the UK use **ac** - as in www.oxford.**ac.uk** -, and government agencies in the UK use **gov** - as in www.numberten.**gov.uk**.

Webopedia.com supplies a list of domain extensions, including GTLDs and country TLDs (http://www.webopedia.com/quick_ref/topleveldomains/countrycodeA-E.asp).

Although there is also a country domain for the United States - **us** -, addresses there tend to end with generic top-level domain or GTLD extensions. The list of GTLDs is as follows:

com	Commercial institutions. Most companies use this TLD, and not just in the United States. It is the most widely used TLD.
mil	The domain for websites belonging to the US military.
net	Designates companies or organisations throughout the world which act as

network providers or administrate networks.

edu Originally intended to designate all educational institutions in the United States, registrations are now restricted to four-year colleges and universities. Schools and two-year colleges are now registered in the country domain **us**.

gov Reserved for agencies of the United States federal government. The Library of Congress also carries this TLD.

org Used by organisations of various kinds, most notably international entities such as the United Nations or the World Bank.

int Intended for organisations established by international agreements. Not many websites have this TLD, important exceptions being NATO, the European Union and the International Telecommunications Union.

A number of new GTLDs have recently been authorised, and will be seen increasingly on the web. These are: **aero**, for the aviation industry; **biz**, for businesses; **coop**, for cooperatives; **info**, for unrestricted use; **museum**, for museums; **pro**, for professionals such as lawyers and doctors. A final new TLD is **name**, as in john.smith.name, which may be registered by any individual.

Beyond the use of TLDs, groups of institutions with a common purpose may well give themselves domain names that are structured in a similar way. Thus in Germany, all *Fachhochschulen* begin their names with **fh-**, followed by the name of the location (e.g. www.**fh-magdeburg.de**), while all universities use **uni-** in the same way (e.g. www.**uni-hamburg.de**). In Switzerland, certain universities also have similarly structured domain names, for example: www.**unizh.ch** for Zurich, www.**unibe.ch** for Berne, or www.**unige.ch** for Geneva.

Finally, the path of a URL can also contain useful information for us. Take the following example from the site of the bank UBS:

http://www.ubs.com/1/e/index.html

The section **/e/index.html** informs us that the index.html is in a directory called **e**. It seems likely that **e** stands for English. Now let's suppose a translator were looking for the German version of this document; and let's also suppose that no direct link were provided to it from the page **index.html**. The quickest way would be to replace the **e** in the path with a **g**:

http://www.ubs.com/1/g/index.html

This is an effective search strategy for finding parallel texts.

Recognising the path of a URL can also help when you get a "404 Page Not Found" error message. Just cut back one or two directories, or to the domain name, and try to trace the document from there.

So, understanding the structure of URLs has two major advantages. Firstly, it helps us to identify the origin of a page or site on the web. And, more importantly for research, it allows us to deduce the addresses and site structures of organisations, thus saving us from having to resort to a search engine. This can considerably speed up search tasks on the web, since all you need to do is use the browser's address bar.

If you do not know, or cannot deduce, the URL of a site, or if you are looking for information but have no idea what site it can be found on, you will obviously have to use the appropriate **search tools**. This unit briefly considers the functionality of the two main types, namely **search engines** and **subject directories**.

There are two basic forms of search engine. **Individual search engines** index and archive the contents of pages and other documents on the web. When you enter search terms in an engine, or use an engine's subject directory, the engine searches its database of indexed documents and then presents you with a list of matches. This usually takes no more than a second or two. **Meta search engines** simultaneously search the indexes of multiple search engines up to a certain cut-off point. This has two important consequences: meta search engines return only a portion of the documents available directly through the individual search engines; the results retrieved can often be highly relevant, since they usually take the first items from the ranked lists of the individual engines being searched.

Of course, there are dozens of search engines to choose from. A list can be found at http://www.internettutorials.net/engines.html. No single engine has indexed the entire web, and each engine has its own indexing method. Some index the entire page, others only part of the page. The larger the index, the more likely the search engine will be a comprehensive record of the web, which is especially useful for those looking for obscure material. The latter applies equally to meta search engines, which are best employed when you have an obscure topic and your search is not complex.

You can find out more about the size and popularity of search engines today by going to *Search Engine Watch* at http://searchenginewatch.com/. Currently, comScore (http://searchengineland.com/080718-183129.php) puts Google (http://www.google.com/) well ahead of Yahoo! (http://search.yahoo.com/) and MSN (http://www.msn.com/) as the engines most used by surfers. Reports on search engine sizes (based on the number of pages indexed) again show Google to be out in front.

But size is not everything, especially since search engines do not necessarily index the same pages. Other general factors in determining the usefulness of an engine are:

- the quality of the search facilities, such as the ability to use operators and wildcards (already introduced in the previous unit "Introduction to Lexicographical Research: Monolingual Resources in English"), the additional features offered in the hit lists etc.
- the regularity with which indexes are updated
- the response time to search queries

So it is always worth trying out more search engines when you are not satisfied with the results yielded by the first one you use. Another point to bear in mind is that **regional versions** of search engines may contain more local information than international ones.

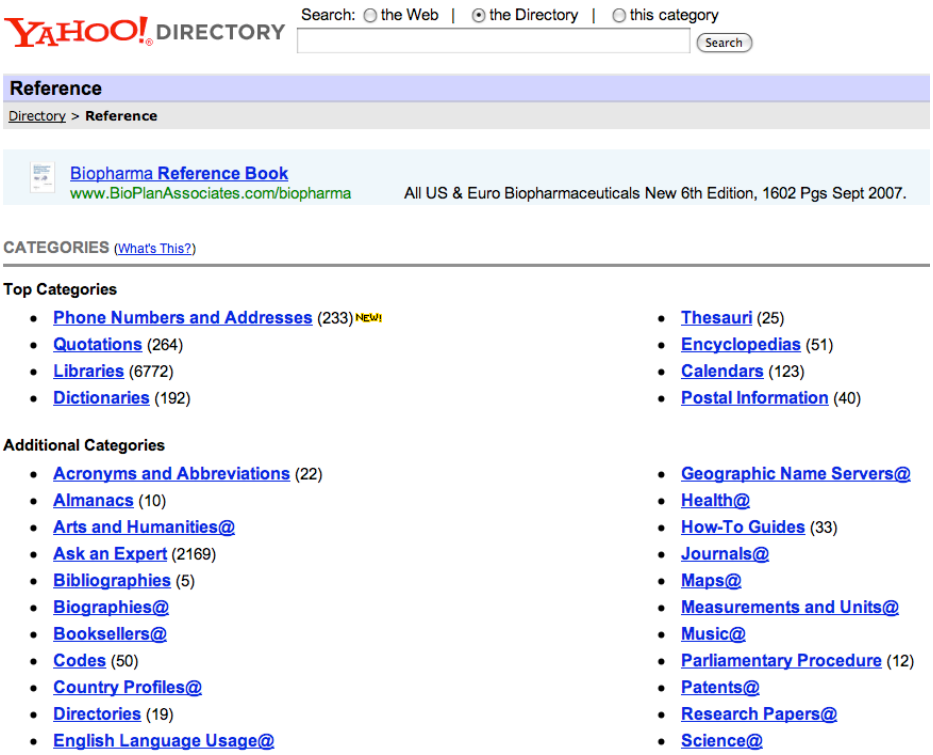
Subject Directory Searches

The basic tool for systematically finding information on the web is the **subject directory**. A directory enables users to look for information by selecting thematic categories and sub-categories in catalogues or subject-trees, and is an effective way of finding very specific information from reliable sources. Subject directories can be divided into **commercial portals** and **academic** or **professional directories**.

The best-known and most popular commercial portal is Yahoo! (http://dir.yahoo.com/). Unlike search engines that are based on automatically generated indexes of sites and pages, Yahoo! uses people to find and categorise information. This slows down the indexing process, but Yahoo! makes up for its relatively small database by cooperating with other search engines like Google. And human indexing of content produces a better quality of search result.

You search the Yahoo! directory by following links to progressively narrow down the field in which a search is conducted. Once you reach the (sub-)category you require, you can either

browse the site listings or enter a term to search either only within the Yahoo! sub-category or across the entire web. Of course, you can also click on further links to even narrower sub-categories. For example, you can see a full listing of general sites related to reference works by visiting Yahoo!'s Reference category (in the left-hand column):



In addition to its international website, Yahoo! offers localised versions of its directory in many languages and with regional content. The list of local Yahoo! international sites can be seen at http://world.yahoo.com/.

However, for subject-specific academic, scientific and professional research it is best to turn to academic and professional directories. These are often created and maintained by subject experts to support the needs of researchers and professionals in their fields. Theses include the *Internet Public Library* at http://www.ipl.org/ run by the University of Michigan. Two other very fine examples are *Infomine*, http://infomine.ucr.edu/, maintained by the University of California Library in conjunction with other university libraries, and the much respected *WWW Virtual Library* at http://vlib.org/.

A list of subject directories can be found at http://www.internettutorials.net/subject.html.

Term Searches in Search Engines

Searching by directory or subject tree is a very efficient way of finding specific information among the immeasurable volume of pages and sites on the web. Directory searches are also very helpful in zoning in on sources which are very reliable. But directories are not the right tools for a comprehensive search of web content.

Indeed, no search tool can list all the information contained in all web documents. However, search engines like Google are capable of finding far more documents related to a topic, word or phrase than a directory like Yahoo!

The searches conducted by search engines are based on words or characters. The engine looks for a string of characters listed in its index which matches the string of characters in your search query. It then lists all the hits or matches for that query, and provides you with a

link to the web pages in question.

You can enter search queries in two different ways, depending on whether you want to perform a simple basic search or a more complex advanced search. Search engines offer different interfaces for the two types of word search. An advanced search gives you a greater range of search options, although basic searches do offer powerful facilities in themselves.

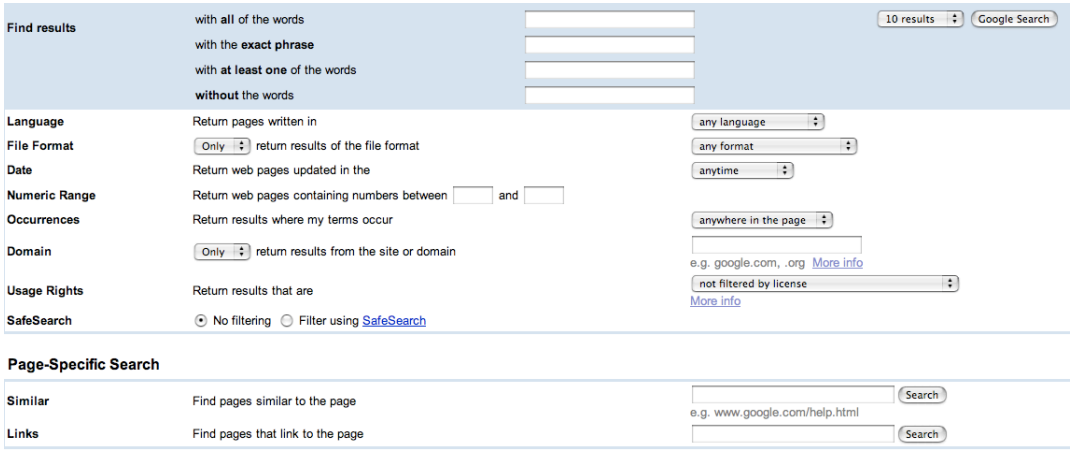
Searching with Google

Google's fine Help file at http://www.google.com/support/?ctx=web and the Google Guide at http://www.googleguide.com/intro.html present tips and information on how best to use Google's features.

First of all, Google's basic search engine interface automatically interprets gaps between words as the operator **AND**, and only returns pages that include all search terms in a string. The order in which the terms are typed will affect the search results. To restrict a search further, just include more terms. Google ignores common words and characters such as "where" and "how", as well as certain single digits and single letters, because they tend to slow down a search. Google searches are not case-sensitive, i.e. capitals are not recognised.

If a common word is essential, however, you can include it by putting a **+** sign in front of it. Another method for doing this is conducting a phrase search, which simply means putting quotation marks around two or more words. Google also supports the operators **-** (signifying **NOT**) and **OR**, and an extensive set of advanced operators (see the Help file at http://www.google.com/help/operators.html and the Google Guide at http://www.googleguide.com/advanced_operators.html). Word wildcards are recognised by Google. A quick reference "Cheat Sheet" for advanced operators, taken from the Google Guide (http://www.googleguide.com/advanced_operators_reference.html), is supplied in the **appendix** to this unit.

The Google advanced search facilities offer additional search parameters - such as date - and a user-friendlier layout for complex searches (for information on how to use this interface, see http://www.googleguide.com/sharpening_queries.html).



Some of the most attractive features of Google are seen in the presentation of results (see http://www.google.com/help/interpret.html). For instance, Google users can call up an older, stored version of a page if that page has been changed since it was indexed or no longer exists (by clicking on "Cached") or call up pages with similar content (by clicking on "Similar pages"). They can also get definitions and other linguistic information related to a search term by using the link above and to the right of the list of search results:

Results **1** - **10** of about **13,700,000** for **entropy** [definition]. (**0.04** seconds)

Here is a short list of the some **useful basic search operators** for a quick search in Google:

word +word word word	Finds pages on which both words occur (putting just a space between words will have the same effect as using the +)
word -word	Finds pages on which the first but not the second word occurs
word OR word	Finds pages on which the first or the second word occurs
"word word word word"	Finds pages on which exactly this phrase occurs
~word	Finds the word and its synonyms
*	Denotes a wildcard (only word wildcards are currently supported)
site: (e.g. site:ac.uk)	Restricts search to certain domains (e.g. UK university sites)
define: (e.g. define:momentum)	Locates definitions of terms, encyclopedia entries etc.
related: (e.g. related:www.linguistlist.org)	Finds sites related to the URL (e.g. www.linguistlist.org)
filetype: (e.g. filetype:pdf)	Searches only for this type of file

Finally, beyond its primary search interface, Google offers a number of **specialised search services**. The most useful for our purposes are:

news.google.com (local site: news.google.ch)	Searches news resources in various languages (scroll to bottom of page for international versions)
directory.google.com (local site: directory.google.ch)	Directory search
books.google.com (local site: books.google.ch)	Full-text search in books
www.google.com/universities.html	Search portal for US university sites
scholar.google.com	Searches scientific and academic literature

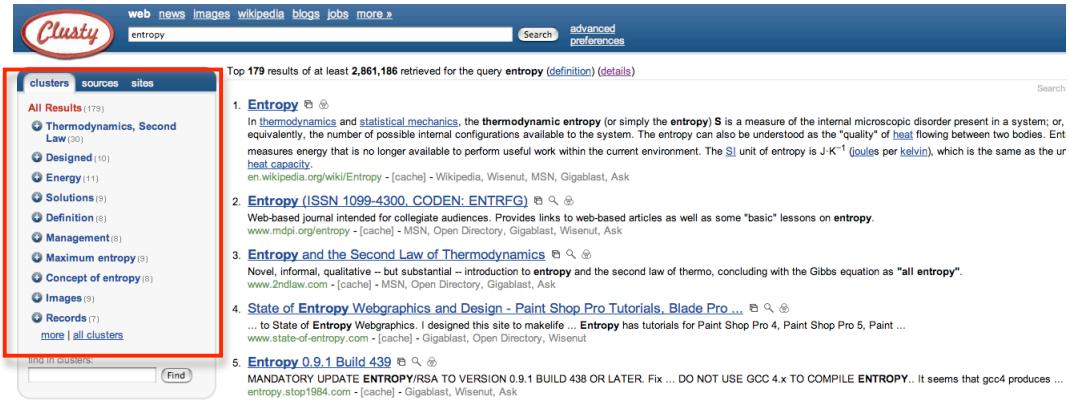
Choosing the Right Search Tools

Yet Google is just one of many search tools available, and you need to choose the right one for the research you must do. An excellent introduction to selecting tools can be seen in the table at http://www.internettutorials.net/started.html. A more specific tabular guide can be viewed at http://www.internettutorials.net/choose.html. The following provides some general tips, based on the tutorial at http://www.internettutorials.net/checklist.html.

1. Use academic or professional directories rather than commercial portals for academic, scientific or professional research.

2. Be aware of how a search engine returns its results. In common with many second generation search engines, Google ranks by the number of links to a page from other pages which it ranks high (i.e. on the same principle of linking that we have adopted when determining the reliability of sources). However, other engines rank by other criteria, such as topic (http://www.ask.com/), or will actually sort results into

categories or *clusters* related to concepts derived from your search (http://clusty.com/):



The advantages of the latter are self-evident.

3. You will often find that multiple pages are returned from a single site because they all contain your search terms. This gives a distorted view of the number of actual hits you get and can be confusing. Some search engines, such as AltaVista (http://www.altavista.com/) and AlltheWeb (http://alltheweb.com/) avoid this by grouping their results, putting all the returns from one site together into a single hit. You are then given the chance to view all the retrieved pages if you want to. You may therefore get a smaller number of results, but each result comes from a different site.

4. Use meta search engines when you have an obscure topic, your search is not complex, you have no success with individual search engines and you want to retrieve a smaller number of results. The latter is because, as we have already mentioned, meta search engines return only a portion of the documents indexed by individual engines, usually those coming first on their ranked lists.

5. Do not use search engines to look for the latest news, since it takes time for their spiders to index pages and they rarely contain the most recent documents posted in the Internet. Instead, use specific sites such as those listed in the previous unit.

Evaluating Information in the Web

Once language professionals have found the information and documents they need for their research, they have to evaluate the reliability of the information they contain. The web is huge and anarchic, and the quality of information on it can vary considerably. It is therefore very important for translators to develop their own evaluation strategies.
Frank Austermühl* lists four principal aspects of any evaluation:

- find out how many links there are to the document by entering the operator **link:** followed by the document's URL in Alta Vista or Google (e.g. **link:www.zhaw.ch**); then consider the number and status of the organisations linking to the document
- evaluate the author's credentials, affiliation with organisations and motivation in publishing the document
- establish where the document has been published, the status of the publishing organisation and the quality of any bibliographies or links
- check factual data in the document against proven reliable sources

* Austermühl, Frank, 2001, *Electronic Tools for Translators*. Manchester: St. Jerome, 64 ff.

To this we can add an important point relating to the quality of language. The web contains a huge number of texts - especially English texts - created by non-native speakers. So use the language and regional filters provided by search engines, your knowledge of country TLDs, and any available biographical information on authors (sometimes even a name can be a good guide) to increase the probability that documents you use for linguistic research have actually been written by native speakers.

The quality of a site may also be gauged by its popularity and/or history. As we have heard and seen in the lecture series, traffic and popularity rankings for sites can be established by systematically using a resource like *alexa.com*, at http://www.alexa.com/, while resources like *WayBackMachine*, at http://www.archive.org/web/web.php, can be useful for tracing the history of a site.

The more detailed your research becomes, the more important it is to verify your sources adequately. This may require you doing even more to check their validity. An expanded and more systematic guide, developed by the University of California, Berkeley (UC Berkeley), can be found at http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/Evaluate.html. UC Berkeley also supplies a brief but very useful checklist for rapidly assessing web pages and documents. The tutorial and checklist are provided in the appendix to this unit.

Language professionals rarely have time to thoroughly evaluate information in the ways suggested here; but regularly applying one or two of these criteria can be a good way of building up a collection of trusted resources.

Summary

This unit has presented the main aspects of effective Internet research and provided a basic introduction to the principal features of web search facilities. Concentrating on the Yahoo! directory search and Google's simple, advanced and specialised search services, it has considered ways in which professionals can optimise research strategies and techniques by using the right tools, syntax and operators. As information literacy vitally depends on using the right resources for the purpose at hand, stress has been placed on techniques and methods of evaluation.