Open Access Journal Publishing at UT Arlington:

An Analysis Using Academic Analytics Data in Combination with DOAJ Data

Clarke Iakovakis

March 2, 2014

1 Objective

To determine the scale of publishing in open access journals by UT Arlington academic departments, and the open access journals in which they publish.

2 Introduction

The scale of open access publishing in both green and gold forms has been steadily increasing. As more open access journals become sustainable and reputable, the scale in which researchers want to publish there increases. Furthermore, institutional open access policies, such as those passed by faculties at Harvard, MIT, Oregon State University, and most recently, the University of California System, will set a standard for university-supported publishing that many smaller institutions will want to copy.

Testimonies by those individuals at institutions which passed OA policies indicate the importance of building momentum and education behind the movement, in order to ensure that faculty voices are leading the call for a change in publication norms, which is critical for success. Therefore in the early stages of building the movement, it will be useful to determine which departments are already publishing in open access journals. These individuals will not only be familiar with the process of publication–which may vary somewhat from publication in toll-access journals–and perhaps more importantly may understand the virtues of open access publishing, from the increase in citation metrics to the larger practical and ethical benefits of making their research more accessible.

3 Academic Analytics Data

In order to discover which open access journals faculty are publishing in, one of course needs data on faculty publications. If your institution happens to subscribe to Academic Analytics (AA), this is one source for this data. Academic

Analytics is a subscription database providing metrics on publication counts, citation counts, research funding, and awards to faculty. According to their homepage, "The Academic Analytics Database (AAD) includes information on over 270,000 faculty members associated with more than 9,000 Ph.D. programs and 10,000 departments at more than 385 universities in the United States and abroad" (*Academic Analytics "What We Do"*, n.d.).

Academic Analytics does not provide any information as to whether or not journals are open access. AA does collect data on specific journals that individual faculty members publish in, as inferenced by their provision of "a *numeric* tally of each faculty members total scholarly productivity in each of the five areas of scholarly research (journal articles, citations, books, research grants and honorific awards)" (emphasis mine); nonetheless, the micro-level data is not accessible. However, AA provides publication data aggregated by academic department, through the "Department Articles Market Share" page. For the purposes of the following study, I used only the following two variables. The variables are not explicitly defined within the table, nor could I find a codebook specifically defining these particular variables, therefore the following definitions are mine:

- Journal Name: This is the primary key for the table. It is the name of the journal in which researchers for that department have published.
- Unit Articles: Number of articles published in the specified journal by UT Arlington researchers, aggregated by unit (i.e. department)

In the case of UT Arlington, this provides data on 48 departments (called "Units" in the table), including variables on the following list.

3.1 Scope of Academic Analytics Journal List

The first need is to establish the coverage of Academic Analytics journals. As Scopus is a well-regarded indexing service, the extent to which the set of AA journals is accounted for in Scopus will be telling.

Scopus is Scopus provides an updated title list (For this particular research question, the only variables of interest are the Title variable, i.e. the title of the publication, and the Type variable, classifying the item type. All data requires some cleaning before it can be analyzed, and will have the following transformations applied:

- Conversion to uppercase
- Leading/trailing whitespace deleted
- Duplicate entries deleted

This is especially important because the journal lists will be compared, and any variations in capitalization or spacing will lead to false negatives. Cleaning the data reveals 156 duplicates, leaving 34,119 unique titles. This can then be tabulated to give us a count of item types in the Scopus data.

```
library(stringr)
# Clean the data: removing whitespace and converting to uppercase
scopus.titles <- data.frame(scopus$Title, scopus$Type) # get
    list of Scopus titles
scopus.titles$scopus.Title <- factor(scopus$Title) # convert to
    factor
scopus.titles$scopus.Title <-
    str_trim(scopus.titles$scopus.Title, side = "both") # trim
    extra spaces on doaj list
scopus.titles$scopus.Title <-
    toupper(scopus.titles$scopus.Title) # convert to upper case
scopus.titles$scopus.Type <- factor(scopus$Type) # convert to
    factor
scopus.titles$scopus.Type <- str_trim(scopus.titles$scopus.Type,
    side = "both") # # trim the whitespace (for Book Series)
scopus.titles$scopus.Type <- toupper(scopus.titles$scopus.Type)
    # convert to upper case
dupe.c <- duplicated(scopus.titles$scopus.Title) # logical
    vector of duplicates
scopus.list <- scopus.titles[!dupe.c,]
```

Here the **scopus.titles** dataframe is created that includes only the Title and Type variables. Whitespace is removed using the *stringr* package, and the *toupper* function is used to convert the titles to uppercase. Finally a logical vector **dupe.c** (YES/NO) of duplicated entries is created using the *duplicated* function, and the **scopus.titles** dataframe is subset using this vector to return the full set of titles.

Figure 1: Clean Scopus Titles

```
scopus.type.table <-
    as.data.frame(table(scopus.list$scopus.Type)) # Get the Freq
    for Type
names(scopus.type.table) <- c("Type", "Count")
```

Figure 2: Tabulate Scopus types

Type	Count
BOOK SERIES	794
CONFERENCE PROCEEDINGS	817
JOURNAL	31719
TRADE JOURNAL	789

Table 1: Content in Scopus

```
doaj.titles <- data.frame(doaj$Title) # get list of DOAJ titles
aa.titles <- data.frame(aa.journals$AAD.2011.Journal.List) # get
    list of AA titles
aa.titles <- factor(aa.titles$aa.journals.AAD.2011.Journal.List)
    # convert to factor
aa.titles <- str_trim(aa.titles, side = "both") # trim extra
    spaces on aa list
aa.titles <- toupper(aa.titles) # convert to upper case
dupe.b <- duplicated(aa.titles) # logical vector of duplicates
aa.list <- aa.titles[!dupe.b] # return all AA journals as
    characters, in caps, without duplicates
aa.list.dupes <- aa.titles[dupe.b]
```

Unlike the Scopus analysis, which returns a dataframe of two variables, this function returns a character vector.

Figure 3: Clean Scopus Titles

The vast majority (93%) of these titles are journals; the distribution of types of titles is in Table 1:

Academic Analytics users can download a full list of publications that they index. Unfortunately, their data is proprietary and therefore this information is the use of University of Texas at Arlington employeess. The data comes in CSV format, with 218,469 observations of two variables: journal title and subject discipline. This data does not use journal title as a primary key, and many journals are classified in more than one disciplines; therefore it must be cleaned to remove duplicate entries, and it must be converted to uppercase to make intersection matching easier. This is done with the same basic code as for the Scopus data above, although for only the journal variable:

This returns a total of 14,586 unique titles.

Of those, how many are accounted for in 34,119 Scopus entries? Academic Analytics only provides data on academic journals, so it will likely only match entries from the 31,719 journals indexed by Scopus, but the full scopus list will be used nonetheless. The answer is easy to determine with a simple intersect function in R:

```
aa.scopus <- intersect(aa.list, scopus.list$scopus.Title)
```

```
doaj.titles <- data.frame(doaj$Title) # get list of DOAJ titles
doaj.titles <- factor(doaj.titles$doaj.Title) # convert to factor
doaj.titles <- str_trim(doaj.titles, side = "both") # trim extra
    spaces on doaj list
doaj.titles <- toupper(doaj.titles) # convert to upper case
dupe.a <- duplicated(doaj.titles) # logical vector of duplicates
doaj.list <- doaj.titles[!dupe.a] # return all DOAJ titles as
    characters, in caps, without duplicates (9,786)
doaj.list.dupes <- doaj.titles[dupe.a] # return all duplicated
    journals from the DOAJ list (18)
```

Figure 4: Delete Leading/Trailing Whitespace & Capitalize DOAJ Titles

There are 10,809 publications indexed by Academic Analytics that also appear in the Scopus data. AA journals account for approximately one third (31.6%) of journals indexed by Scopus journals, and if trade journals are included, AA journals account for 33.3%. AA journal selection criteria is not available, therefore comment or observation is not possible. Because there is no authoritative data on UT Arlington publications aside from Academic Analytics, it is also impossible to make substantive commentary as to how well AA data covers UT Arlington publications.

4 Directory of Open Access Journals in Academic Analytics

The overall purpose of the current study is to determine the extent of publication by UT Arlington researchers in open access journals. The Directory of Open Access Journals provides the broadest scope of open access journal publications. DOAJ is How well are Directory of Open Access Journals covered in Academic Analytics? This we find using the same procedure as above. The DOAJ data can be freely downloaded on the Web from The initial dataset includes 9,804 observations and a number of variables, including ISSN, The only pertinent variable for this study is the journal name.

First, the DOAJ data must be cleaned, The first two problems can be solved using the same procedure thus far followed: trim the whitespace using the stringr package and capitalize all the letters in the string. After removing duplicates, there are 9,786 unique titles in the DOAJ. To find all journals in common between the two lists, *all* spaces can be deleted in both DOAJ and AA lists (by replacing spaces with nothing): This finds 1,226 titles, which is the maximum number of journals the two lists have in common. Therefore, roughly 12.5% of DOAJ journals are indexed in Academic Analytics. Unfortunately, the above function returns journal names that look like

```
]CELLULARPHYSIOLOGYANDBIOCHEMISTRY
```

```
doaj.repl <- str_replace_all(doaj.list, pattern = " ", repl="")
    # delete ALL spaces on doaj list
aa.repl <- str_replace_all(aa.list, pattern = " ", repl="") #
    delete ALL spaces on aa list
doaj.aa.repl <- intersect(doaj.repl, aa.repl) # 1,226 common
    journals after deleting all spaces
```

Figure 5: Intersection of Trimmed DOAJ & AA Lists

```
doaj.aa.trim.repl <- str_replace_all(doaj.aa.trim, pattern = "
    ", repl="") # delete ALL spaces on trimmed list list
all.repl <- doaj.aa.repl %in% doaj.aa.trim.repl # get all values
    in the deleted space list that are in the trimmed space list
missing <- doaj.aa.repl[!all.repl] # Number of journals that
    appear in both lists but have punctuation problems aside
    from leading/trailing whitespace
```

Figure 6: Delete Leading/Trailing Whitespace & Capitalize DOAJ Titles

or

`EURASIPJOURNALONAUDIO,SPEECH,ANDMUSICPROCESSING.`

These titles will later be used to generate graphs; therefore, it would be better to avoid using that list. Fortunately, the number of journals falling into issue three are very small (27), and can be returned into their own vector by first deleting all spaces in the list, and comparing that list to the complete list of By comparing the list of common journals that have been found only with trimmed space with the journals that have been found with both trimmed and deleted space, the remainder will be those that are on both lists, but are only on the list of journals with deleted space:

Therefore, when the comparison between UT Arlington publications in Academic Analytics is run against the DOAJ list, it will be run first against the doaj.list variable, and second against a very small

doaj.aa.trim ¡- intersect(doaj.list, aa.list)

The intersection function finds 1,199 common journals.

Unfortunately, there is not an easy way to fix the third issue and still preserve the integrity of the character string, which is important in this study because the journal names will be graphed for each discipline. Fortunately, in this case, the number of journals that fall into this category are small. Therefore the AA data can be examined therefore a separate vector of these can be created to run a second loop for the function.

Then, another string

1. Delete all whitespace in DOAJ list

2. Compare that list to the AA list

After accounting for duplicates, Scopus provides an updated title list (http://www.elsevier.com/online-tools/scopus/content-overview), which as of February 2014 included 34,276 titles. The vast majority of these titles are journals; the distribution of types of titles is below:

Academic Analytics provides a full list of their journals through their database. The following data is proprietary and for internal University of Texas at Arlington staff use only.

After reading in the CSV file, it requires a bit of cleanup. First we create a vector of the title names from the Academic Analytics file, and coerce the vector to a factor. We convert it to all upper case, as some capital letter naming conventions vary

I first

it's vital to establish the extent to which the Academic Analytics journal list includes journals indexed by the Directory of Open Access Journals. To do this I downloaded the full list of journal names indexed by AA

The Department Articles Market Share page provides data on 48 departments (called Units in the table) at UT Arlington, including the variables on

5 Literature Review

The UT Arlington Libraries is a strong advocate for open access to scholarly information; that is, "digital, online, free of charge, and free of most copyright and licensing restrictions." (Suber, 2012)

References

Academic Analytics "what we do". (n.d.). http://www.academicanalytics.com/Public/WhatWeDo. (Accessed: 2014-02-22)

Suber, P. (2012). *Open access.* Cambridge, Mass.: MIT Press.

A R Code for Open Access Journal Coverage in Academic Analytics