Goals for this session

Understand what the “Internet” is and the range of information that is available

Learn how to navigate the “World Wide Web” (“web”)

Learn how to use search engines, portals, and online databases

Learn how to formulate searches to produce optimal results
Using the Internet is a bit like eating an elephant—where do you take the first bite?

- Internet Basics
- Navigation
- Search
- Portals
- Databases

The “Internet” is a network of publicly accessible computers that communicate with each other through established software protocols.

- “Server”
  - Computer connected to the Internet network, storing information in electronic files for remote clients

- “Client”
  - Personal computer browser software

HTTP
FTP
Email
Telnet
The World Wide Web (web) is that part of the Internet that contains multimedia information.

Files contain text, graphics, audio, video, databases, interactives.

Files are transferred from server to client via hypertext transfer protocol (http).

Some, but not all, web sites are indicated by ‘www’ in the address.

Anyone with the necessary hardware (personal computer), software (browser), and Internet connection can use the web.

The web has no central authority to regulate content.

Information is user-controlled—anyone with access can put anything on a server that is part of the web—no “peer review”

Determining the quality and accuracy of web information is the responsibility of the user.

Just because you found something on the web doesn’t mean it’s true or accurate—users must be skeptical and cautious in evaluating web resources.
The amount of information on the web is staggering, which poses special problems for web users

In 2003, 45 million servers were connected

Google claims to have indexed >3 billion pages

>20 000 scientific journals are available on the web in full or in part

Effective search strategies can compensate for limited access to PCs; save bandwidth and money

There are four main methods to find what you’re looking for

“Surfing”—time-consuming, often ineffective, hard to remember what you’ve looked at and where you’ve been

Search engines—huge databases of indexed information; some are better than others

Site-specific searches—useful to search within sites for desired information; faster than searching the whole web

Portals—information gateways; narrowly focused, quality controlled, and targeted for specific audiences
Search engine or portal? Think of a book…

A portal is the table of contents

A search engine is the index

First, learn how to read “URLs”

http://www.physics.uiuc.edu

http:// indicates a web address

www.physics identifies a specific server

uiuc is the “domain” (the organization)

edu identifies one of the main U.S. nets

edu = a university or educational organization

http://www.harvard.edu = Harvard University

org = a not-for-profit organization

http://www.acs.org = American Chemical Society

mil = a U.S. military organization

http://www.nrl.navy.mil = U.S. Naval Research Lab

gov = a U.S. government organization

http://www.lanl.gov = Los Alamos National Lab

com = a commercial firm or company

http://www.adobe.com = Adobe Software

/xxx/yyyy/zz after the “net” are folder names
Non-U.S. URLs are different

The first part is the same

http://server.domain…

Next, 2-letter country code (vs. 3-letter net)

uk = United Kingdom  fr = France

jp = Japan  de = Germany

kr = Korea  il = Israel

us = United States—new net for smaller orgs

Some countries indicate universities by ac (“academic”) following the domain name

http://www.ox.ac.uk = Oxford University

Japan uses go for a government agency

http://www.jaeri.go.jp

Many URLs are intuitive; let’s see what you’ve learned…

http://www.berkeley.edu
http://chemistry.caltech.edu/

http://www.asbmb.org
http://www.llnl.gov

http://www.pasteur.fr/recherche/banques/

- **Sequences databases** maintained at the Pasteur Institute
- ARCSSF databases - ABC systems. Information on Sequence Structure and Evolution
- Base de données de génotoxicité
- CARDB
- Cobri, genome browser for Escherichia coli
- Collection de l'Institut Pasteur
- CROBA_basse de données Arbovirus
- Dictionnaires (en travaux)
- Encyclopédia nicotinica (en travaux)
- Gendolist, mother of all genomic browsers,
- Genome de Acanthamoeba castellanii
- Leprana, genome browser for Mycobacterium leprae
- Ligand-Gated Ion Channel subunits database
- LibriGen, multi-genome browser for the two genomes of *Listeria monocytogenes* and *Listeria innocua*
- MvDB: An integrated *Mycobacterial* database
Search engines (SEs) are good for…
Finding organizations and people
Finding very specific information
Doing quick and extensive searches
Locating images and video

…but they’re only as good as your choice of keywords and search strategy

In addition to general SEs, a variety of specialized SEs are available
Multi-engine searches combines the results of several general SEs; they’re usually slower but may be more comprehensive
http://www.metacrawler.com/

Special science search engines
http://www.scirus.com/
http://www.atgenetics.com/genetics/indexb.php3
http://www.biolinks.com
http://www.cora.justresearch.com
Before you can retrieve information with an SE, you must have a search strategy

Identify the key concepts

Determine alternative terms for these concepts, if needed

Refine your search to dates, geographic locales, experimental methods if appropriate

Practice helps—strategies and styles differ according to personal experience and professional discipline

Search engines use “keywords” to retrieve information

1. Type in keywords

Your choice of keywords is critical for successful web searching

2. Check the results
Use “web math” to improve your search results

The + symbol narrows a search
- biophysics = 33 730 matches
  +biophysics +ion +channel = 1 100 matches

The – symbol excludes terms
+ion +channel +Na –Ca –K = 97 matches

Use “ ” to search for specific phrases
+sodium +ion +channel = 1 600 matches
  “sodium ion channel” = 11 matches

Use “wildcards” (*) to search for plurals or variations of words
+sodium +ion +channel* = 1 620 matches
  picks up “channels,” “channeling,” and “channelling”

Boolean operators can also be used, and they have one advantage over web arithmetic

Boolean operators define relationships among keywords

AND and NOT work like + and –
  tuberculosis AND hepatitis AND C

OR retrieves documents that contain at least one of the specified search terms

Use OR when you want to pull together articles on similar subjects
  parasites OR helminths

Always type Boolean operators in CAPS
Good results from simple searches

Some sites have site-only searches
Faster than searching the whole web
Good strategy if you think that site should contain the info that you need

http://www.nih.gov
Here are four ways to improve your search results

Use multiple search strategies—different keywords, in different combinations, in different search engines

Use web math and Boolean operators

Be very specific

- imaging = 1,561,200 matches
- medical imaging = 85,400 matches
- ultrasound imaging = 9,800 matches
- ultrasound Doppler sonography = 15 matches

Check your spelling

- autofluorescence = 390 matches
- autofluorecence = 1 match

“404: Document Not Found”

The web changes daily; what was there yesterday is gone tomorrow

Go back up the directory tree of the URL until you find a good link

Go to the main homepage URL and do a search or try to navigate to the page

Send an e-note to the “webmaster” of the referring page, asking for the correct URL

Use one of the search engines and do a keyword search
Another way to find information is through “portals”

Most search engines rely on robots to create their indexes; portals use human beings

Subject experts select, evaluate, describe, and organize the information that goes into the portal—ensures some quality control

Portals provide coherence and fast access to quality information on very specific topics

Portals organize information and allow users to browse by subject

Information is arranged topically so it is easy to see at a glance the range of information that is available

Within topics, information is arranged hierarchically so that you can tell what’s important and see relationships

Because they are labor-intensive and thus expensive to create, portals are not widespread and are limited to specialized topics
Here are key bio portals to explore; some require “registration” but all may be used free of charge

Nature’s Genome Gateway  
http://www.nature.com/genomics/

Bio Netbook (Institut Pasteur)  
http://www.pasteur.fr/recherche/BNB/bnb-en.html

BioVisa (protocols, free books, free journals, fora)  
http://biovisa.net/

EurasiaHealth Knowledge Network  
http://www.eurasiahealth.org/

BioExplorers Net  
http://www.biolinks.net.ru/
Web databases are a cross between SEs and portals; they provide access to immense collections of factual data

U.S. National Institutes of Health

Technical Database Services, Inc. (chemistry)
http://www.tds-tds.com/ind_hum2.htm

Metabolic Database
http://cgsc.biology.yale.edu/metab.html

Like SEs, web databases use keywords to retrieve information

New molecular sequence retrieval system developed at the National Center for Biotechnology Information (NCBI)

Entrez provides access to several linked databases, including PubMed, nucleotide and protein sequence databases, three-dimensional macromolecular structures, complete genome assemblies, population study data sets, organisms in GenBank, and Online Mendelian Inheritance in Man

Let’s review what we’ve learned…

✓ What the Internet and the World Wide Web are
✓ How to navigate the “World Wide Web” (“web”)
✓ How to use search engines, portals, and online databases
✓ How to formulate searches to produce optimal results