stephane.plaisance@vib.be

2011-09-16

NCBI eBot – an online perl script generator to capture complex NCBI search queries

NCBI-Ebot allows capturing steps normally followed by clicking on the different buttons and links present in NCBI search pages and generate a perl script which you can keep and adapt or at least reuse with identical results.

Main advantages over performing manual searches over time:

- You store your method in a self-contained script and are able to re-run the script on regular basis to obtain updated results (**reproducibility**)
- When adapting the code, you caninclude it in your own pipeline to build a software solution, based on the complex query but extending further (**flexibility**).
- The Perl code provided by NCBI is clear enough and is expected to be stable in time unless NCBI drastically change their platform (**simplicity**).

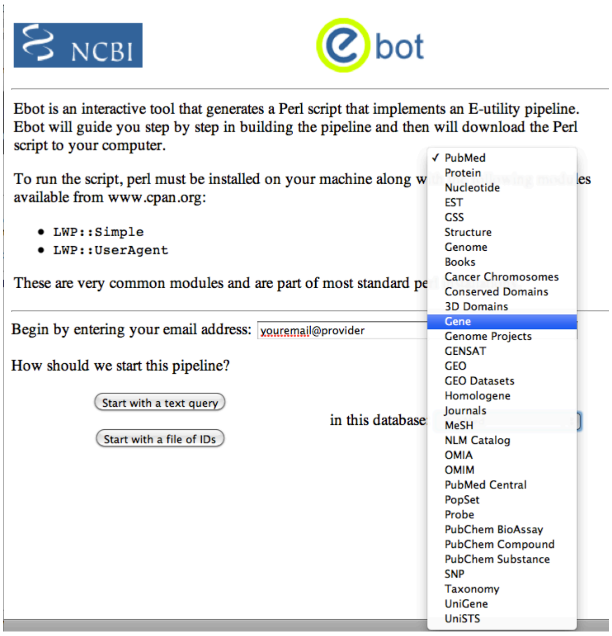
Example usage: Search for a gene symbol (SOD) in the Gene database, then link-out to all SOD-related pubmed articles and perform a new search to limit this list to relevant papers. The final result is a list of articles which you may want to get as a reference list or as counts depending on your needs.

The two steps approach is more efficient than including the gene symbol 'SOD' directly in a standard PubMed query as it will avoid ambigous acceptation obtained from homonymous terms (some gene names are very common english words and can be taken out of the gene-name context).

We present below a series of screen shots describing the few intuitive steps involved in creating the perl code (ony the top part of the code is added in appendix as the remaining ~1000+ code lines are standard subroutines shared by all eBot scripts).

I) eBot demonstration run

Step-1: Initial database selection



Ebot is an interactive tool that generates a Perl script that implements an E-utility pipeline. Ebot will guide you step by step in building the pipeline and then will download the Perl script to your computer.

To run the script, perl must be installed on your machine along with the following modules available from www.cpan.org:

- `LWP::Simple`
- `LWP::UserAgent`

These are very common modules and are part of most standard perl installations.

Begin by entering your email address: youremail@provider

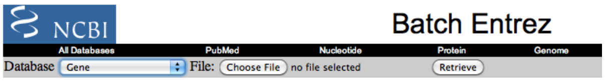
How should we start this pipeline?

in this database: Gene

we take 'Gene' here in order to identify records associated with a given gene. If you want to add more query terms, please keep them for a later stage as all queries will be associated by 'AND'.

Note that you could input here a single gene query **OR** a list of UIDs.

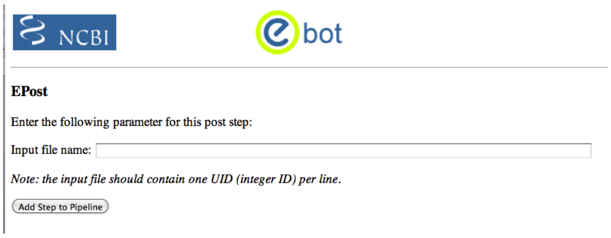
UIDs can be obtained from a list of entrezIDs or Genbanck accessions using NCBI online tools accessible at: http://www.ncbi.nlm.nih.gov/sites/batchentrez)



Batch Entrez

Database: Gene

File: no file selected



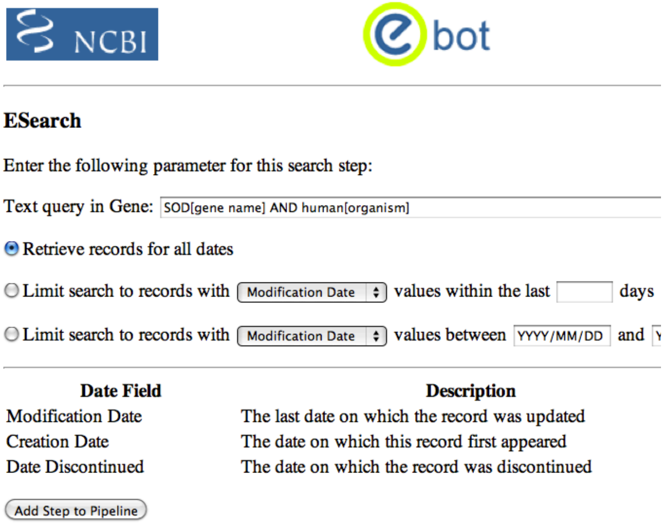
EPost

Enter the following parameter for this post step:

Input file name:

Note: the input file should contain one UID (integer ID) per line.

Step-2: Initial database selection (single gene query)



ESearch

Enter the following parameter for this search step:

Text query in Gene: SOD[gene name] AND human[organism]

● Retrieve records for all dates

○ Limit search to records with Modification Date values within the last ____ days

○ Limit search to records with Modification Date values between YYYY/MM/DD and Y

Date Field	Description
Modification Date	The last date on which the record was updated
Creation Date	The date on which this record first appeared
Date Discontinued	The date on which the record was discontinued

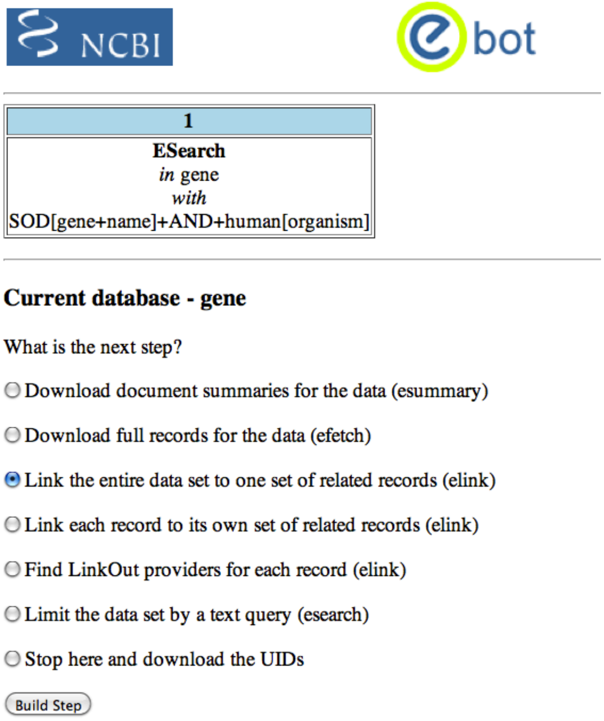
You have obviously many more choices here

Add Step to Pipeline

Step-3: Choosing what to do with the results of Query#1

We choose here to link out to another data type within NCBI

Build Step



1
ESearch *in* gene *with* SOD[gene+name]+AND+human[organism]

Current database - gene

What is the next step?

○ Download document summaries for the data (esummary)

○ Download full records for the data (efetch)

● Link the entire data set to one set of related records (elink)

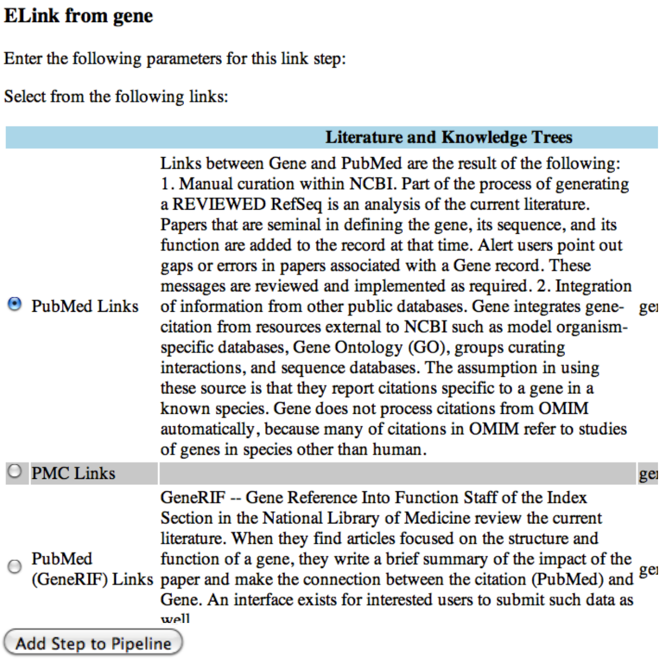
○ Link each record to its own set of related records (elink)

○ Find LinkOut providers for each record (elink)

○ Limit the data set by a text query (esearch)

○ Stop here and download the UIDs

Step-4: Selecting the second database



ELink from gene

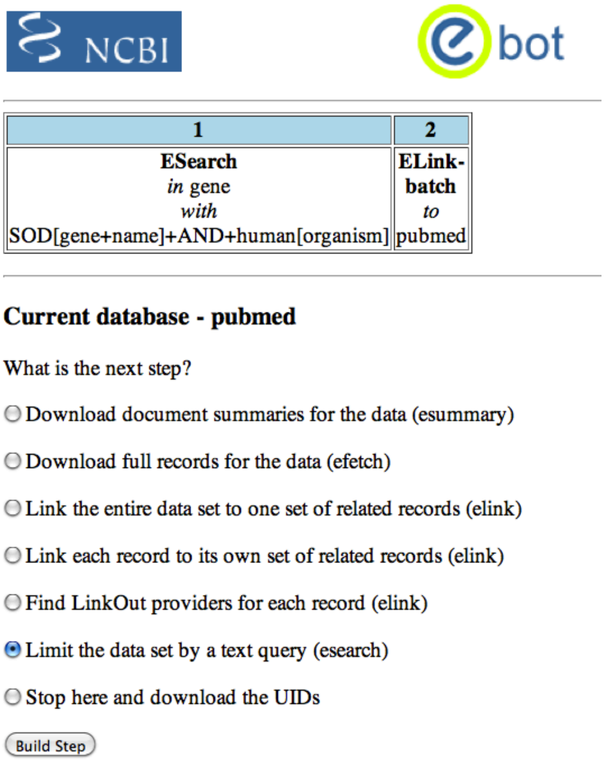
Enter the following parameters for this link step:

Select from the following links:

Literature and Knowledge Trees			
●	PubMed Links	Links between Gene and PubMed are the result of the following: 1. Manual curation within NCBI. Part of the process of generating a REVIEWED RefSeq is an analysis of the current literature. Papers that are seminal in defining the gene, its sequence, and its function are added to the record at that time. Alert users point out gaps or errors in papers associated with a Gene record. These messages are reviewed and implemented as required. 2. Integration of information from other public databases. Gene integrates gene-citation from resources external to NCBI such as model organism-specific databases, Gene Ontology (GO), groups curating interactions, and sequence databases. The assumption in using these source is that they report citations specific to a gene in a known species. Gene does not process citations from OMIM automatically, because many of citations in OMIM refer to studies of genes in species other than human.	ge
○	PMC Links		ge
○	PubMed (GeneRIF) Links	GeneRIF -- Gene Reference Into Function Staff of the Index Section in the National Library of Medicine review the current literature. When they find articles focused on the structure and function of a gene, they write a brief summary of the impact of the paper and make the connection between the citation (PubMed) and Gene. An interface exists for interested users to submit such data as well	ge

Note: the screen shot is here limited but all data types present on teh right of aNCBI search page are represented here.

Step-5: Initiate a new text query



1	2
ESearch *in* gene *with* SOD[gene+name]+AND+human[organism]	**ELink-batch** *to* pubmed

Current database - pubmed

What is the next step?

○ Download document summaries for the data (esummary)

○ Download full records for the data (efetch)

○ Link the entire data set to one set of related records (elink)

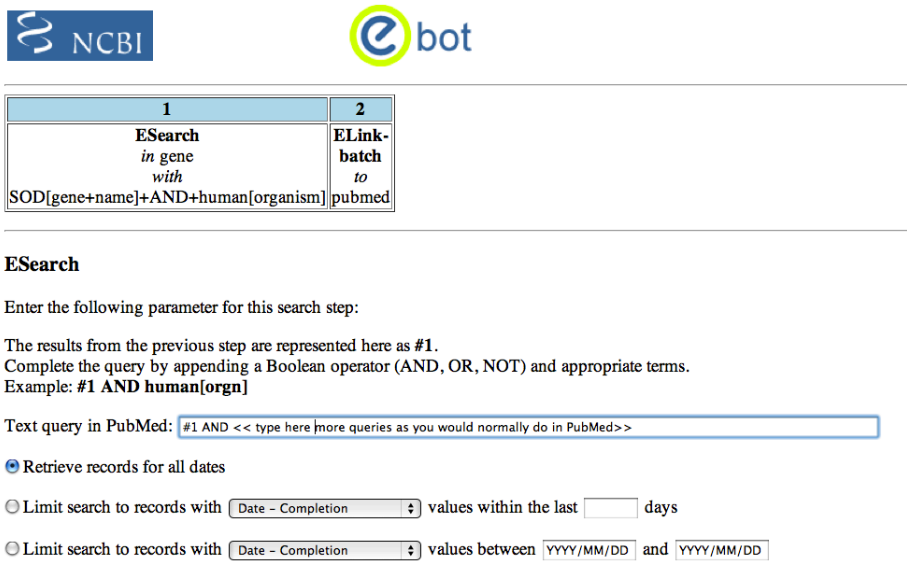
○ Link each record to its own set of related records (elink)

○ Find LinkOut providers for each record (elink)

● Limit the data set by a text query (esearch)

○ Stop here and download the UIDs

Step-6: Initial database selection (& retreive all results)



1	2
ESearch *in* gene *with* SOD[gene+name]+AND+human[organism]	**ELink-batch** *to* pubmed

ESearch

Enter the following parameter for this search step:

The results from the previous step are represented here as **#1**.
Complete the query by appending a Boolean operator (AND, OR, NOT) and appropriate terms.
Example: **#1 AND human[orgn]**

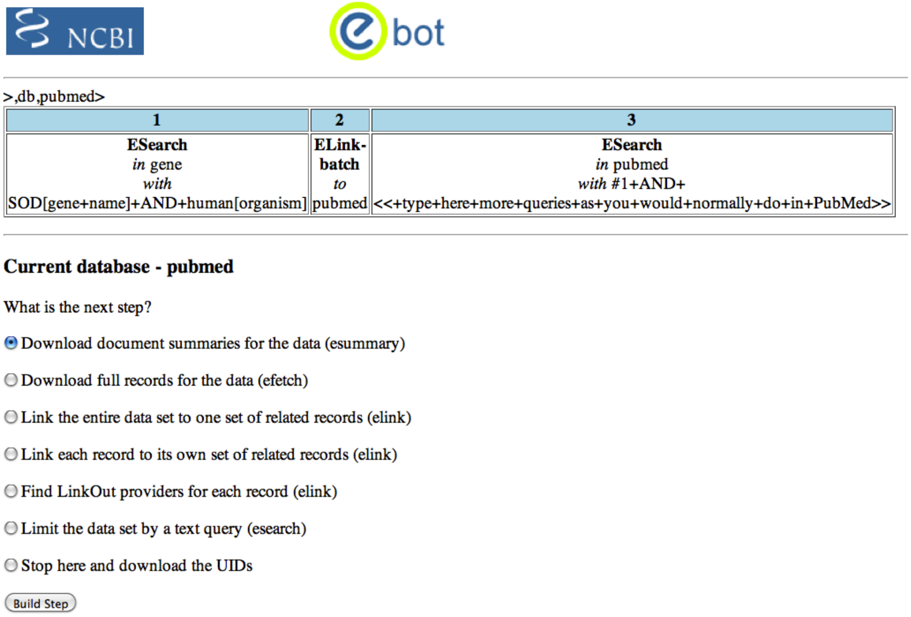
Text query in PubMed: #1 AND << type here more queries as you would normally do in PubMed>>

● Retrieve records for all dates

○ Limit search to records with Date - Completion values within the last ____ days

○ Limit search to records with Date - Completion values between YYYY/MM/DD and YYYY/MM/DD

Step-7: Finish and retreive results (*here as a summary*)



>,db,pubmed>

1	2	3
ESearch *in* gene *with* SOD[gene+name]+AND+human[organism]	**ELink-batch** *to* pubmed	**ESearch** *in* pubmed *with* #1+AND+ <<+type+here+more+queries+as+you+would+normally+do+in+PubMed>>

Current database - pubmed

What is the next step?

● Download document summaries for the data (esummary)

○ Download full records for the data (efetch)

○ Link the entire data set to one set of related records (elink)

○ Link each record to its own set of related records (elink)

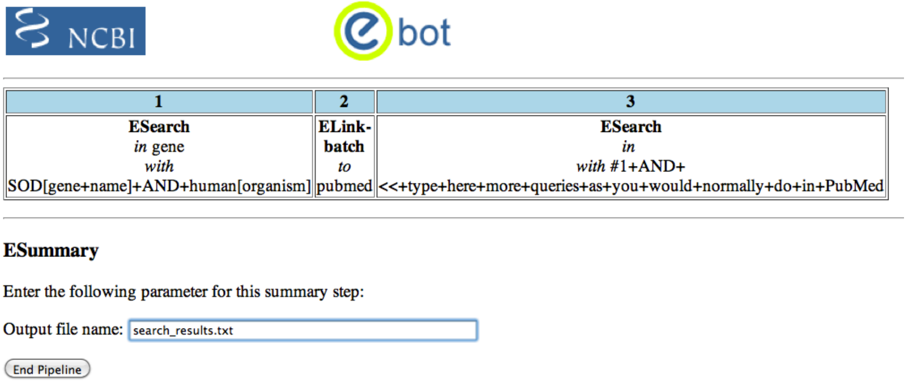
○ Find LinkOut providers for each record (elink)

○ Limit the data set by a text query (esearch)

○ Stop here and download the UIDs

Build Step

Step-8: name the output file



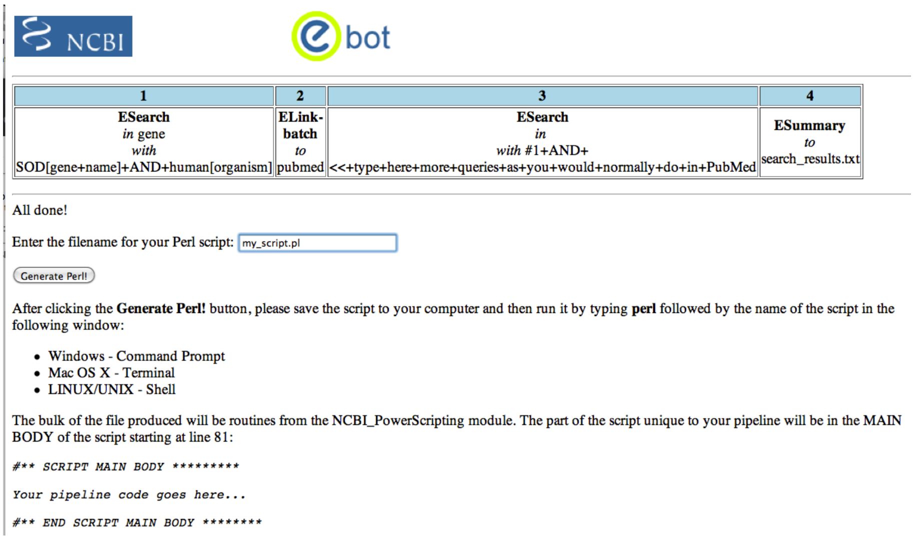
1	2	3
ESearch *in* gene *with* SOD[gene+name]+AND+human[organism]	**ELink-batch** *to* pubmed	**ESearch** *in* *with* #1+AND+ <<+type+here+more+queries+as+you+would+normally+do+in+PubMed

ESummary

Enter the following parameter for this summary step:

Output file name: search_results.txt

Step-9: Name and generate teh perl code on the fly



1	2	3	4
ESearch *in* gene *with* SOD[gene+name]+AND+human[organism]	**ELink-batch** *to* pubmed	**ESearch** *in* *with* #1+AND+ <<+type+here+more+queries+as+you+would+normally+do+in+PubMed	**ESummary** *to* search_results.txt

All done!

Enter the filename for your Perl script: my_script.pl

After clicking the **Generate Perl!** button, please save the script to your computer and then run it by typing **perl** followed by the name of the script in the following window:

- Windows - Command Prompt
- Mac OS X - Terminal
- LINUX/UNIX - Shell

The bulk of the file produced will be routines from the NCBI_PowerScripting module. The part of the script unique to your pipeline will be in the MAIN BODY of the script starting at line 81:

```
#** SCRIPT MAIN BODY *********

Your pipeline code goes here...

#** END SCRIPT MAIN BODY ********
```

Step-10: generate the perl code on the fly

Download the script to your computer and start using it

Appendix: Top of the resulting code as example:

```
#!/usr/bin/perl

#This script was generated by Ebot (http://www.ncbi.nlm.nih.gov/Class/PowerTools/eutils/ebot.cgi)
#Ebot is part of the materials of the NCBI PowerScripting course
#This script contains the routines of the NCBI_PowerScripting.pm module used in the course



# ===========================================================================
#
#                            PUBLIC DOMAIN NOTICE
#               National Center for Biotechnology Information
#
#  This software/database is a "United States Government Work" under the
#  terms of the United States Copyright Act.  It was written as part of
#  the author's official duties as a United States Government employee and
#  thus cannot be copyrighted.  This software/database is freely available
#  to the public for use. The National Library of Medicine and the U.S.
#  Government have not placed any restriction on its use or reproduction.
#
#  Although all reasonable efforts have been taken to ensure the accuracy
#  and reliability of the software and data, the NLM and the U.S.
#  Government do not and cannot warrant the performance or results that
#  may be obtained by using this software or data. The NLM and the U.S.
#  Government disclaim all warranties, express or implied, including
#  warranties of performance, merchantability or fitness for any particular
#  purpose.
#
#  Please cite the author in any work or product based on this material.
#
# ===========================================================================
#
# Author:  Eric W. Sayers  sayers@ncbi.nlm.nih.gov
# http://www.ncbi.nlm.nih.gov/Class/PowerTools/eutils/course.html
#
#
# ---------------------------------------------------------------------------



#Contains the following subroutines:
#read_params
#egquery
#esearch
#esearch_links
```

```
#esummary
#esummary_links_by_id
#efetch
#efetch_batch
#efetch_links_by_id
#elink
#elink_history
#elink_batch
#elink_batch_to
#elink_by_id
#elink_by_id_to
#elink_out
#epost_uids
#epost_file
#epost_set
#print_summary
#print_links
#print_link_summaries
#get_uids
#read_index
#get_linknames
#get_link_report
#extract_links
#get_ftp_file


use LWP::Simple;
use LWP::UserAgent;
use Net::FTP;


my $delay = 0;
my $maxdelay = 3;
my $base = "http://eutils.ncbi.nlm.nih.gov/entrez/eutils/";

#***************************************************************
#** SCRIPT MAIN BODY ******************************************

$params{email} = "stephane.plaisance@vib.be";
$params{db} = "gene";
$params{tool} = "ebot";
$params{term} = "SOD[gene+name]+AND+human[organism]";
%params = esearch(%params);

$params{linkname} = "gene_pubmed";
%params = elink_batch_to('pubmed', %params);

$params{db} = "pubmed";
$params{term}                                                                                          =
"%23$params{query_key}+AND+((clinical[Title/Abstract]+AND+trial[Title/Abstract])+OR+clinical+trials[MeSH
+Terms]+OR+clinical+trial[Publication+Type]+OR+random*[Title/Abstract]+OR+random+allocation[MeSH+Te
rms]+OR+therapeutic+use[MeSH+Subheading])";
%params = esearch(%params);
```

```
$params{outfile} = "results";
esummary(%params);

#** END SCRIPT MAIN BODY ************************************
#***********************************************************
#** BEGIN NCBI_PowerScripting MODULE ROUTINES ***************
#************************************************************
```

code shortenned here, all ermaining lines are subroutines shared by all eBot scripts

//