Introduction to NCBI EDirect Software

Learn how to programmatically search and compile PubMed and related data in a Unix Shell

Fall 2022 UA Libraries, Data Services

Vincent F. Scalfani
vfscalfani@ua.edu

Disclaimer: Workshop is not affiliated with NCBI/NLM/NIH

Outline

1. Overview of NCBI EDirect
2. Brief Unix Shell Review
3. Searching PubMed and related databases with EDirect
4. Retrieving PubMed XML and extracting information
5. Creating scripts for repeated tasks and workflows

Further Reading

Key References and Further Reading:

1. Software Carpentry: The Unix Shell
2. Official NCBI Manual for EDirect - Entrez Direct: E-Utilities on the Unix Command Line
3. NLM EDirect for PubMed Recordings and Materials - EDirect for PubMed
4. NLM EDirect Documentation on xtract
5. NCBI EDirect Cookbook
6. Our PubMed/PubChem EDirect Cookbook, EDirectChemInfo and Unix Workshops

Appropriate EDirect and NCBI Data Usage Notes

Read the NCBI Website and Data Usage Policies and Disclaimers:
https://www.ncbi.nlm.nih.gov/home/about/policies/

See information about abstract copyright in PubMed:
https://www.nlm.nih.gov/databases/download.html

And PubMed Central Copyright Notice:
https://www.ncbi.nlm.nih.gov/pmc/about/copyright/

If you have questions about copyright and fair-use for your particular use-case, contact Elaine Walker, UA Scholarly Communications Librarian: elaine.walker@ua.edu

What is EDirect [1]?

- Free command-line program from National Center for Biotechnology Information (NCBI) that allows (E-utilities) programmatic access to NCBI databases such as PubMed, PubChem, Gene, Taxonomy, etc. directly within a Unix terminal window.

- Can be installed on Unix, Unix-like (e.g., GNU/Linux) distributions, Mac OS, and Windows with Cygwin Unix-emulation.

Example EDirect Use

```
$ esearch -email name@xx.edu -db pubmed -query "\"ionic liquids\"[MESH] AND imidazolium" | \
> efetch -format xml | \
> xtract -pattern PubmedArticle -element MedlineCitation/PMID -first Author/LastName \
> Author/Initials ISOAbbreviation PubDate/Year Volume Issue MedlinePgn \
> -block ArticleId -if ArticleId@IdType -equals doi -doi ArticleId
33396149                    Hu       LX      Ecotoxicol Environ Saf        2021    208     111629
https://doi.org/10.1016%2Fj.ecoenv.2020.111629
33346267            Kaur      M       Phys Chem Chem Phys    2021    23      1       320-328 https://doi.org/10.1039%2Fd0cp04513f
33253998                 Tashakkori  P       J Chromatogr A 2021    1635    461741      https://doi.org/10.1016%2Fj.chroma.2020.461741
...
...
```

[1] https://www.ncbi.nlm.nih.gov/books/NBK179288/

Why Would I Use EDirect?

1. For accessing NCBI data programmatically, EDirect has a lower learning curve than trying to write your own scripts in a programming language. EDirect constructs the programmatic web URLs for you and includes programs to help you format and process the data into tabular formats. *Can also combine/process data with other Unix utilities.*

2. You want to compile bibliographic data or molecular/biological datasets.

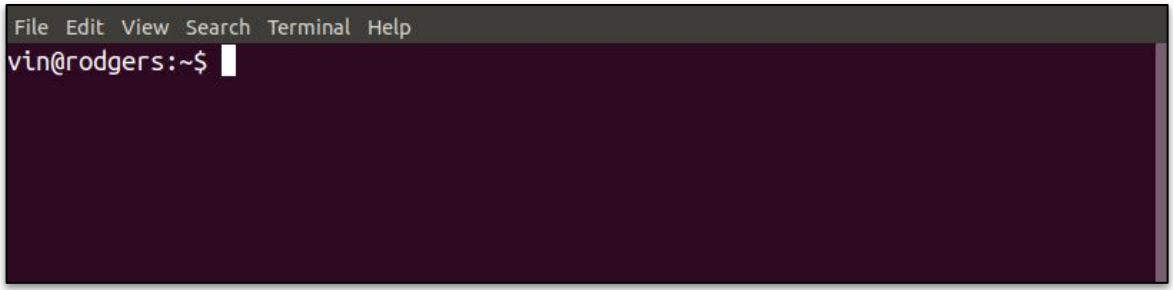
3. You have many search queries (e.g., PubMed) to perform or need to repeat the search frequently. Easy to precisely record your database searches and analysis, which may be useful for systematic reviews.

4. You can quickly answer interesting and specific questions like “Who are the most common authors studying peanut allergies?” or “What is the most common journal indexed in PubMed for research on synthetic chemistry “total synthesis”?

5. Searching NCBI databases in a Unix terminal is a lot of fun.

Unix Terminal

A Unix terminal is a text input/output environment [1]:



From the terminal input, a shell then interprets the commands (i.e., a command line interpreter).

> *Most Unix-like operating systems such as GNU/Linux distributions (e.g., Ubuntu) are using the GNU Bash Shell.*

[1] Unix Stack Exchange Thread: What is the exact difference between a 'terminal', a 'shell', a 'tty' and a 'console'?

Unix Programs and Utilities

To run a Unix program, you generally type the name of the program, followed by (optional) -arguments.

Type **`-help`** or **`--help`** after the program name or **`man`** (for manual) / **info** before the program name to see specific usage. Example with GNU utilities `cut`:

```
$ cut --help
Usage: cut OPTION... [FILE]...
Print selected parts of lines from each FILE to standard output.

With no FILE, or when FILE is -, read standard input.
Mandatory arguments to long options are mandatory for short options too.
  -b, --bytes=LIST        select only these bytes
  -c, --characters=LIST    select only these characters
  -d, --delimiter=DELIM    use DELIM instead of TAB for field delimiter
  -f, --fields=LIST       select only these fields;  also print any line
                          that contains no delimiter character, unless
                          the -s option is specified
...
$ man cut
```

(outputs manual page for cut, more detailed description, not shown)

Unix Shell Pipelines, Redirect, and Loops [1]

With the Unix shell, we can use pipelines to create sequences of commands. Each command output is piped into the next command:

```
$ command1 | command2 | command3
```

We can redirect our output from a command or sequence of commands to a file:

```
$ command1 > myfile1.txt
$ command1 | command2 | command3 > myfile3.txt
```

Unix shell is also a programming language, and, for example, we can create loops to repeat tasks:

```
$ for item in list_of_items
> do
>     something_using $item
> done
```

[1] See the Software Carpentry Unix Shell and Bash Reference Manual

Install EDirect [1]

After logging into our SUSE Linux server, you need to install EDirect into your profile:

```
sh -c "$(curl -fsSL ftp://ftp.ncbi.nlm.nih.gov/entrez/entrezdirect/install-edirect.sh)"
```

Then verify installation by checking if an EDirect folder was created:

```
ls
```

And then check the version:

```
esearch -help
```

[1] https://www.ncbi.nlm.nih.gov/books/NBK179288/

EDirect Unix Programs [1]

EDirect contains several individual programs. We will review the following today:

1. **einfo -** prints fields and links indexed in each database
2. **esearch** - performs an NCBI Entrez database search based on a specified database and query
3. **efetch** - downloads the esearch query results in a specified format such as XML
4. **xtract** - extracts selected data values from XML
5. **elink** - finds associated records within a specified database
6. **efilter** - limits results (e.g., by date, information type, etc.)

Typical use-case is to connect these programs with unix pipelines:

```
$ esearch | efetch | xtract
$ esearch | elink | efilter | efetch | xtract
$ esearch | elink | efilter | efetch | xtract | sort | uniq -c
$ esearch | elink | efilter | efetch | xtract > myfile.txt
```

[1] See the official manual for other EDirect programs and examples: https://www.ncbi.nlm.nih.gov/books/NBK179288/

EDirect program syntax and Usage Notes

EDirect programs all have similar syntax:

```
eprogram -argument input

eprogram -email name@xx.edu -argument input
```

All of the EDirect programs accept your email as an option too, this is a really good idea to add so that if you are accidentally causing server issues or violating their usage policies, NCBI can contact you.

See earlier slide 4 entitled, "Appropriate EDirect and NCBI Data Usage Notes."

einfo

einfo - prints fields and links indexed in each database

```
einfo -help

einfo -dbs

einfo -db pubmed -fields

einfo -db pubmed -links
```

EDirect manual: https://www.ncbi.nlm.nih.gov/books/NBK179288/

esearch

esearch - performs an NCBI Entrez database search based on a specified database and query

```
esearch -help

esearch -db pubmed -query "17630804"[UID]

esearch -db pubmed -query "imidazolium AND bacteria"
```

As queries become more complex, use the `-debug` flag to check the query translation:

```
esearch -db pubmed -query "hydrogel-based drug delivery" -debug

nquire -url https://eutils.ncbi.nlm.nih.gov/entrez/eutils/ esearch.fcgi -retmax 0 \
-usehistory y -db pubmed -term "hydrogel-based drug delivery"
```

EDirect manual: https://www.ncbi.nlm.nih.gov/books/NBK179288/

esearch

esearch - performs an NCBI Entrez database search based on a specified database and query

Escape , \, internal quotes and use parentheses for complex searches:

```
esearch -db pubmed -query "\"Artificial Intelligence\"[MESH] AND \"drug discovery\"[ALL]"

esearch -db pubmed -query "(university of alabama[AFFL]) \
NOT (birmingham[AFFL] OR huntsville[AFFL])"
```

Again, try the -debug flag for testing, and it is also helpful to build queries online with the PubMed Advanced Search Builder: https://pubmed.ncbi.nlm.nih.gov/advanced/, though **e-utilities based search results may be different than the new web based PubMed**.

EDirect manual: https://www.ncbi.nlm.nih.gov/books/NBK179288/

efetch

efetch - downloads the esearch query results in a specified format such as XML

```
efetch -help

esearch -db pubmed -query "17630804"[PMID] | \
efetch -format abstract


esearch -db pubmed -query "17630804"[PMID] | \
efetch -format xml


esearch -db pubmed -query "\"Artificial Intelligence\"[MESH] AND \"drug discovery\"[ALL]" | \
efetch -format xml
```

EDirect manual: https://www.ncbi.nlm.nih.gov/books/NBK179288/

xtract

xtract - extracts selected data values from XML

```
xtract -help
```

Very powerful tool, we will look at some basics today. Use `-outline` option to help identify fields.

```
esearch -db pubmed -query "17630804"[PMID] | \
efetch -format xml

esearch -db pubmed -query "17630804"[PMID] | \
efetch -format xml | \
xtract -outline
```

Let's take a closer look at the PubMed XML

EDirect manual: https://www.ncbi.nlm.nih.gov/books/NBK179288/

PubMed XML example

```xml
<?xml version="1.0" encoding="UTF-8" ?>
<!DOCTYPE PubmedArticleSet PUBLIC "-//NLM//DTD PubMedArticle, 1st January 2019/
<PubmedArticleSet>
  <PubmedArticle>
    <MedlineCitation Status="MEDLINE" Owner="NLM">
      <PMID Version="1">17630804</PMID>
      <DateCompleted>
        <Year>2007</Year>
        <Month>10</Month>
        <Day>25</Day>
      </DateCompleted>
      <DateRevised>
        <Year>2007</Year>
        <Month>08</Month>
        <Day>10</Day>
      </DateRevised>
      <Article PubModel="Print-Electronic">
        <Journal>
          <ISSN IssnType="Print">0022-3263</ISSN>
          <JournalIssue CitedMedium="Print">
            <Volume>72</Volume>
            <Issue>17</Issue>
            <PubDate>
              <Year>2007</Year>
              <Month>Aug</Month>
              <Day>17</Day>
            </PubDate>
          </JournalIssue>
          <Title>The Journal of organic chemistry</Title>
          <ISOAbbreviation>J Org Chem</ISOAbbreviation>
        </Journal>
        <ArticleTitle>Total synthesis and absolute configuration determination
        <Pagination>
          <MedlinePgn>6621-3</MedlinePgn>
        </Pagination>
```

```
PubmedArticle
  MedlineCitation
    PMID
    DateCompleted
      Year
      Month
      Day
    DateRevised
      Year
      Month
      Day
    Article
      Journal
        ISSN
        JournalIssue
          Volume
          Issue
          PubDate
            Year
            Month
            Day
        Title
        ISOAbbreviation
      ArticleTitle
      Pagination
        MedlinePgn
```

Xtract [1]

Basic usage today:

$ xtract **-pattern** A **-element** B C...

Key concepts:

1. pattern defines new rows (e.g., PubMedArticle, Author)

2. element defines new columns (e.g., ArticleTitle, Volume, Issue...)

3. Attributes of XML elements (e.g., <PMID **Version**="1">17630804</PMID>) can be selected with @:
   a. PMID@Version

4. In cases where elements have the same name (e.g., Year), use a / to define your selection as Parent/Child hierarchy
   a. PubDate/Year versus DateRevised/Year

[1] https://dataguide.nlm.nih.gov/edirect/xtract.html

xtract Examples

extract author names as 1 author per line [1]

```
esearch -db pubmed -query "17630804"[PMID] | \
efetch -format xml | \
xtract -pattern Author -element Author/LastName Author/Initials
```

extract author names as 1 article per line [1]

```
esearch -db pubmed -query "17630804"[PMID] | \
efetch -format xml | \
xtract -pattern PubmedArticle -element Author/LastName Author/Initials
```

[1] https://dataguide.nlm.nih.gov/edirect/xtract.html

xtract Examples

Extract PMID and other bibliographic information [1]

```
esearch -db pubmed -query "17630804"[PMID] | \
efetch -format xml | \
xtract -pattern PubmedArticle -element MedlineCitation/PMID \
ISOAbbreviation PubDate/Year Volume Issue MedlinePgn
```

Add in author names [1]

```
esearch -db pubmed -query "17630804"[PMID] | \
efetch -format xml | \
xtract -pattern PubmedArticle -element MedlineCitation/PMID Author/LastName \
Author/Initials ISOAbbreviation PubDate/Year Volume Issue MedlinePgn
```

[1] https://dataguide.nlm.nih.gov/edirect/xtract.html

xtract Examples

Reformat author names using the xtract -block argument [1]

```
esearch -db pubmed -query "17630804"[PMID] | \
efetch -format xml | \
xtract -pattern PubmedArticle -element MedlineCitation/PMID ISOAbbreviation \
PubDate/Year Volume Issue MedlinePgn \
-block Author -element LastName Initials
```

extract only first author names using the xtract -first argument [2]

```
esearch -db pubmed -query "17630804"[PMID] | \
efetch -format xml | \
xtract -pattern PubmedArticle -element MedlineCitation/PMID -first Author/LastName \
Author/Initials ISOAbbreviation PubDate/Year Volume Issue MedlinePgn
```

[1] https://dataguide.nlm.nih.gov/edirect/xtract.html; [2] https://github.com/vfscalfani/EDirectChemInfo

xtract Examples

Good idea to use the xtract default field (-def) value to handle missing fields [1]

```
esearch -db pubmed -query "25818947"[PMID] | \
efetch -format xml | \
xtract -pattern PubmedArticle -def "N/A" -element MedlineCitation/PMID -first \
Author/LastName Author/Initials ISOAbbreviation PubDate/Year Volume Issue \
MedlinePgn
```

Add in the DOI with the -block and conditional -if [1]

```
esearch -db pubmed -query "25818947"[PMID] | \
efetch -format xml | \
xtract -pattern PubmedArticle -def "N/A" -element MedlineCitation/PMID -first \
Author/LastName Author/Initials ISOAbbreviation PubDate/Year Volume Issue MedlinePgn \
-block ArticleId -if ArticleId@IdType -equals doi -doi ArticleId
```

[1] https://dataguide.nlm.nih.gov/edirect/xtract.html

xtract Examples

the same xtract commands can work for queries that return multiple articles [1,2]

```
esearch -db pubmed -query "Anthraquinones/chemical synthesis"[MESH] | \
efetch -format xml | \
xtract -pattern PubmedArticle -def "N/A" -element MedlineCitation/PMID -first Author/LastName \
Author/Initials ISOAbbreviation PubDate/Year Volume Issue MedlinePgn \
-block ArticleId -if ArticleId@IdType -equals doi -doi ArticleId
```

[1] https://dataguide.nlm.nih.gov/edirect/xtract.html; [2] https://github.com/vfscalfani/EDirectChemInfo

elink

elink - finds associated records within a specified database

Citation information from NIH Open Citation Collection

```
elink -help

elink -cited      (references to this article)
elink -cites      (article reference list)
elink -related    (neighbors in same database)
```

```
# Get article citations for PMID 31254167

esearch -db pubmed -query "31254167[PMID]" | \
elink -cited | \
efetch -format xml | \
xtract -pattern PubmedArticle -element MedlineCitation/PMID -first Author/LastName \
Author/Initials ISOAbbreviation PubDate/Year Volume Issue MedlinePgn
```

EDirect manual: https://www.ncbi.nlm.nih.gov/books/NBK179288/

elink Example

elink - finds associated records within a specified database

Can also specify another database:

```
elink -target new_database -name related-link
```

Recall: `einfo -db pubmed -links`

```
# Get related PubChem compounds from a PubMed search [1]

esearch -db pubmed -query "Anthraquinones/chemical synthesis"[MESH] | \
elink -target pccompound -name pubmed_pccompound | \
efetch -format docsum | \
xtract -pattern DocumentSummary -element IsomericSmiles CID InChIKey
```

[1] https://github.com/vfscalfani/EDirectChemInfo

efilter

efilter - limits results (e.g., by date, information type, etc.)

```
efilter -help
```

Basic example use:

```
efilter -query
efilter -pub review
efilter -mindate 2017
```

Sometimes these can be incorporated directly into esearch.

EDirect manual: https://www.ncbi.nlm.nih.gov/books/NBK179288/

efilter examples

limit PubMed results to review articles only

```
esearch -db pubmed -query "Anthraquinones/chemical synthesis"[MESH] | \
efilter -pub review | \
efetch -format xml | \
xtract -pattern PubmedArticle -def "N/A" -element MedlineCitation/PMID -first \
Author/LastName Author/Initials ISOAbbreviation PubDate/Year Volume Issue \
MedlinePgn -block ArticleId -if ArticleId@IdType -equals doi -doi ArticleId
```

limit PubMed results from a linked PubChem search to a specific Journal [1]

```
esearch -db pccompound -query 174076[uid] | \
elink -target pubmed -name pccompound_pubmed | \
efilter -query "Phys Chem Chem Phys"[JOUR] | \
efetch -format xml | \
xtract -pattern PubmedArticle -def "N/A" -element MedlineCitation/PMID -first \
Author/LastName Author/Initials ISOAbbreviation PubDate/Year Volume Issue \
MedlinePgn -block ArticleId -if ArticleId@IdType -equals doi -doi ArticleId
```

[1] https://github.com/vfscalfani/EDirectChemInfo

Creating a For Loop for multiple Queries [1,2]

Let's say I have a list of PMIDs and want bibliographic information for each one:

```
for refs in \
      "20426451" \
      "21982300" \
      "21948594" \
      "12653513" \
      "11259830" \
      "10592235" \
      "16796559" \
      "27899562" \
      "26400175" \
      "8709122"
do
      esearch -db pubmed -query "$refs[PMID]" |
      efetch -format xml |
      xtract -pattern PubmedArticle -def "N/A" -element MedlineCitation/PMID -first Author/LastName \
      Author/Initials ISOAbbreviation PubDate/Year Volume Issue MedlinePgn \
      -block ArticleId -if ArticleId@IdType -equals doi -doi ArticleId
      sleep 1
done
```

[1] EDirect manual: https://www.ncbi.nlm.nih.gov/books/NBK179288/; [2] https://github.com/vfscalfani/EDirectChemInfo

Creating a For Loop for multiple Queries [1,2]

```
# Or maybe we want the number of cited references for each PMID:

for refs in \
      "20426451" \
      "21982300" \
      "21948594" \
      "12653513" \
      "11259830" \
      "10592235" \
      "16796559" \
      "27899562" \
      "26400175" \
      "8709122"
do
      esearch -db pubmed -query "$refs[PMID]" |
      elink -cited |
      xtract -pattern ENTREZ_DIRECT -lbl "$refs" -element Count
      sleep 1
done
```

[1] EDirect manual: https://www.ncbi.nlm.nih.gov/books/NBK179288/; [2] https://github.com/vfscalfani/EDirectChemInfo

Answering Specific Questions

most common UA chemistry authors indexed in PubMed [1]:

```
esearch -db pubmed -query "(university of alabama[AFFL] AND tuscaloosa[AFFL])" | \
efetch -format xml | \
xtract -pattern Author -if Affiliation -contains chemistry -and Affiliation -contains \
tuscaloosa -element LastName Initials | \
sort-uniq-count-rank
```

[1] N.B. affiliation query and xtract pattern is not perfect, see more here:
https://github.com/vfscalfani/EDirectChemInfo/blob/master/05_EDirect_PubMed_Recipes.md

[2] sort-uniq-count-rank: https://dataguide.nlm.nih.gov/edirect/sort-uniq-count-rank.html

Answering Specific Questions

most Frequent Journals for a PubMed Query [1]

```
esearch -db pubmed -query "\"Artificial Intelligence\"[MESH] AND \"drug discovery\"[ALL]" | \
efetch -format xml | \
xtract -pattern PubmedArticle -element ISOAbbreviation | \
sort-uniq-count-rank
```

[1] https://github.com/vfscalfani/EDirectChemInfo

Answering Specific Questions

how many records are being added to PubMed by create date each month? [1]

```
for date in \
     "2021/01" \
     "2021/02" \
     "2021/03" \
     "2021/04" \
     "2021/05" \
     "2021/06"
do
     esearch -db pubmed -query "$date[CRDT]" |
     xtract -pattern ENTREZ_DIRECT -lbl "$date" -element Count
     sleep 1
done
```

[1] https://github.com/vfscalfani/EDirectChemInfo

Answering Specific Questions

number of records for a PubMed query that are available in PubMed Central [1]:

```
esearch -db pubmed -query "J Chem Inf Model[JOUR]" | \
elink -target pmc -name pubmed_pmc | \
efetch -format docsum | \
xtract -pattern DocumentSummary -element PubDate | \
cut -d " " -f 1 | \
sort-uniq-count-rank | \
sort -k2,2
```

[1] https://github.com/vfscalfani/EDirectChemInfo

Answering Specific Questions

```
# most frequent article title words from a PubMed query

esearch -db pubmed -query "J Cheminform[JOUR]" | \
efetch -format xml | \
xtract -pattern PubmedArticle -element -lower ArticleTitle | \
tr '\n' ' ' | \
word-at-a-time | \
filter-stop-words | \
sort-uniq-count-rank | \
head -n50
```

And many more possibilities, use your imagination and look at linked resources on the next slide.

word-at-a-time: https://dataguide.nlm.nih.gov/edirect/word-at-a-time.html

Thanks!

Key References and Further Reading:

1. Software Carpentry: The Unix Shell
2. Official NCBI Manual for EDirect - Entrez Direct: E-Utilities on the Unix Command Line
3. NLM EDirect for PubMed Recordings and Materials - EDirect for PubMed
4. NLM EDirect Documentation on xtract
5. NCBI EDirect Cookbook
6. Our PubMed/PubChem EDirect Cookbook, EDirectChemInfo and Unix Workshops

Need help?

Let me know!

Vincent F. Scalfani
vfscalfani@ua.edu