**Open Access and Institutional Research Impact: Toward the Effect of Institutional Efforts in Promoting Open Access**

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

**I. Introduction**

Open access (OA) is a type of access to online digital goods. The development of the Internet is one condition that makes it possible, since the Internet affords the near zero marginal cost of the distribution of digital goods. In the academic domain, making publications open access reduces the barrier that price inflation of scholarly publishing and copyright restriction erect (Hess & Ostrom, 2007). It is claimed that open access publishing can bring research products to a larger audience and thus increase their impact. Existing research indicates that journals transforming to open access have a rise in their citation impact (Suber, 2012). Then, increased citation impact as a potential outcome should be able to motivate scholars to make their work open access. However, many scholars are still not familiar with their open access options.

Generally, there are two open access options available for scholars: Gold and Green OA. Gold OA is the term for open access journals, while Green OA refers to OA delivered by repositories. That is, scholars can choose to publish in OA journals or deposit their research products in OA repositories. In the meantime, funding agencies, research institutions, and universities with their mission to advance research usually have the same motivation to increase the availability and impact of their funded and supported research. In the United States, more than 130 institutions have adopted strong open access policies (ROARMAP, 2019). Some of them develop OA mandates, which are stronger policies that require researchers to make their work open access. Mandates only work for Green OA as only about a quarter of peer-reviewed journals are open access. If Gold OA is required, then researchers' publishing freedom will be greatly constrained. The weaker type of institutional OA policies is recommendation or encouragement policies. They target both Green and Gold OA equally (Suber, 2012).

Institutions sometimes host their own OA repositories. This effort is not only for assisting open access deposit, but also serves as an infrastructural work for preserving and promoting the research work produced at a given institution. Overall, institutions engage in open access expect a positive change in their research impact. Thus, the research question comes: Do institutional efforts in open access have an effect on their research impact?

**II. Operationalizing Research Impact and Institutional Efforts in Open Access**

Educational and research institutions take the responsibilities of generating and transforming knowledge for the benefits of the whole society. Their research impact is often evaluated for their managerial aims, informing funding decisions, and accounting for their values to relevant stakeholders and the general public (Penfield et al., 2014). The measurement of research impact of an institution usually depends on the purpose of evaluation. The assessment of research impact can be a proxy for academic quality and productivity, or it can demonstrate the socio-economic benefits that the research outcome generates.

Traditionally, bibliometric methods are used to derive indicators of academic impact. They translate research publications and citations into quantifiers that reflect scientific productivity or quality. For instance, the citation count of an academic publication is a direct measure of its usefulness to other scholars (Bornmann & Marx, 2012). Such quantifiers and techniques are used to assess the productivity of researchers, research institutes, and journals.

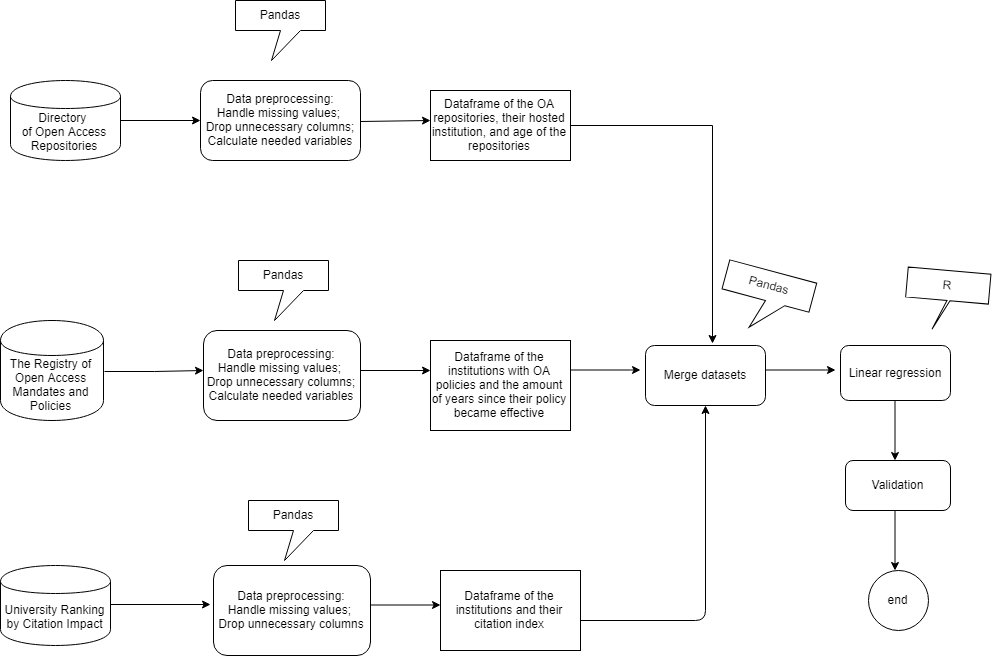
Given that the current discussion about open access is focused on the academic domain, the operationalization of research impact should also have a scholarly orientation. In many university rankings, citations are used to estimate the research impact of higher education institutions. Thus, in this study, the institutional research impact is operationalized as its citation impact. Since the indicators of citation impact are usually calculated only for higher educational institutions, in this study I only analyze the association between OA repositories and policies and the research impact for universities. Other institutions such as research institutes and funding agencies are not included in this study.

Regarding institutional efforts in open access, it has been discussed earlier in the article that institutions adopt OA policies and host OA repositories. Since both approaches will need time to take effect, I use the time span since the initial date of institutional OA policies or repositories as measures of institutional efforts in open access.

**III. Dataset**

**1. Dataset Compiling**

Figure 1 demonstrates the workflow for dataset compiling and analysis. The final merged dataset for analysis contains *institutional names*, their *citation impact indicators*, *age of* *OA repositories*, and *the number of years since the OA policy is adopted*. The merged data table and a data dictionary explaining the variables in the table can be found in Appendix A.



*Figure 1: Workflow for dataset compiling*

The dataset is compiled from 3 online data sources:

1. The citation impact indicator is obtained from *the Ranking Web of Universities (Cybermetrics Lab, Spanish National Research Council, CSIC)*. This research lab updates the worldwide university rankings every 6 months. The retrieved dataset was updated in January 2019.

2. The age of OA repository is computed using the date of repository creation from the *Directory of Open Access Repositories (OpenDOAR).* It is a global directory of academic open access directories hosted by University of Nottingham and Lund University. The directory has its criteria for collecting and receiving information about OA repositories. Only repositories without any form of access control on full-text resources will be accepted by the directory.

For repository records the date of creation is absent, the date when the repository record is registered into the directory is used to replace the missing value.

3. The number of years since the OA policy became effective is computed using the year when institutional OA policy became effective. This information is obtained from the Registry of Open Access Repository Mandates and Policies (ROARMAP). It is an international registry where over three thousands of institutional and cross-institutional repositories have been registered.

When the year when OA policy became effective is missing, the year when OA policy was adopted is used to replace the missing value.

Note that all three retrieved datasets are about universities in the United States. The web URL for retrieving the three datasets can be found in Appendix B.

**2. Data Analysis**

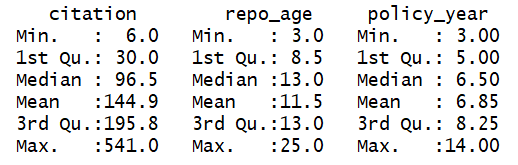
Multiple regression model is used to model the association between age of OA repositories, number of years since OA policy became effective, and the citation impact of institutions:

*Citation* = α + β1 • *RepoAge* + β2 • *YearEffective*

**IV. Results**

**1. Descriptive Analysis**

Figure 2 shows the summary statistics of major variables.

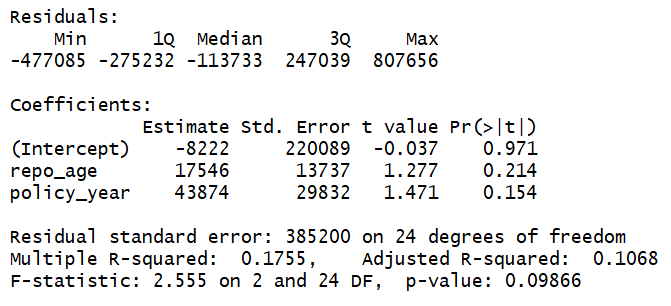


*Figure 2: Summary Statistics of Major Variables*

The summary statistics of major variables demonstrate that, for the universities of interest, on average the age of OA repositories they host is approximately twice the time length of their effective OA policies.

**2. Regression Analysis**

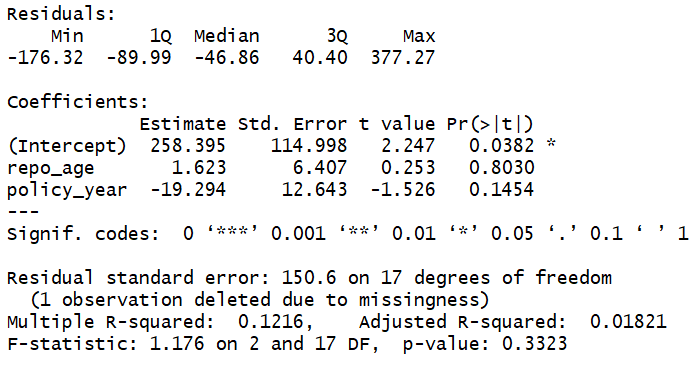
Figure 3 shows the regression results of the multiple linear regression model. According to the results, the association between two variables representing institutional engagement in open access and the citation indicator is not significant.



*Figure 3: Results of the Regression Model of the Compiled Dataset*

**V. Validation**

Another dataset of university ranking is used for running the same regression model in order to check the quality of the citation impact data. The validation data set is the citation impact indicators from the university ranking in the United States in 2018-2019, retrieved from the *Center for World University Ranking (CWUR)*. Using this dataset instead of the citation impact indicators retrieved from the *Ranking Web of Universities*, the regression results are demonstrated in Figure 4:



*Figure 4: Results of the Regression Model of the Validation Dataset*

From the validation results, we can see that there is an increase in *p-value* and a decrease in the *F-statistic*, which may indicate that our dataset of citation impact has better quality than that of the validation dataset. However, the citation impact from the two datasets is calculated in different ways. The Ranking Web of Universities, which is the source of the citation impact indicator being used for analysis, does not disclose its full methodology of measure computation. It is just mentioned on its website that the OA repositories of institutions are taken into consideration when computing the citation impact indicators. It remains unknown whether the citation impact indicator from the Ranking Web of Universities contains the component similar to the age of institutional OA repositories. Thus, the validity of the used dataset is uncertain.

The merged validation dataset and the corresponding data dictionary can be found in Appendix A. The web URL of the source of citation impact data for validation can be found in Appendix B.

**VI. Discussion and Conclusion**

I suspect two main possible reasons for the insignificance of regression analysis. First, the quality of the retrieved dataset is a concern. The citation impact indicator is a composite measure and its components are unknown. Also, after data preprocessing and merging, the available sample size for regression is relatively small (N = 20). All these factors could lead to invalid regression.

Second, more variables may need to be introduced to explain the change of institutional research impact. In this regard, a more thorough literature review should be conducted to bring control variables into the current model. The current regression model is also highly constrained by the availability of existing datasets. It fails to tell a well-rounded story in response to my research question.

**References**

Bornmann, L., & Marx, W. (2013). How good is research really?: Measuring the citation impact of publications with percentiles increases correct assessments and fair comparisons. *EMBO Reports*, *14*(3), 226–230. <https://doi.org/10.1038/embor.2013.9>

Hess, C., & Ostrom, E. (2007). *Understanding Knowledge as a Commons: From Theory to Practice*. MIT Press.

Suber, P. (2012). *Open Access*. MIT Press.

The State of OA: A large-scale analysis of the prevalence and impact of Open Access articles. ProQuest. Retrieved May 10, 2019, from <https://search.proquest.com/docview/1952466686?pq-origsite=gscholar>

**Appendix A. Tables and Data Dictionaries.**

***Table 1****: Institutions and Their Citation Impact, Age of OA Repository, and Number of Years Since OA Policy Became Effective*

|  |  |  |  |
| --- | --- | --- | --- |
| **institution** | **citation** | **repo\_age** | **policy\_year** |
| Boston University | 26 | 13 | 4 |
| California Institute of Technology | 31 | 13 | 5 |
| Cornell University | 40 | 13 | 14 |
| Duke University | 14 | 11 | 9 |
| Emory University | 63 | 3 | 8 |
| Florida State University | 139 | 11 | 3 |
| Georgia Institute of Technology | 222 | 13 | 6 |
| Massachusetts Institute of Technology | 6 | 13 | 10 |
| Northern Illinois University | 327 | 7 | 5 |
| Oregon State University | 187 | 15 | 6 |
| Princeton University | 27 | 4 | 8 |
| Rice University | 108 | 20 | 7 |
| University of Colorado Boulder | 59 | 3 | 5 |
| University of Delaware | 405 | 15 | 4 |
| University of Kansas | 121 | 13 | 10 |
| University of Massachusetts Amherst | 121 | 9 | 3 |
| University of Pennsylvania | 13 | 25 | 8 |
| University of Rhode Island | 541 | 13 | 6 |
| University of Virginia | 85 | 6 | 9 |
| Utah State University | 363 | 10 | 7 |

*Data Dictionary of Table 1:*

|  |  |  |  |
| --- | --- | --- | --- |
| **variable\_name** | **data\_type** | **description** | **source** |
| institution | string | name of high education institution | Directory of Open Access Repositories (DOAR); The Registry of Open Access Repositories Mandates and Policies (ROARMAP); The Center for World University Ranking (CWUR) |
| citation | integer | indicator of the citation impact of high education institutions | The Center for World University Ranking (CWUR) |
| repo\_age | float | the number of years since the OA repository was established | Calculated by the year\_established in Directory of Open Access Repositories (DOAR) |
| policy\_year | float | the number of years since the institution's open access policy became effective | Calculated by the policy\_effective in the Registry of Open Access Repositories Mandates and Policies (ROARMAP) |

***Table 2:*** *Validation Dataset*

|  |  |  |  |
| --- | --- | --- | --- |
| **institution** | **citation** | **repo\_age** | **policy\_year** |
| Abilene Christian University | 5425 | 3 | 2 |
| Allegheny College | 14851 | 2 | 6 |
| Boston University | 1203027 | 13 | 4 |
| California State University Northridge | 59422 | 6 | 6 |
| Connecticut College | 18650 | 13 | 6 |
| Cornell University | 773925 | 13 | 14 |
| Drake University | 21191 | 13 | 6 |
| Duke University | 1072992 | 11 | 9 |
| Emory University | 597192 | 3 | 8 |
| Florida State University | 399057 | 11 | 3 |
| Georgia Institute of Technology | 870242 | 13 | 6 |
| Lafayette College | 23596 | 9 | 8 |
| Massachusetts Institute of Technology | 1234793 | 13 | 10 |
| Northeastern Illinois University | 9665 | 0 | 3 |
| Northern Illinois University | 59617 | 7 | 5 |
| Oregon State University | 304501 | 15 | 6 |
| Princeton University | 1065331 | 4 | 8 |
| Rice University | 776833 | 20 | 7 |
| University of Colorado Boulder | 556075 | 3 | 5 |
| University of Delaware | 255812 | 15 | 4 |
| University of Kansas | 339411 | 13 | 10 |
| University of Massachusetts Amherst | 793467 | 9 | 3 |
| University of North Texas | 227760 | 8 | 7 |
| University of Pennsylvania | 785709 | 25 | 8 |
| University of Rhode Island | 207914 | 13 | 6 |
| University of Virginia | 443637 | 6 | 9 |
| Utah State University | 134644 | 10 | 7 |

*Data Dictionary of Table 2*

|  |  |  |  |
| --- | --- | --- | --- |
| **variable\_name** | **data\_type** | **description** | **source** |
| institution | string | name of high education institution | Directory of Open Access Repositories (DOAR); The Registry of Open Access Repositories Mandates and Policies (ROARMAP); Ranking Universities by cybermetrics Lab, Spanish National Research Council |
| citation | integer | indicator of the citation impact of high education institutions | Ranking Universities by cybermetrics Lab, Spanish National Research Council |
| repo\_age | float | the number of years since the OA repository was established | Calculated by the year\_established in Directory of Open Access Repositories (DOAR) |
| policy\_year | float | the number of years since the institution's open access policy became effective | Calculated by the policy\_effective in The Registry of Open Access Repositories Mandates and Policies (ROARMAP) |

**Appendix B. URLs for Retrieving Source Datasets.**

Directory of Open Access Repositories (OpenDOAR). <http://v2.sherpa.ac.uk/opendoar/>

CWUR World University Rankings 2018-2019. <https://cwur.org/2018-19/usa.php>

Universities National | Ranking Web of Universities. <http://www.webometrics.info/en/North_america/natx>

Registry of Open Access Repository Mandates and Policies (ROARMAP). <https://roarmap.eprints.org/>