SAS® GLOBAL FORUM 2016

IMAGINE. CREATE. INNOVATE.

Accessing PubMed Data using SAS and the Entrez Programming Utilities

Craig Hansen, PhD

South Australian Health and Medical Research Institute, Australia.

Click to start

MENU (click on heading)

#SASGF

Accessing PubMed Data using SAS and the Entrez Programming Utilities

Craig Hansen, PhD
South Australian Health and Medical Research Institute

What is PubMed?

PubMed (http://www.ncbi.nlm.nih.gov/pubmed) is a free search engine within MEDLINE and has become one of the standard databases to search for scientific abstracts.

MEDLINE is a suite of indexed databases developed and maintained by the National Center for Biotechnology Information (NCBI) at the United States National Library of Medicine (NLM).

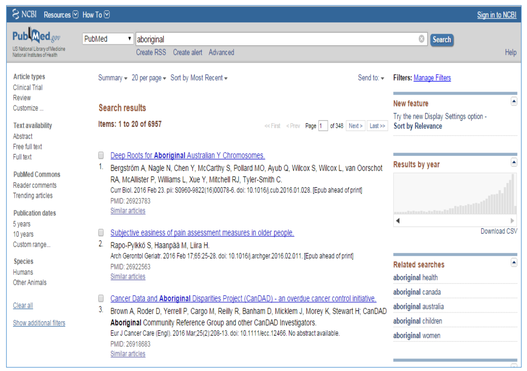

Figure 1. Pubmed search results

Figure 2. Journal article abstract

PROBLEM TO SOLVE: *I want to create a database of all the journal articles based on a search term. I don't want to do this manually!*

SOLUTION TO PROBLEM: *Entrez Programming Utilities allows you to extract all this information into data formats (then convert to SAS datasets)*

Entrez Programming Utilities (APIs)

Entrez is the information retrieval system that gives you direct access to the 40 databases with over 1.3 billion records within the NCBI.

You can access these records by using the eight e-utilities (einfo, esearch, summary, efetch, elink, einfo, epost and egquery) - the NCBI application programming interfaces (APIs).

Base URL: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/

eInfo: Provides information about each database
http://eutils.ncbi.nlm.nih.gov/entrez/eutils/einfo.fcgi?db=pubmed

eSearch: Provides a list of the UIDs (e.g. IDs for records in a particular database) for a search term
http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=asthma

eFetch: Provides formatted output for a list of UIDs (for PUBMED it will be PMIDs). This example we will fetch the information for the first PMID listed in the XML that was generated.
http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&id=26432873&retmode=xml

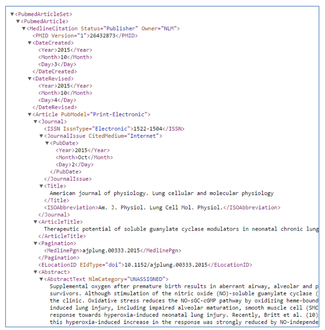

Figure 3. Journal article XML file

- **Use the API to search journal articles**
- **Extract the XML files using SAS PROC HTTP**
- **Map the XML file to dataset using SAS XML Mapper**

Accessing PubMed Data using SAS and the Entrez Programming Utilities

Craig Hansen, PhD
South Australian Health and Medical Research Institute

1. Run eSearch to get max # of records

```
/** SET UP MACRO VARIABLES **/

* - XML file to save;
%LET FILE = C:\SAS\Global Forum 2016\eSearchHistory.xml;
* - Search term URL;
%LET URL = http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?
db=pubmed%nrstr(&term)=aboriginal[TIAB]%nrstr(&RETMAX)=1;

/** RUN eSearch TO GET THE MAXIMUM NUMBER OF RECORDS TO RETURN **/

* - Proc HTTP;
FILENAME test1 "&FILE." encoding="UTF-8";
PROC HTTP
        OUT=test1
        URL="&URL."
        METHOD="get";
RUN;
* - Map the XML file;
FILENAME  Maxsrch 'C:\SAS\Global Forum 2016\eSearchMax.xml';
FILENAME  MyMap 'C:\SAS\Global Forum 2016\PubmedIDs.map';
LIBNAME   Maxsrch xmlv2 xmlmap=MyMap ACCESS=READONLY;
        * - Put the count variable into a macro variable;
PROC SQL NOPRINT;
SELECT DISTINCT
        PUT(Count,BEST12.), QueryTranslation
        INTO :MYCOUNT TRIMMED, :QUERY TRIMMED
FROM MaxSrch.eSearchResult;
QUIT;
```

Need to create an XML map in SAS XML mapper first

Get the max# records from the eSearchResults table and create macro variables for the next step

2. Run eSearch - “usehistory=y” parameter

```
* - Run eSearch with the "&usehistory=y" paremeter,
        this will save all the PMIDs in the ENTREZ database for later use;

%LET FILE = C:\SAS\Global Forum 2016\eSearchHistory.xml;
%LET URL = http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?
db=pubmed%nrstr(&term)=aboriginal[TIAB]%nrstr(&retmax)=&MYCOUNT.%nrstr(&usehistory)=y;

FILENAME test1 "&FILE." encoding="UTF-8";
PROC HTTP
        OUT=test1
        URL="&URL."
        METHOD="get";
RUN;
* - Map the XML file;
FILENAME  eSearch       'C:\SAS\Global Forum 2016\eSearchHistory.xml' ;
FILENAME  MyMap         'C:\SAS\Global Forum 2016\eSearchHistory.map';
LIBNAME   eSearch       xmlv2 xmlmap=MyMap ACCESS=READONLY;
* - Get the "&query_key" and "&WebEnv" paremeters;
PROC SQL NOPRINT;
SELECT PUT(querykey,BEST12.), WebEnv INTO :QK TRIMMED, :WEBENVKEY TRIMMED
FROM eSearch.eSearchResult;
QUIT;
```

Get the “querykey’” and “WebEvn” and create macro variables to use in the next step

*** The step above is optional – however it is best to use this when you have a large output of articles ***

*** Using the “usehistory=y” saves the results (e.g. PMIDs) in the Entrez server ready for extraction by running a query with the QueryKey and WebEnv values given in the eSearchResults table ***

Accessing PubMed Data using SAS and the Entrez Programming Utilities

Craig Hansen, PhD
South Australian Health and Medical Research Institute

3. Run eFetch to get the final results

```
* - Run eFetch with the Query_Key and WebEnv paremeters;
%LET FILE = C:\SAS\Global Forum 2016\eSearchHistoryResults.xml;
%LET URL = http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?
db=pubmed%nrstr(&query_key)=&QK.%nrstr(&WebEnv)=&WEBENVKEY.%nrstr(&usehist
ory)=y%nrstr(&retmode)=xml;
FILENAME test1 "&FILE." encoding="UTF-8";
PROC HTTP
        OUT=test1
        URL="&URL."
        METHOD="get";
RUN;
FILENAME  eFetch 'C:\SAS\Global Forum 2016\eSearchHistoryResults.xml' ;
FILENAME  MyMap         'C:\SAS\Global Forum 2016\eFetch.map';
LIBNAME   eFetch        xmlv2 xmlmap=MyMap ACCESS=READONLY;
```

4. Create final tables from XML

Create 'Authors' dataset

```
* - Authors;
PROC SQL;
CREATE TABLE eFetch_AUTHORS AS
SELECT DISTINCT
        A.PMID,
        B.Author_ORDINAL,
        B.AuthorList_ORDINAL,
        B.LastName,
        B.ForeName,
        B.Initials
FROM eFetch.PMID AS A, eFetch.Author AS B
WHERE A.MedlineCitation_ORDINAL=B.AuthorList_ORDINAL
ORDER BY PMID, AuthorList_ORDINAL;
QUIT;
```

4. (continued) Create final tables from XML

Create 'Articles' dataset

```
* - Articles;
PROC SQL;
CREATE TABLE eFetch_ARTICLE AS
SELECT DISTINCT
        A.MedlineCitation_Status,
        A.MedlineCitation_Owner,
        B.PMID,
        C.YEAR AS CREATED_YEAR,
        C.MONTH AS CREATED_MONTH,
        C.DAY AS CREATED_DAY,
……More code here

FROM            eFetch.MedlineCitation  AS A
LEFT JOIN       eFetch.PMID AS B ON A.PubmedArticle_ORDINAL=B.MedlineCitation_ORDINAL
……More code here
LEFT JOIN       eFetch.PubmedData AS M ON
B.MedlineCitation_ORDINAL=M.PubmedData_ORDINAL
……etc
;
QUIT;
```

Create 'Abstract' dataset

- `* - Abstract;`

……Use similar SQL to create ABSTRACT table

Create 'Affiliations' dataset

- `* - Author Affiliations;`

……Use similar SQL to create AUTHOR AFFILIATIONS table

Accessing PubMed Data using SAS and the Entrez Programming Utilities

Craig Hansen, PhD
South Australian Health and Medical Research Institute

RESULTS: SAS Datasets

The main SAS datasets generated from the XML file and SQL joins are:

- ARTICLE
- AUTHOR
- AUTHOR AFFILIATION
- ABSTRACT

Relational database design with lots of information

Articles

PMID	ArticleTitle	JOURNAL_TITLE	Volume	Issue	MedlinePgn	PUBDATE_YEAR
1282258	Effects of maternal smoking upon neuropsychological development in early childhood: importance of taking account of social and environmental factors.	Paediatric and perinatal epidemiology	6	4	403-15	1992

Authors

PMID	Author_ORDINAL	LastName	ForeName	Initials
1282258	1	Baghurst	P A	PA

Author Affiliations

PMID	Author_ORDINAL	AffiliationInfo_ORDINAL	Affiliation
1282258	1	1	CSIRO Division of Human Nutrition, Adelaide, Australia.

Abstracts

PMID	AbstractText_NlmCategory	AbstractText
1282258		Data from a prospective study of 548 children followed from birth to 4 years of age were analysed to determine whether maternal smoking during and/or after pregnancy affects children's neuropsychological development. The differences in mean developmental test scores between children whose mothers smoked and those of mothers who did not smoke were slight, with subscale scores only 2.4 to 4.1% lower in children whose mothers smoked. These differences were not statistically significant after adjustment for socio-economic status, quality of home environment and mother's intelligence, suggesting that the social and environmental factors are major confounders of the association of exposure to maternal smoking and neuropsychological development in childhood. In order to gain a better understanding of this area, more precise measures of exposure to environmental tobacco smoke and comprehensive consideration of confounders will be required.

Figure 4. Example SAS datasets generated – these can be linked via the PMID field

CONCLUSION

PROS

- Using the APIs with PROC HTTP is a very efficient method to get data from PubMed
- Create PubMed datasets automatically based on different searches
- Extract large amounts of data in one go (e.g. no looping required for limits)
- Can create macros to perform multiple searches and append results

CONS

- Documentation on the Entrez Programming Utilities could be improved with more information on different parameters used in the URL
- There are many tables generated by the XML mapper and it takes a while to workout the linkages
- No bibiolmetric data in PubMed (e.g. times a journal article is cited, and by who etc)
- Truncation of data fields without knowing the length

REFERENCES AND READING

Introduction to E-Utilities. http://www.ncbi.nlm.nih.gov/books/NBK25501/
E-Utilities Introduction (YouTube). https://www.youtube.com/watch?v=BCG-M5k-gvE
SAS PROC HTTP Documentation.
http://support.sas.com/documentation/cdl/en/proc/61895/HTML/default/viewer.htm#a003286672.htm
McNeill, B. (2013). The Ins and Outs of Web-Based Data with SAS., SAS Institute Inc., Cary, NC
https://support.sas.com/resources/papers/proceedings13/024-2013.pdf
Litton, I & Ottesen, R. (2013). %GrabTweet: A SAS® Macro to Read JSON Formatted Tweets.
http://www.lexjansen.com/wuss/2013/103_Paper.pdf
Martell, C. (2008). SAS® XML Mapper to the Rescue.
http://www2.sas.com/proceedings/forum2008/099-2008.pdf

SAS® GLOBAL FORUM 2016

IMAGINE. CREATE. INNOVATE.

THANK YOU FOR YOUR INTEREST

Craig Hansen, PhD
South Australian Health and Medical Research Institute
Craig.Hansen@sahmri.com

LAS VEGAS | APRIL 18-21

#SASGF