Scraping and mining XML data with SAS

Herve J. Momo
Manulife Financial
TASS, SEP-25th, 2015
Toronto

Motivation

- Breast MRI Literature analysis on PUBMED
  - Which journals actually publish in the field?
  - Which countries publish the most?
  - Where is the journal located?

  - What is the publication status in the field
    - Initial publication,
    - Recent publication,
    - Evolution through the years.

What is XML data

- XML = eXtensible Markup Language
- Use as a medium for data exchange
- Just another Text file but with hierarchical structure
- Understanding the structure is the key to successful processing
- PUBMED data available in XML format

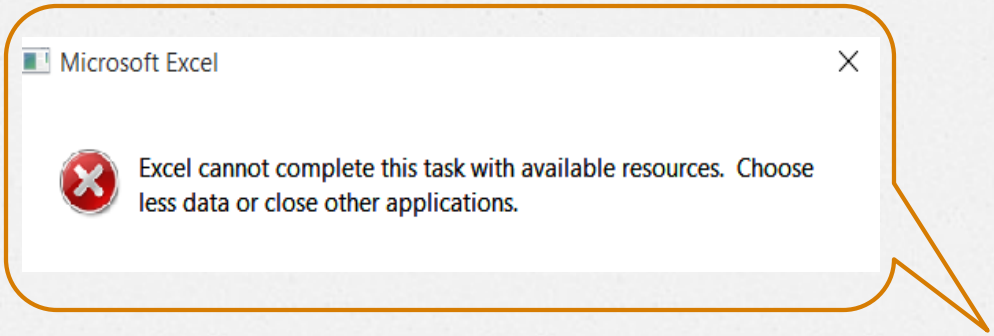
PUBMED is an archive of biomedical and life sciences journal literature

What is XML data

```
<Books>
  <Article>
    <Title>SAS and Machine Learning for Beginner</Title>
    <Year>2015</Year>
    <Month>9</Month>
    <Day>25</Day>
  </Article>
  <Article>
    <Title>Sam Teach Yourself SAS in 21 days</Title>
    <Year>2013</Year>
    <Month>9</Month>
    <Day>25</Day>
  </Article>
</Books>
```

Title	Year	Month	Day
SAS and Machine Learning for Beginner	2015	9	25
Sam Teach Yourself SAS in 21 days	2013	3	6

Thinking Excel?



- File contains about 2 millions records

sas tools for reading XML files

- Using XML engine on a libname statement

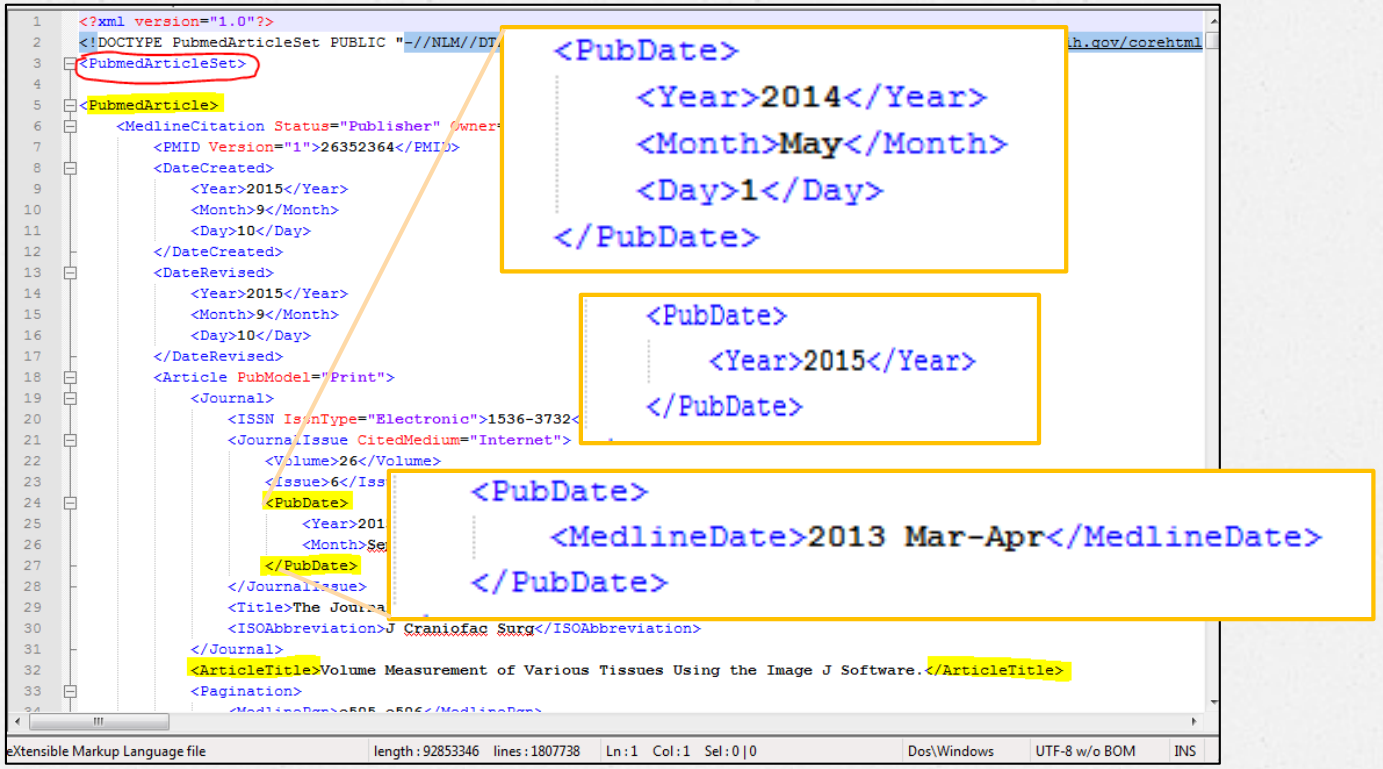
```
libname xmlread xml "Absolute path to xmlfile";
```

- XML Mapper for more complex file
  - GUI interface that create map for reading
- SAS Programing for even more complex file
  - The option available to me

Four Steps for Reading complex XML

- Reading in complicated XML files into SAS can be accomplished using a four step process:
  1. Explore and understand the structure
  2. Decide what you want to extract
  3. Write the program
  4. Submit your program and validate the extract

Explore and understand



Decide what you want

- Publication date
- Name of journal
- Title of Article
- Authors affiliation
- Country of affiliation

Writing your program: SAS Function

- Some important character processing functions for the program
  - Index,
  - Scan,
  - Substr,
  - Tranwrd
  - Strip
  - compress
  - Prxparse,
  - Prxmatch

Writing your program: Algorithm

- Reading a record

```
filename inputfile "C:\Users\momoher\Downloads\pubmed_result.xml" lrecl=2048 ;
data pmidc;
            INFILE &inputfile  truncover;
            INPUT record $2048.;
            retain rtclid 1;
            if index(record, '<ArticleId IdType="pubmed">')=1 then output ;
            if index(record, '</PubmedArticle>')=1 then  rtclid+1;
run;
```

Very important

- Cleaning a record

```
data pmidn(keep= pmid rtclid);
            set pmidc;
            pos1=length('<ArticleId IdType="pubmed">');
            pos2=find(record, '</ArticleId>');
            pmid1=substr(record, pos1+1, pos2-pos1-1);
            pmid=input(pmid1, 12.);
run;
```

Writing your program: Pattern Matching

```
data affiliation;
set affiliationtemp;
length eml_add $60 affiliation $512;
             if _n_=1 then do;
rx1=prxparse("/( +\w*\.\w\w*\@\w\w*)|( +\w\w*\@\w\w*)/");
rx2=prxparse("/Electronic address:/");
end;
retain rx1 rx2;
posn1=prxmatch(rx1, record); posn2=prxmatch(rx2, record);
if posn1 > 0 and posn2 >0 then do;
eml_add=strip(substr(record, posn2+length("Electronic address:")+1));
toremove= cat("Electronic address:","",eml_add);
affiliation=strip(TRANWRD(record, toremove, ''));
end;
else if posn1 > 0 then do;
eml_add=strip(substr(record, posn1));
affiliation=strip(TRANWRD(record, eml_add, ''));
end;
else do;
affiliation=strip(record); email="";
end;
run;
```

Match email pattern

Remove email from records

Validating the extract: Geographical Data

```
/*Join all tables */
proc sql;
create table &outputlib..alldata as
select p.*, a.*, j.*, d.*, t.articletitle
from &outputlib..pmidn as p, author_country as a, &outputlib..journal
as j, &outputlib..pubdaten as d, &outputlib..article_title as t
where p.rtclid=a.rtclid and a.rtclid=j.rtclid and j.rtclid=d.rtclid and
d.rtclid=t.rtclid ;
quit;
```

```
/*Validate countries*/
proc sort data=&outputlib..alldata out=alldata nodupkey;
by country pmid journal;
run;
data real_country unknown_country;
          merge alldata (in=f1) GEOGRAPHICAL_LOOKUP(in=f2);
          by country;
          if f1;
          if f1 and f2 then output known_country;
          else output unknown_country;
run;
```

Require several iterations

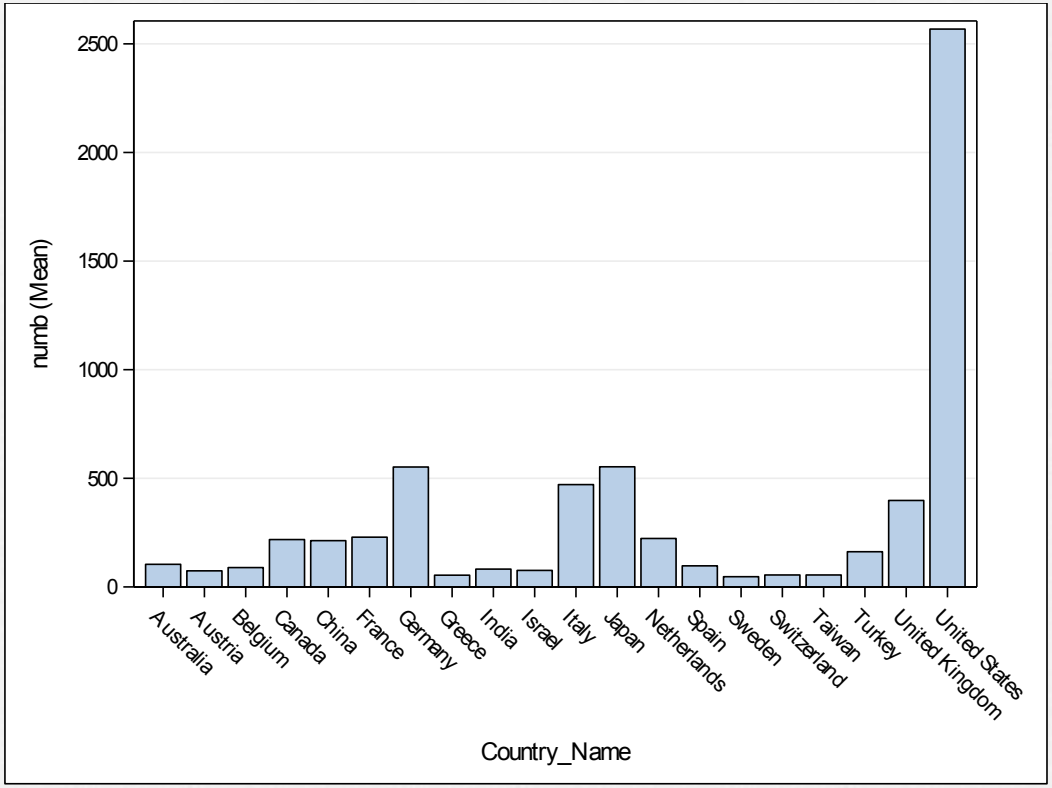
Sample Data

	pmid	country	journal	pub date	affiliation	eml_add	articletitle	ISO_2_Letter_Code
1948	20799873	India	Future Oncol	2010	Department of General S...	profanandkumar...	Color Doppler ult...	IN
1949	20803241	Italy	Breast Cancer Res. Treat.	2010	Dipartimento di Scienze...	f.sardanelli@gru...	Additional finding...	IT
1950	20808988	Turkey	Hell J Nucl Med	2010	Department of Nuclear...	tozulker@ttmail....	The efficacy of (9...	TR
1951	20811744	Australia	World J Surg	2010	Department of Surgery,...	kylie.musgrave...	Surgical decision...	AU
1952	20825223	Italy	ACS Nano	2010	Dipartimento di Scienze...		Single-domain pr...	IT
1953	20828980	Ireland	Can Assoc Radiol J	2012	Department of Radiology...	naomicampbell8...	Imaging patterns...	IE
1954	20829043	Canada	Breast	2011	Department of Medical O...	Daphne_tsoi@h...	Willingness of br...	CA
1955	20830650	Germany	Rofo	2011	Institut fÃ¼r Diagnostisc...	dietzelmatthias2...	Magnetic resona...	DE
1956	20833496	Japan	Magn Reson Imaging	2011	Department of Medical P...	nishiura@butsur...	Evaluation of tim...	JP
1957	20833844	Japan	Radiographics	2010	Department of Diagnosti...	yamataka@aa.c...	Radiologic-patho...	JP
1958	20839000	Canada	Eur Radiol	2011	Division of Breast Imagin...	pavel.crystal@u...	High-risk lesions...	CA
1959	20845796	Italy	Tumori	2010	Unit of Diagnostic Radiol...	giovanna.trecate...	Is there a specifi...	IT
1960	20847879	India	Breast Care (Basel)	2009	Regional Cancer Centre,...		Magnetic Reson...	IN
1961	20850641	Saudi Arabia	J. Pediatr. Surg.	2010	Division of Pediatric Sur...	aaljazaeri@ksu....	Unilateral breast...	SA
1962	20852328	Italy	Interact Cardiovasc Thorac Surg	2010	Division of Cardiac Surg...	ceresa77@hotm...	Right atrial lipom...	IT
1963	20853046	Canada	Ann. Surg. Oncol.	2010	Division of Experimental...		Impact of preope...	CA
1964	20863686	Germany	Eur. J. Cancer	2011	Department of Medical P...		Sorafenib tosylat...	DE
1965	20871661	Japan	J Oncol	2010	Department of Radiology...		Apparent Diffusio...	JP
1966	20878973	Germany	NMR Biomed	2010	Philips Research Europe...		Fast T(2) relaxo...	DE
1967	20884914	Italy	Radiology	2010	Department of Radiologi...	federica.pediconi...	Role of breast M...	IT
1968	20885929	Taiwan	J Oncol	2010	Department of Radiology...		Characterization...	TW
1969	20927653	Italy	Radiol Med	2011	Dipartimento di Radiolog...	giravero@yahoo.it	Inflammatory bre...	IT
1970	20932740	Belgium	Eur. J. Cancer	2010	Multidisciplinary Breast...	Frederic.amant...	Breast cancer in...	BE
1971	20937600	China	Jpn. J. Clin. Oncol.	2011	Department of Breast Su...		Clinical features...	CN
1972	20938085	Spain	Cancer Biomark	2010	IECSCYL, Spain.		Autoantibody pro...	ES
1973	20942728	Germany	Acta Radiol	2010	Institute of Diagnostic an...	Tibor.Vag@med....	Kinetic characteri...	DE
1974	20945289	Germany	Zentralbl Chir	2011	UniversitÃ¤tsmedizin Ma...		[Leiomyosarcom...	DE
1975	20956982	Canada	Nat Rev Clin Oncol	2010	Womens College Resear...	steven.narod@w...	BRCA mutations...	CA
1976	20958072	United States	ACS Nano	2010	JHU ICMIC Program, Th...	congli@fudan.ed...	Nanoplex deliver...	US
1977	20959375	Germany	Br J Radiol	2011	Department of Diagnosti...	till.heusner@uni-...	Diagnostic accur...	DE
1978	20964170	United Kingdom	Med Phys	2010	Cancer Research UK Ep...		Automated regist...	GB
1979	20964216	China	Med Phys	2010	Beijing City Key Laborat...		Feasibility of hig...	CN

Some challenges encountered

- Affiliation challenges
  - Same author with different contacts in the same country
  - Affiliation ending with email address after country
  - Affiliation ending with country
  - Country information missing
  - Several authors in the same country for the same article

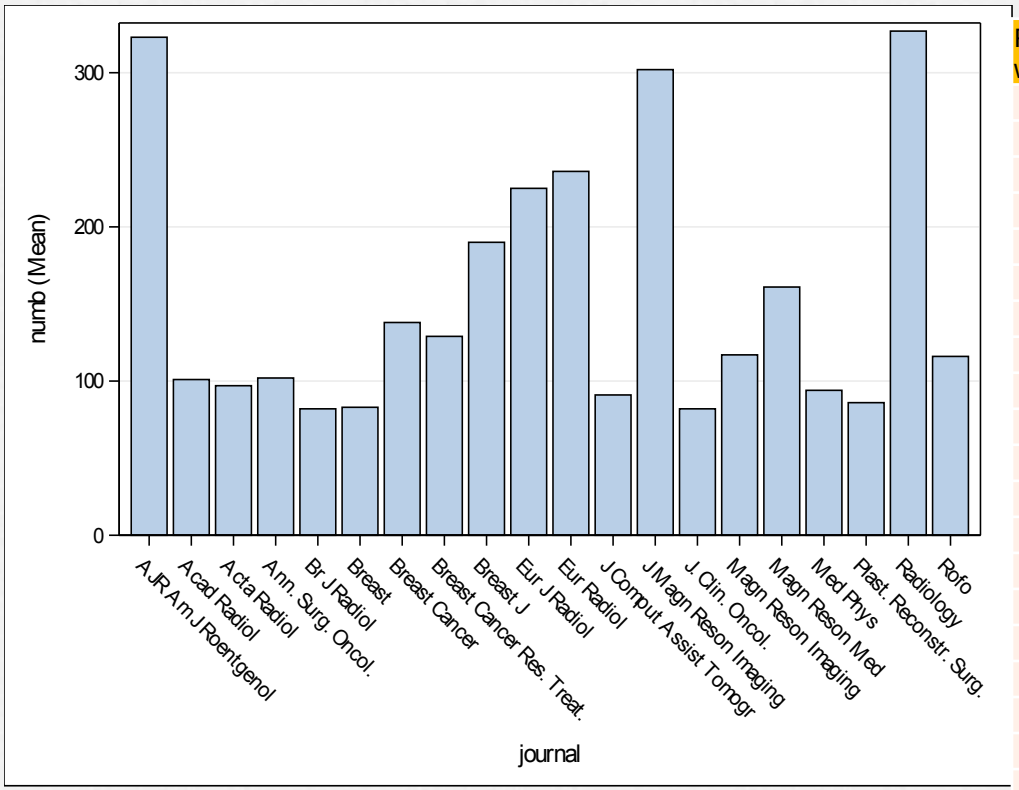
Country Analytics



It's Official USA publishes more than anyone else

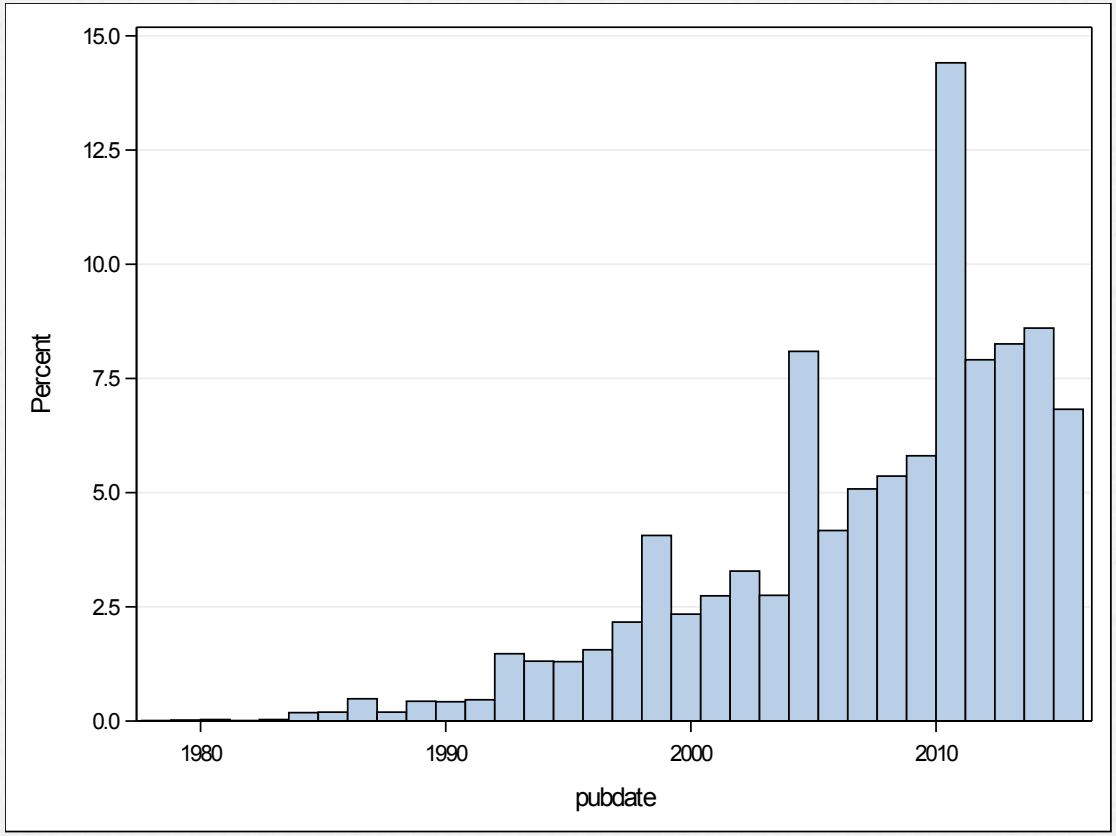
Row	Country_Name	numb
1	United States	2568
2	Japan	553
3	Germany	552
4	Italy	471
5	United Kingdom	398
6	France	229
7	Netherlands	223
8	Canada	218
9	China	213
10	Turkey	162
11	Australia	104
12	Spain	97
13	Belgium	89
14	India	82
15	Israel	76
16	Austria	74
17	Taiwan	55
18	Switzerland	55
19	Greece	54
20	Sweden	47

Journal Analytics



Row	Journal	Numb
1	Radiology	327
2	AJR Am J Roentgenol	323
3	J Magn Reson Imaging	302
4	Eur Radiol	236
5	Eur J Radiol	225
6	Breast J	190
7	Magn Reson Med	161
8	Breast Cancer	138
9	Breast Cancer Res. Treat.	129
10	Magn Reson Imaging	117
11	Rofo	116
12	Ann. Surg. Oncol.	102
13	Acad Radiol	101
14	Acta Radiol	97
15	Med Phys	94
16	J Comput Assist Tomogr	91
17	Plast. Reconstr. Surg.	86
18	Breast	83
19	J. Clin. Oncol.	82
20	Br J Radiol	82

Publication Date Analytics



Thank you

- Herve Momo
  - SAS Certified Base programmer for SAS 9
  - SAS Certified Advanced Programmer for SAS 9
  - Herve.momo@live.com
  - http://www.hervemomo.info/
  - Tel: 647-308-6075