

## Course 2: The Internet and Biological Research

The internet is gradually revolutionizing researchers' practices, both in terms of their literature searches and the dissemination of their work. While the opportunity to access a colossal amount of information "with a simple click" is a particularly attractive prospect for any biology researcher, discovering relevant information in this "ocean" is actually quite difficult, despite the growing number of search tools available to internet users.

The number of web pages accessible through various information search tools is constantly increasing. This abundance of information available on the web is, of course, beneficial to users, as the increased size of the database queried raises the likelihood of finding relevant information on a given topic.

The way internet search tools work can be classified into three categories:

- search engines (or robots) whose objective is to index web pages in "full text", without human intervention and without associated quality criteria, in a database (also called an index) in which users can perform keyword searches, *via* a specific interface.
- meta-search engines, whose principle is to simultaneously query, *via* a single interface (website or software installed on a client computer), a greater or lesser number of different search engines.
- Directories (or directories) are characterized by their hierarchical organization and the ability to search by browsing through different categories. Unlike web pages indexed by search engines, the pages listed in directories are selected based on quality criteria and are not indexed in full text; they are descriptive entries created by publishers.

Furthermore, the knowledge areas covered by search engines and directories vary. A distinction is made between general tools, which cover all knowledge areas, and selective or specific tools that cover a given scientific or disciplinary field.

### 2. Research tools studied

Because the number of information search tools available on the internet is very high, for reasons of feasibility, only a limited number of 8 tools were included in this study: 5 search engines, 1 meta-search engine and 2 directories.

## 2.1. Search Engines

- *Altavista* ( <http://www.altavista.com> ) is, probably due to its age (1995), one of the most popular and frequently tested general-purpose search engines.
- *Google* ( <http://www.google.com> ), which appeared quite recently (1998), is the general search engine that currently has the largest index in terms of the number of indexed pages (more than 1.5 billion).
- *Bioview* ( <http://www.bioview.com> ) is a search engine specifically dedicated to biologists with an index composed of web pages selected only in the field of biology.
- *Scirus* ( <http://www.scirus.com> ) allows users to query the general index of *Fast* ( <http://www.fast.com> ), which is limited to web pages with scientific content. It also allows users to query four Elsevier databases, access to which requires a subscription. This latter option was not used for this study because the information provided is not freely accessible.
- *Search4science* ( <http://www.search4science.com> ) is a search engine specifically dedicated to finding information in science. It offers the possibility of searching in two different indexes: (1) the *Northern Light index when the user chooses the Dynamic search option Search* . This interface offers a series of synonyms for each search term entered by the user, allowing them to broaden or narrow their search. Its unique feature is that it groups results into directories to facilitate their use; (2) *Google 's search results when the user chooses the Direct search option* . This latter option is very similar to a direct query of *Google* via its own interface; only the *Dynamic option differs*. The search has been evaluated.

## 2.2. Meta-engine

- *Copernic* ( <http://www.copernic.com> ) is a client-side metasearch engine , meaning its use requires prior installation of software on the user's computer. Two versions are available: a paid version offering advanced search capabilities by domain or type of information sought and primarily aimed at information professionals, and a free version allowing simultaneous queries of 10 general search engines. Since the vast majority of users in biology do not have access to the paid version, only the free version was tested in this study.

### 2.3. Directories

- *Infomine* ( <http://infomine.ucr.edu/Main.html> ) is a specialized directory developed by a network of Californian university libraries. It lists approximately 20,000 authoritative websites in the sciences for an audience of academics and researchers.
- *The Open Directory Project (ODP)* ( <http://www.dmoz.org> ) is a general directory whose web page selection and indexing are carried out by volunteer editors, each a specialist in their field. It is one of the general directories with the largest index size.