Open Access and Institutional Research Impact

PA 397C Data Management and Research Life Cycle | Spring 2019

CAIFAN DU
School of Information, The University of Texas at Austin

What is Open Access (OA)?

- *Open Access (OA) literature*: "digital, online, free of charge, free of most copyright and licensing restrictions" (Suber, 2012)
  - Reduces access barriers (price & copyright)
  - Claimed benefits: Increased visibility of research work & larger audience

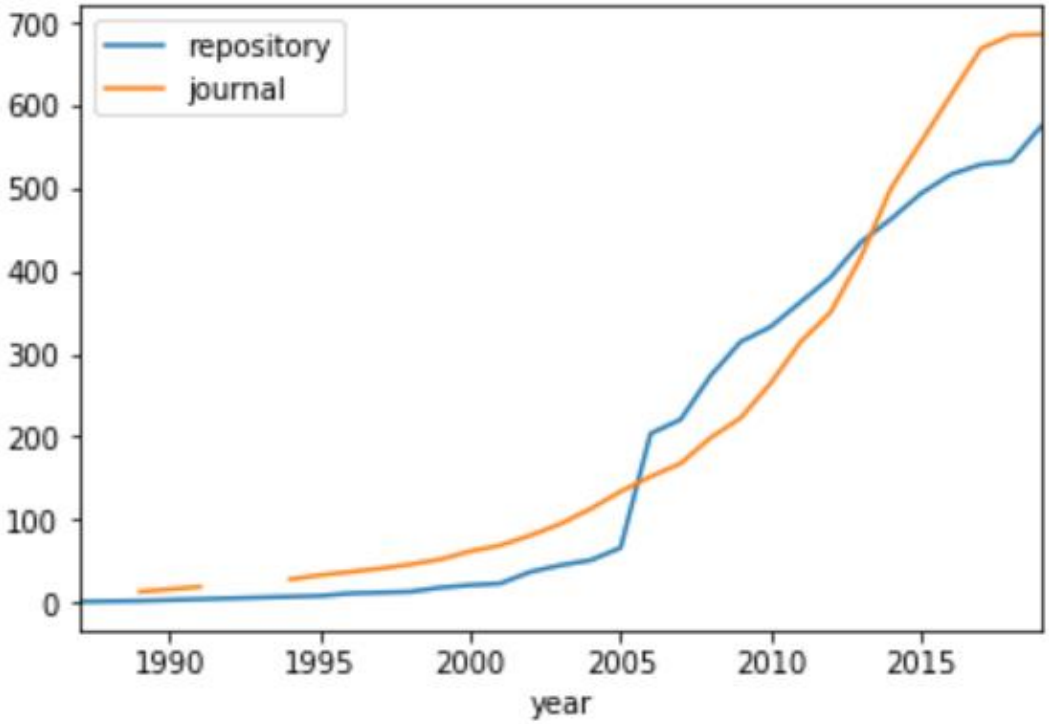
How can OA work out?

- Existing research indicates that journals turning to open access have a rise in their citation impact.
- Researchers produce academic work for impact instead of financial rewards.
- However, many academics are still not familiar with their OA options.

OA Mandates and Policies

- **Policy-makers**:
  - funding agencies and higher education institutions
- **2 OA options**:
  - *Gold OA* (journals) & *Green OA* (repositories)

Gold OA & Green OA



OA Mandates and Policies

- **Mandates** as *requirement*:
  - Only works for Green OA

  (As only about ¼ of peer-reviewed journals are OA)
- Institutions also host their own OA repositories
  - For knowledge management, digital preservation, and research impact, and so on

Research Question

- Do institution-hosted OA repositories and their OA policies have an effect on institutional research impact?

IV: 1) age of institution-hosted OA repositories

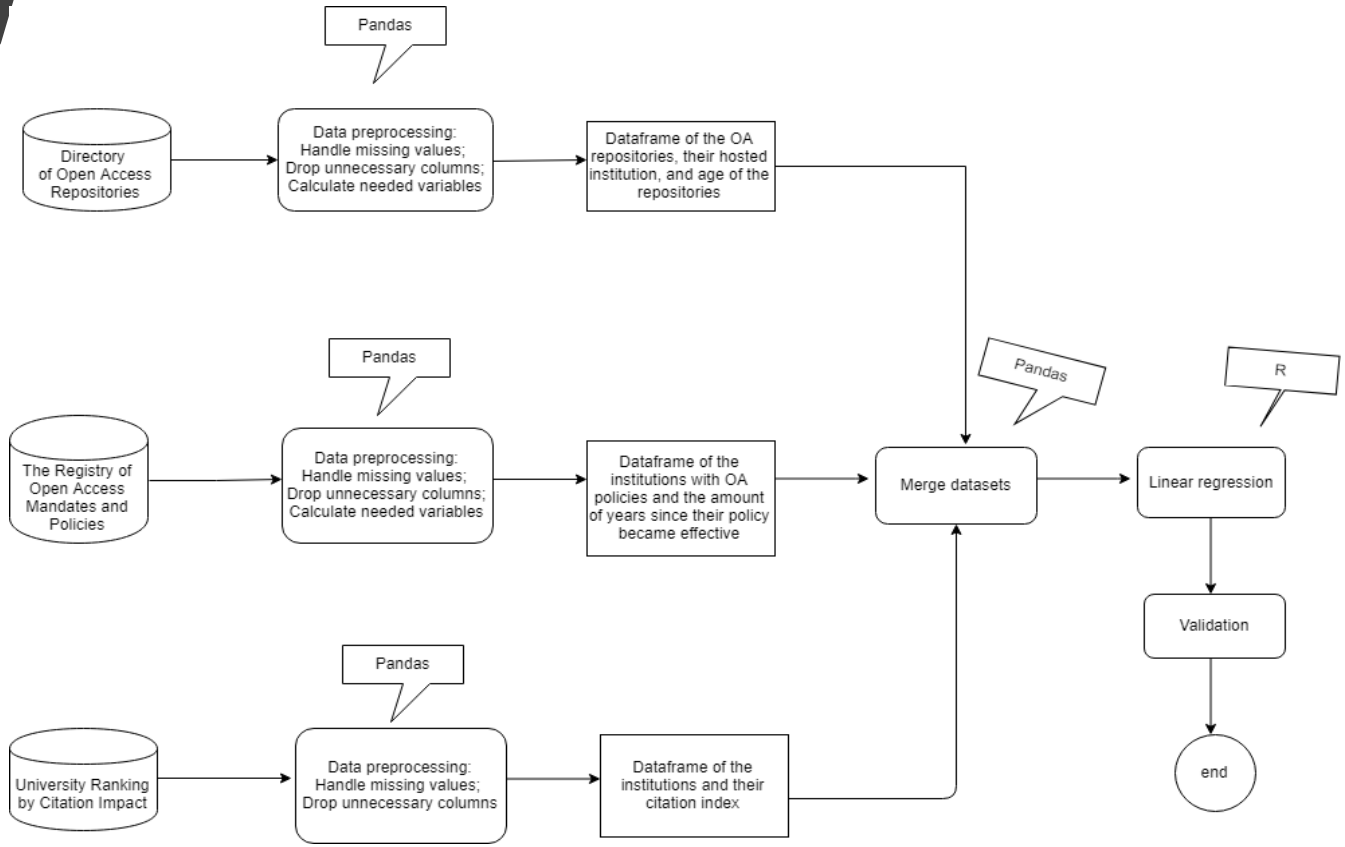
2) # of years since institutional OA policies become effective

DV: institutional research impact

Data Sources

- Directory of Open Access Repositories (DOAR) **(self-reported)**
  - Institution, year_established
- Registry of Open Access Mandates & Policies **(self-reported)**
  - Institution, policy_effective
- Ranking Web of Universities (Cybermetrics Lab, Spanish National Research Council)
  - Institution, citation_impact
- The Center for World University Rankings (CWUR) **(for validation)**
  - Intitution, citation_impact

Workflow


Pandas
Directory of Open Access Repositories
Data preprocessing: Handle missing values; Drop unnecessary columns; Calculate needed variables
Dataframe of the OA repositories, their hosted institution, and age of the repositories
Pandas
The Registry of Open Access Mandates and Policies
Data preprocessing: Handle missing values; Drop unnecessary columns; Calculate needed variables
Dataframe of the institutions with OA policies and the amount of years since their policy became effective
Pandas
University Ranking by Citation Impact
Data preprocessing: Handle missing values; Drop unnecessary columns
Dataframe of the institutions and their citation index
Pandas
Merge datasets
R
Linear regression
Validation
end

Data Preprocessing

```
import pandas as pd

# 1.1 import first dataset: university rankings
urank = pd.read_csv("https://raw.githubusercontent.com/caifand/DMRL_THA/master/FinalPaper/0_data_preprocessing/0_raw/cwur.csv")
```

```
# 2.3 Deal with missing values

# get needed data format
oa_repo.dtypes
oa_repo['date_created'] = oa_repo['date_created'].astype('datetime64')
oa_repo['date'] = pd.DatetimeIndex(oa_repo['date_created']).year
```

```
# replace missing values in 'year_established' column with corresponding values in 'date' column
oa_repo.year_established.fillna(oa_repo.date, inplace=True)
```

```
# 2.4 Add new values

# calculate the age of the OA repositories

oa_repo['age'] = 2019 - oa_repo['year_established']
```

```
# group by insitution and sort out the first entry within each group
cite_repo = cite_repo.groupby('institution', as_index=False)
cite_repo = cite_repo.first()
```

```
# merge the third dataset oap
repo_p = cite_repo.merge(oap, on='institution', how='inner')
repo_p
```

Data Management

- FinalPaper
  - 0_data_preprocessing
    - 0_raw
    - 1_code
    - 2_output
  - 1_analysis
  - 2_draft
  - 3_presentation

File Organization & Naming Conventions

Branch: master DMRL_THA / FinalPaper /

cfandu add presentation files

..	
0_data_preprocessing	add validation data and workflow chart
1_analysis	add validation data and workflow chart
2_draft	add validation data and workflow chart
3_presentation	add presentation files
workflow.png	add validation data and workflow chart

Back-up & Syncing

	A	B	C
1	variable_name	data_type	description
2	institution	string	name of high education institution
3	citation	integer	indicator of the citation impact of high education institutions
4	repo_age	float	the amount of years since the OA repository was established
5	policy_year	float	the amount of years since the institution's open access policy

DMRL_THA

This repository is dedicated to all the class assignments completed in the course Data Management and Research Cycle, Spring 2019.

Files in the repo

1. THA1 contains sample data compiled from three source datasets. These datasets are mainly about open access journals and journal rankings. In response to THA1, I also created a notebook file documenting profiles of these datasets and the sample compilation process. See THA1.ipynb.

Documentation & Data Dictionaries

Data Analysis

```{r}
reg_v <- lm(citation~repo_age+policy_year, data=vd)
summary(reg_v)
```

```
Call:
lm(formula = citation ~ repo_age + policy_year, data = vd)

Residuals:
    Min      1Q  Median      3Q     Max 
-477085 -275232 -113733  247039  807656 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    -8222     220089  -0.037    0.971
repo_age       17546      13737   1.277    0.214
policy_year    43874      29832   1.471    0.154

Residual standard error: 385200 on 24 degrees of freedom
Multiple R-squared:  0.1755,	Adjusted R-squared:  0.1068 
F-statistic: 2.555 on 2 and 24 DF,  p-value: 0.09866
```

Validation

```{r multiple regression}
reg <- lm(citation~repo_age+policy_year, data)
summary(reg)
```

```
Call:
lm(formula = citation ~ repo_age + policy_year, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-176.32  -89.99  -46.86   40.40  377.27 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)  258.395    114.998   2.247   0.0382 *
repo_age       1.623      6.407   0.253   0.8030  
policy_year  -19.294     12.643  -1.526   0.1454  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 150.6 on 17 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.1216,	Adjusted R-squared:  0.01821 
F-statistic: 1.176 on 2 and 17 DF,  p-value: 0.3323
```

***Note**: Even though the validation result seems imply the better dataset is being used, in the two compared datasets, citation impact is computed in different ways*

Reflections

- Challenge 1 (*operationalization*):
  - How to find the empirical counterpart to your conceptualization?
  - How to find the conceptual counterpart of your empirical evidence?
- Challenge 2 (*technical*):
  - String matching is a pain!

Reflections

- Challenge 3 (*data management*):
  - The concern about quality and methodological transparency of online data sources
- Positive side:
  - Data management habits are built bit by bit
- For my research questions…

Thank you!