CHAPTER 15

Conducting a Baseline Diversity, Equity, and Inclusion Assessment of Institutional Repository Content

Rebekah Kati

INTRODUCTION

Diversity, equity, and inclusion (DEI) are central to repository and overall library work. Although DEI principles have been incorporated into many repository programs for digital collections, institutional repository initiatives have lagged. However, DEI principles can and should be applied to institutional repository collections to ensure equity and representation.

The Carolina Digital Repository (CDR) is the institutional repository for the University of North Carolina at Chapel Hill (UNC-CH) and aims to collect scholarly material that is representative of the research conducted at the university. In support of UNC-CH's open access policy, the university's libraries have launched three major content recruitment initiatives in the past four years: a large vendor-supplied open access article batch upload, ongoing CV review for faculty, and annual highly cited researchers batch uploads.[1] After loading content from these projects, it was apparent that the initiatives identified articles concentrated

in the sciences. UNC-CH has a strong humanities and social sciences focus, which it was feared would be obscured by the large import of science content. Additionally, it was suspected that the CDR might now reflect demographics that were not aligned with those of the university, and this would affect the CDR's mission of scholarly representation.

In 2021, the libraries conducted a baseline assessment of the CDR's content projects to see if they aligned with the demographics of the university. The outcomes of the assessment will inform resource allocation for future projects that promote DEI principles. This assessment looked at subject area coverage, author gender, and author self-identified race in all three CDR content initiatives. Results were compared with official UNC-CH demographics to determine if articles loaded as part of the projects reflected the demographics of the university and thus are broadly representative of the university's scholarly output. A white paper describing the findings in detail is available in the CDR.[2] This chapter explores the process used to conduct the assessment and reflect on lessons learned. It also presents key takeaways for readers interested in conducting their own assessments.

BACKGROUND: THE PROJECTS

In support of UNC-CH's Open Access Policy, the UNC-CH Libraries Open Access Implementation Team was tasked in 2017 with increasing the amount of faculty scholarship in the CDR.[3] The team identified three strategies, which were collectively named Content Liberation:[4]

1. Author citations/1foldr: Originally, CDR staff planned to conduct affiliation searches in the UNC-CH Libraries subscription databases. After the UNC-CH Libraries purchased a 1foldr report from 1Science, this project was adapted to load content from that report.
2. CV review: CDR staff planned to review faculty CVs for deposit-eligible scholarship on an as-needed basis. During the COVID-19 pandemic, this project was adapted into a work-from-home project for library workers and students.
3. Highly cited researchers: Using Clarivate's Highly Cited Researchers lists, CDR staff identified high-impact, deposit-eligible scholarship.

Although the team made great progress in increasing faculty content in the CDR, they wondered if the content that was added was representative of UNC-CH as an institution. For example, the 1foldr report enabled the deposit of over 28,000 articles into the CDR, but much of the content came from PubMed Central, which focuses on research in the biomedical and life sciences fields. While UNC-CH has a strong program in biomedical and life sciences, they are only one aspect of the university's research profile. By loading 28,000 articles in the biomedical and life sciences fields into the CDR, the team might have skewed the subject focus of the repository.

The CDR is charged with storing, preserving, and providing access to university scholarship, therefore the Open Access Implementation Team believed that the content in the CDR should be representative of the university. It follows that the content in the CDR should reflect the subject area, gender, and racial makeup of the university. These are only three aspects of diversity, but they are a starting point upon which further work can be built.

To discover the gaps in coverage, an analysis was conducted of the subject areas, gender, and racial makeup of authors included in the three approaches to content identification. The findings of the analysis were compared to official UNC-CH statistics to benchmark the CDR's performance. To be clear, this analysis should not be regarded as comprehensive, and limitations are noted in the sections below where appropriate. The goal of this project was to reveal general trends and inequities in CDR coverage that could be addressed in future initiatives.

SUBJECT AREA COVERAGE

UNC-Chapel Hill is the twelfth-largest research university in the United States.[5] Research occurs across the university's schools and colleges including medicine, public health, arts and sciences, education, pharmacy, and more. The purpose of the subject area assessment was to determine if the Content Liberation projects contained work from all the colleges and schools at UNC-CH and were therefore representative of the work of the university.

This part of the assessment used a consistent methodology for each Content Liberation project. Each author was assigned a subject classification based on their College or School affiliation within the university, based on the author's primary departmental affiliation listed in the article. For clarity, the College of Arts and Sciences was further subdivided into subject areas according to the categories listed on the college's website. If an article author only listed a university-level affiliation, they were assigned to an "Unknown" category.

The 1foldr portion of the assessment required a large amount of data cleaning, as CDR staff ingested over 28,000 articles to the CDR in 2020. Duplicate authors and authors who did not list a UNC-CH affiliation were removed. Since the same dataset was used for both the subject and gender analysis, authors who listed initials, rather than full first names were also removed. This process generated a dataset of 11,102 unique UNC-CH-affiliated authors. The subject analysis determined that 7,214 of these researchers work in the sciences. Only 195 researchers work in the humanities, social sciences, business, and law fields.

The CV review portion of the assessment generated a much smaller sample set of 426 faculty CVs. Library workers were asked to choose departments for review based on their own interests; 289 chosen researchers worked in the

humanities, social sciences, education, journalism, or social work fields, which were under-represented in the CDR.

Clarivate's Highly Cited Researchers list identified seventy-one unique UNC-CH researchers. Sixty-four out of seventy-one authors wrote in the sciences or medicine and only seven authors wrote in social sciences, business, or journalism.

WORK OF BLACK FACULTY, FACULTY OF COLOR, AND INDIGENOUS FACULTY IN THE CDR

In 2020, 73.9 percent of tenure and tenure-track faculty at UNC-CH identified as white. Only 11.8 percent of tenure and tenure-track faculty identified as Asian. The numbers were much smaller for tenure and tenure-track faculty who identify with other racial minority groups: 5.7 percent identified as Black, 5.3 percent identified as Hispanic, 0.9 percent identified as multiracial, 0.5 percent identified as American Indian or Alaskan Native and only 0.1 percent identified as Native Hawaiian or Pacific Islander.[6] For this part of the DEI assessment, the goal was to determine if scholarship produced by faculty who identify as Black, indigenous, or a person of color (BIPOC) had been deposited into the CDR as part of the Content Liberation projects.

On June 22, 2020, UNC-CH faculty publicly published a document titled "Black Faculty, Faculty of Color and Indigenous Faculty Roadmap for Racial Equity at the University of North Carolina at Chapel Hill."[7] This roadmap was signed by 815 supporters, including 144 faculty members who self-identified as BIPOC. This dataset was used as the basis of the analysis of BIPOC faculty because signatories voluntarily signed the widely circulated statement and publicly self-identified as a member of a minority community. A list of BIPOC faculty was compiled based on the self-identified signatories of the roadmap as well as from websites of UNC-CH-affiliated racial and ethnic affinity groups in which members had listed their names publicly, which brought the list to 154 authors in total. The BIPOC faculty were given the option to list their departmental affiliations, which were categorized into School and Colleges using the same process as the subject area assessment. Of course, this method of self-identification does not identify all members of BIPOC communities at UNC-CH, only those members who signed the roadmap and/or publicly identified themselves. It is very likely that this analysis under-represents contributions by BIPOC faculty to the CDR.

For this portion of the assessment, a consistent methodology was used for all three projects. The dataset of BIPOC faculty was small enough that it was possible to compare the list with searches in the CDR and the CV review and

Highly Cited Researchers lists. It was determined that 283 articles deposited in the CDR had been written by faculty on the BIPOC faculty list. Of the 1foldr articles, 198 of the 283 articles authored by BIPOC faculty were a part of the upload project. Thirty-six authors on the BIPOC faculty list had their CVs reviewed. None of the seventy-one UNC-CH Highly Cited Researchers appear on the BIPOC faculty list. Since UNC-CH has such a small percentage of tenure and tenure-track BIPOC faculty, these results are sadly not surprising. Frustratingly, UNC-CH's demographics also align with overall academic employment in the sciences, where white people make up 49.4 percent of tenured doctoral scientists and engineers.[8]

Results for the CV review project were more encouraging. Out of the 154 authors on the BIPOC faculty list, thirty-six had their CVs reviewed, representing 23 percent of the overall BIPOC list. The CV Review results may be due to the large number of humanities and social sciences researchers present on the list, which aligned with the interests of library workers working on the project. Furthermore, the prevalence of humanities and social sciences researchers on the BIPOC faculty list may explain their under-representation on the 1foldr report and Highly Cited Researchers list, as both the report and the list trended toward the sciences.

GENDER IN CDR

In 2020, 58.9 percent of tenure and tenure-track faculty at UNC-CH identified as male. Only 41.1 percent of tenure and tenure-track faculty identified as female.[9] For this part of the assessment, the aim was to determine whether scholarship produced by women had been deposited into the CDR as part of the Content Liberation projects.

Answering the research question for the 1foldr report proved to be tricky, as a list like the Roadmap for Racial Equity was not available for gender. Additionally, the 1foldr dataset is very large and contains older articles, which complicated the choice of methodology. To determine an appropriate approach, articles that asked similar research questions of large datasets were reviewed. Many of these articles used gender prediction services for all or part of their analysis. Gender prediction services are a common bibliometrics tool for investigating gender for large datasets. These services query large datasets containing name and gender data to determine the probability that a given name matches a particular gender. For each query, the service will typically return the number of records queried, a prediction of gender based on the query, and a score indicating the probability that the name matches the service's gender prediction.

There was initial reluctance to use such a service, as they can replicate or introduce inequities. The most used services from the survey did not account for genders other than male or female and may not reflect an individual's gender

identity. Additionally, while testing services, it was observed that several had trouble identifying gender for non-Western names, names that contained spaces or accent markers, and had low probability scores for gender-neutral names. Nevertheless, it was not feasible to manually identify gender for the large 1foldr dataset, so the decision was made to proceed with the gender prediction service while transparently disclosing their limitations.

The service genderize.io was chosen as it was free for up to 1,000 names per day and has a large dataset that seems to be updated regularly. In their assessment of gender prediction tools, Santamaría and Mihaljević determined the error rate of genderize.io to be under 15 percent.[10] Nevertheless, genderize.io was unable to predict a gender for 499 out of 11,102 names in the 1foldr report. When genderize.io returned a null value for a name, two other gender prediction services, GenderAPI and NamSor, were queried. If GenderAPI and NamSor did not agree on a likely gender for the name, the gender that had the highest probability score was chosen.

The process for the CV Review and Highly Cited Researchers projects was more straightforward and more equitable. Since the lists of researchers used in both projects were of a manageable size, web searches were conducted for faculty biographical statements and departmental news stories to determine the faculty member's preferred pronouns. This approach enabled the analysis to reflect the individual's preferred gender identity in a professional setting.

The gender breakdown of Content Liberation project content generally follows the gender trends in UNC-CH tenure and tenure-track faculty. The 1foldr report contains 960 more male-predicted names than female-predicted names. Fifty-four percent of the names in the 1foldr report were male-predicted, which is slightly less than the 58.9 percent of tenure and tenure-track male faculty at UNC-CH. The Highly Cited Researchers project had the starkest disparities, as fifty-six researchers (78.8 percent) used male pronouns and thirteen used female pronouns. As mentioned above, this is likely due to the subject breakdown of the 1foldr and Highly Cited Researchers report. Larivière et al. found that women tend to publish more in the social sciences, whereas men publish more in the sciences and humanities.[11] Given that the 1foldr and Highly Cited Researchers projects concentrated on content in the sciences, it is unsurprising that the results would be male-dominated. The gender distribution on the CV review project was much closer. Only eighteen more researchers used male pronouns than researchers using female pronouns.

KEY TAKEAWAYS AND NEXT STEPS

The Content Liberation initiatives replicate existing inequities in the academy in that they primarily deposited scholarship authored by white men in the sciences. This focus came about inadvertently during the inception of the

Content Liberation projects. In particular, the early days of the Content Liberation initiative focused on the Highly Cited Researchers project. The Open Access Implementation Team felt that highly cited content was a high priority for preservation and hoped that contacting prominent researchers might lead to an increased awareness of the CDR among faculty. While the team did preserve high-impact research, contacting prominent faculty did not lead to an increase in self-deposit. The team may also have added to the imbalance of scholarship. The DEI analysis shows that the Highly Cited Researchers from UNC-CH were overwhelmingly men doing research in the sciences who did not self-identify as BIPOC on the faculty list. This tracks with findings from the literature, which determined that articles with female first authors were cited less than male first authors.[12] It is expected that fewer female authors would be included on Clarivate's Highly Cited Researchers list.

It is difficult to admit that projects that were created with good intentions contribute to inequity. During the assessment and reporting process, there was a tendency to feel slightly defensive and protective of the projects that had been created. It quickly became clear that it is vital to move past personal feelings about the projects for them to grow and become more equitable for all. Depersonalization of one's work is important in order to keep improving, and this will continue to be in mind during future assessments.

Transparency in analysis was important. Not all assessments will be equitable, and it is important to disclose the process by which the assessment was completed and its limitations. This was most apparent during the gender assessment when using a gender prediction service. Although this was not an ideal approach to take, transparent disclosure of the strengths and weaknesses of the approach enabled readers to place the results in their proper context. Also, although the BIPOC faculty data sample was not fully representative of the BIPOC faculty at the university, disclosing the methodology and limitations also helped the reader to contextualize the results.

Additionally, the importance of library worker labor ran throughout the assessment. Many library workers participated in the CV review process, which identified and collected metadata and PDFs for inclusion in the CDR. However, more labor is needed to verify articles against the CDR's inclusion criteria and to prepare the metadata and PDFs for upload to the CDR. This will be an ongoing and time-consuming process but one that will hopefully increase the amount of research performed by members of underrepresented groups in the CDR.

Furthermore, populating an institutional repository can be an outreach opportunity for subject-area librarians. In late 2021, liaison librarians were engaged to help with the deposit process. They contacted researchers in their assigned subject areas to seek permission to deposit work. Liaisons were most comfortable contacting all researchers in their subject areas rather than researchers who

belonged to a particular demographic group. Since most of the participating liaisons work in humanities and/or social sciences areas, a broad approach works well to broaden the scope of CDR's content. The impact of this initiative on CDR's content will be assessed in the future.

Hopefully, the approaches above will be a first step toward broadening the subject area, race, and gender focus of the CDR, which will bring the CDR more in line with UNC Libraries' Reckoning Initiative. The team will continue to assess the progress of the Content Liberation initiatives and aim to make their outputs as equitable as possible.

NOTES

1. "Open Access Policy," University of North Carolina at Chapel Hill, accessed April 25, 2022, https://policies.unc.edu/TDClient/2833/Portal/KB/ArticleDet?ID=132180.
2. Rebekah Kati, "Subject Coverage, Gender and Race in Carolina Digital Repository Content," accessed April 25, 2022, https://doi.org/10.17615/a9v4-z352. Note that portions of this chapter were previously published in the white paper and are used here with permission.
3. The Open Access Implementation Team consists of the scholarly communications officer (Anne Gilliland), head, Repository Services Department (Julie Rudder), institutional repository librarian (Rebekah Kati) and open access librarian (currently vacant but was filled by Jennifer Solomon during the task force planning period in 2017).
4. A more detailed summary of each of these projects can be found at Rebekah Kati, "Content Liberation Project Summary," accessed April 25, 2022, https://doi.org/10.17615/xs4b-w127. Additional information on the 1foldr report is in Rebekah Kati, "Content Liberation Update, August 2020," accessed April 25, 2022, https://doi.org/10.17615/rmy0-kw79.
5. "About," UNC Office of Sponsored Research, accessed May 9, 2022, https://research.unc.edu/about/.
6. For faculty race/ethnicity demographic statistics, see "Permanent Full-Time Faculty and Post-Doctoral Fellows by Race/Ethnicity and Tenure Status, Fall 2010-2020" at https://oira.unc.edu/reports/.
7. "Black Faculty, Faculty of Color and Indigenous Faculty Roadmap for Racial Equity at the University of North Carolina at Chapel Hill," accessed April 25, 2022. https://docs.google.com/document/d/e/2PACX-1vQhpLlf5nWdUzTeD-CAB9wtS-cBd-Bk0V4uPllmEH5zwH6vszmXigDIUV3MmMAcwwkPzPEWRxzihZrn9/pub.
8. "Tenure and academic positions," National Center for Science and Engineering Statistics, Survey of Doctorate Recipients, 2019, accessed May 9, 2022, https://ncses.nsf.gov/pubs/nsf21321/report/academic-careers#tenure-and-academic-positions.
9. For gender demographics at UNC-CH, see "Permanent Full-Time Faculty and Post-Doctoral Fellows by Gender and Tenure Status, Fall 2010-2020," Office of Institutional Research & Assessment, https://oira.unc.edu/reports/.
10. Lucia Santamaría and Helena Mihaljević, "Comparison and benchmark of name-to-gender inference services," *PeerJ Computer Science* 4:e156, https://doi.org/10.7717/peerj-cs.156.
11. Vincent Larivière, Chaoqun Ni, Yves Gingras, Blaise Cronin, and Cassidy R. Sugimoto, "Bibliometrics: Global gender disparities in science," *Nature* 504:7479 (2013): 211–13, https://dx.doi.org/10.1038/504211a.

12. Larivière et al., "Bibliometrics."

BIBLIOGRAPHY

Kati, Rebekah. "Content Liberation Project Summary." Accessed April 25, 2022. https://doi.org/10.17615/xs4b-w127.

———. "Content Liberation Update, August 2020." Accessed April 25, 2022. https://doi.org/10.17615/rmy0-kw79.

———. "Subject Coverage, Gender and Race in Carolina Digital Repository Content." Accessed April 25, 2022. https://doi.org/10.17615/a9v4-z352.

Larivière, Vincent, Chaoqun Ni, Yves Gingras, Blaise Cronin, and Cassidy R. Sugimoto. "Bibliometrics: Global gender disparities in science." *Nature* 504:7479 (2013): 211–13. https://dx.doi.org/10.1038/504211a.

National Center for Science and Engineering Statistics, Survey of Doctorate Recipients. "Tenure and academic positions." 2019. Accessed May 9, 2022. https://ncses.nsf.gov/pubs/nsf21321/report/academic-careers#tenure-and-academic-positions.

Santamaría, Lucia, and Helena Mihaljević. "Comparison and benchmark of name-to-gender inference services." *PeerJ Computer Science* 4:e156. 2018. https://doi.org/10.7717/peerj-cs.156.

UNC-CH Office of Institutional Research and Assessment. "Permanent Full-Time Faculty and Post-Doctoral Fellows by Gender and Tenure Status, Fall 2010–2020." Accessed May 9, 2022. https://oira.unc.edu/reports/.

———. "Permanent Full-Time Faculty and Post-Doctoral Fellows by Race/Ethnicity and Tenure Status, Fall 2010–2020." Accessed May 9, 2022. https://oira.unc.edu/reports/.

UNC Office of Sponsored Research. "About." Accessed May 9, 2022. https://research.unc.edu/about/.

University of North Carolina at Chapel Hill. "Black Faculty, Faculty of Color and Indigenous Faculty Roadmap for Racial Equity at the University of North Carolina at Chapel Hill." Accessed April 25, 2022. https://docs.google.com/document/d/e/2PACX-1vQhpLlf5nWdUzTeDCAB9wtS-cBd-Bk0V4uPll-mEH5zwH6vszmXigDIUV3MmMAcwwkPzPEWRxzihZrn9/pub.

———. "Open Access Policy." Accessed April 25, 2022. https://policies.unc.edu/TDClient/2833/Portal/KB/ArticleDet?ID=132180.