Subject Coverage, Gender and Race in Carolina Digital Repository Content

Rebekah Kati
Institutional Repository Librarian
University of North Carolina at Chapel Hill

Introduction

The Carolina Digital Repository (CDR) is the institutional repository for the University of North Carolina at Chapel Hill (UNC-CH). The CDR is built on Samvera Hyrax and accepts all types of scholarly material from the UNC-CH community, including articles, book chapters, posters, presentations, video and research data.

In support of UNC-CH's Open Access Policy, the UNC Libraries Open Access Implementation Team was tasked in 2017 with increasing the amount of faculty scholarship in the CDR.[1] The Team identified three strategies, which we collectively named "Content Liberation":

1. Author citations/1foldr: Originally, CDR staff planned to conduct affiliation searches in the UNC Libraries subscription databases. After the UNC 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.

A more detailed explanation of the three projects can be found in the Content Liberation Project Summary. Details regarding the 1foldr report can be found in the August 2020 update report.

Although the Team has made great progress in increasing faculty content in the CDR, I have wondered if the content that we have added is representative of UNC-CH as an institution. For example, the 1foldr report enabled the deposit of over 28,000 articles into the CDR, but I observed that much of the content came from PubMed Central, which provides access to works in the biomedical and life sciences fields.[2] While UNC-CH has a strong program in biomedical and life sciences, they are only one aspect of the research in which the university is engaged. By loading 28,000 articles in the biomedical and life sciences fields into the CDR, we might have skewed the subject focus of the repository.

The CDR is the institutional repository for UNC-Chapel Hill and is charged with storing, preserving, and providing access to university scholarship, therefore the Open Access Implementation Team believes 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, I conducted an analysis of the subject areas, gender and racial makeup of authors included in the three approaches to content identification. The research questions in this analysis are:

1. Does CDR content fully represent the scholarly output of the university?
2. Do CDR's initiatives and content sources ameliorate or replicate existing inequities in the academy?

---

[1] 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).

[2] https://www.ncbi.nlm.nih.gov/pmc/

3. How can CDR bring attention to the work of marginalized groups?

To be clear, this analysis should not be regarded as comprehensive. I note limitations in sections below where appropriate. The goal of this project is to reveal general trends and inequities in CDR coverage that can be addressed in future initiatives. I welcome comment and further conversation.

Subject Area Coverage

UNC-Chapel Hill is the twelfth largest research university in the United States.[3] Research occurs across the university's schools and colleges including medicine, public health, arts and sciences, education, pharmacy and more. For this portion of the assessment, I wanted to determine if the Content Liberation projects contained work from all the colleges and schools in the university.

I used a consistent methodology for each of the three Content Liberation projects. I assigned each author a subject classification based on their College or School affiliation within the University, based on their primary departmental affiliation listed in the article. For clarity, the College of Arts and Sciences was further subdivided into subject areas according to their website.[4] "Unknown" refers to authors who only had a university affiliation and therefore could not be classified into a subject area.

1foldr

The CDR Team ingested over 28,000 articles from the 1foldr report into the CDR in 2020. For this portion of the analysis, I removed duplicate authors and authors who did not list a UNC-CH affiliation. Since this dataset was used for both a subject and gender analysis, I also removed authors who listed initials, rather than full first names. This process generated a dataset of 11,102 unique UNC-CH affiliated authors.

[3] UNC Office of Sponsored Research (2021). About. https://research.unc.edu/about/

[4] See "Departments, Curricula, Centers and Institutes" at https://college.unc.edu/news-and-features/departments-curricula-centers-institutes/

College or School	Number of Researchers
Arts and Sciences – Fine Arts and Humanities	3
Arts and Sciences – Natural Sciences and Mathematics	1,325
Arts and Sciences – Social Sciences and Global Programs	91
Business	18
Dentistry	191
Education	7
Government	0
Information and Library Science	42
Interdisciplinary	174
Journalism	9
Law	3
Medicine	4,508
Nursing	47
Pharmacy	409
Public Health	1,143
Social Work	22
Unknown	3,110

CV Review

As of January 2021, library workers and students have reviewed 426 faculty CVs. I asked library workers and students to choose departments for CV review based on their own interests.

College or School	Number of Researchers
Arts and Sciences – Fine Arts and Humanities	96
Arts and Sciences – Natural Sciences and Mathematics	13
Arts and Sciences – Social Sciences and Global Programs	138
Business	0
Dentistry	0
Education	1
Government	0
Interdisciplinary	12
Information and Library Science	8
Journalism	16
Law	0
Medicine	67
Nursing	1
Pharmacy	3
Public Health	41
Social Work	30
Unknown	0

Highly Cited Authors

Clarivate's Highly Cited Researchers list identified 71 unique UNC-CH researchers.

College or School	Number of Researchers
Arts and Sciences – Fine Arts and Humanities	0
Arts and Sciences – Natural Sciences and Mathematics	14
Arts and Sciences – Social Sciences and Global Programs	3
Business	3
Dentistry	0
Education	0
Government	0
Interdisciplinary	0
Information and Library Science	0
Journalism	1
Law	0
Medicine	31
Nursing	1
Pharmacy	4
Public Health	14
Social Work	0
Unknown	0

Work of Black Faculty, Faculty of Color and Indigenous Faculty in the CDR

In 2020, 73.9% of tenure and tenure-track faculty at UNC-CH identified as white. Only 11.8% 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% identified as Black, 5.3% identified as Hispanic, 0.9% identified as multiracial, 0.5% identified as American Indian or Alaskan Native and only 0.1% identified as Native Hawaiian or Pacific Islander.[5] For this part of the DEI assessment, I wanted to determine whether 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 published the Black Faculty, Faculty of Color and Indigenous Faculty Roadmap for Racial Equity at the University of North Carolina at Chapel Hill. This roadmap was signed by 815 supporters, including 144 faculty members who self-identified as BIPOC. I compiled a list of BIPOC faculty 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 affiliation, which I categorized into School and Colleges.

[5] 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/

College or School	Number of BIPOC Researchers from List
Arts and Sciences – Fine Arts and Humanities	54
Arts and Sciences – Natural Sciences and Mathematics	13
Arts and Sciences – Social Sciences and Global Programs	29
Business	1
Dentistry	0
Education	4
Government	1
Interdisciplinary	0
Information and Library Science	3
Journalism	6
Law	1
Medicine	21
Nursing	1
Pharmacy	1
Public Health	12
Social Work	4
Unknown	3

Of course, this method of self-identification does not capture all members of BIPOC communities at UNC-CH, just those members who signed the petition or publicly identified themselves. UNC-CH lists 432 tenured and tenure-track faculty employed by UNC-CH and we have deposited research authored by only 35% of those faculty in the CDR.[6] It is very likely that this analysis under-represents contributions by BIPOC faculty to the CDR.

1foldr

I compared members of the BIPOC faculty list with CDR contributions. I determined that 283 articles deposited in the CDR had been written by faculty on the BIPOC faculty list. Of these articles, 198 or 69% were part of the 1foldr upload project.

CV Review

I compared the BIPOC faculty list to the list of authors whose CVs had been reviewed as of January 2021. Of the 154 authors on the BIPOC faculty list, 36 authors or 23% have had their CVs reviewed. It should be noted that due to staffing shortages, 35 of these authors have not yet had their research deposited in the CDR.

[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/

Highly Cited Authors

I compared the BIPOC faculty list to the list of Clarivate's Highly Cited Researchers from UNC-CH. None of the 71 UNC-CH Highly Cited Researchers appear on the BIPOC faculty list.

Gender in CDR

In 2020, 58.9% of tenure and tenure-track faculty at UNC-CH identified as male. Only 41.1% of tenure and tenure-track faculty identified as female.[7] For this part of the assessment, I wanted to determine whether scholarship produced by women had been deposited into the CDR as part of the Content Liberation projects.

1foldr

Unfortunately, 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. Therefore, I decided to use a gender prediction service for the 1foldr portion of the gender 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 to 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.

I chose genderize.io as it was free for up to 1000 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%.[8] 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, I queried two other gender prediction services, GenderAPI and NamSor. If GenderAPI and NamSor did not agree on a gender for the name, I chose the gender which had the highest probability score.

Of course, gender prediction services are only as good as their datasets. I discovered that genderize.io had trouble identifying non-Western names, likely because their data sources are skewed towards Western Europe.[9] Additionally, I observed that genderize.io did not identify names containing spaces or accents and had low probability scores for gender neutral names. Finally, gender prediction services do not account for genders other than male or female and do not reflect an individual's gender identity.

---

[7] For gender demographics at UNC-CH see "Permanent Full-Time Faculty and Post-Doctoral Fellows by Gender and Tenure Status, Fall 2010-2020" at https://oira.unc.edu/reports/

[8] Santamaría L, Mihaljević H. 2018. Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science* 4:e156 https://doi.org/10.7717/peerj-cs.156

[9] Genderize.io publishes a breakdown of data entries by country at: https://genderize.io/our-data. Four out of the top five data sources as of April 2021 originate from countries in Western Europe.

Gender Prediction Service Prediction	Number of Researchers
Male	6,031
Female	5,071
Unable to ascertain	0

CV Review

I compiled a list of CV reviews completed as of January 2021. This list was of a manageable size and consisted of current faculty. Therefore, I decided to conduct web searches for faculty biographical statements and departmental news stories to determine the faculty member's preferred pronouns. This approach enables the analysis to reflect the individual's preferred gender identity in a professional setting.

Preferred Pronouns	Number of Researchers
He/him	214
She/her	196
Unable to ascertain	16

Highly Cited Researchers

I used names from the list of Highly Cited Researchers at UNC-CH to conduct web searches for faculty biographical statements and departmental news stories to determine the faculty member's preferred pronouns. As with the CV review project, it should be noted that these pronouns reflect the individual's gender identity in a professional setting.

Preferred Pronouns	Number of Researchers
He/him	56
She/her	13
Unable to ascertain	2

Discussion

This analysis revealed several trends in scholarship deposited to the CDR via the Content Liberation projects. I will discuss these trends below in consideration of the research questions.

Research Question 1: Does CDR content fully represent the scholarly output of the university?

Subject areas

The subject content of the CDR overwhelmingly represents the science and medicine areas of the university. Researchers in the sciences comprise 7,214 of the 11,102 authors in the 1foldr report. Researchers in the humanities, social sciences, business and law comprise only 195 of the 11,102 authors in the 1foldr report. This discrepancy could be due to the nature of the 1foldr report, which provides access to openly available versions of papers.[10] Given that the open access movement has been slower to take off in the humanities, it is not surprising that most of the articles in the 1foldr report would be concentrated in the sciences.[11]

The Highly Cited Researchers project also contributed mostly science and medicine content to the CDR. 64 out of 71 authors wrote in the sciences or medicine. Only seven authors wrote in social sciences, business or journalism. Clarivate bases their methodology on the publication of multiple papers in the last decade.[12] This approach would seem to favor researchers in the sciences where the primary mode of scholarship is scholarly articles. Fields which favor other forms of scholarship such as books, presentations, and white papers would not be well represented in Clarivate's list.

In contrast, the CV review project skewed more towards the humanities and social sciences. 125 of the researchers included in the CV review project published in the sciences or medicine. 289 researchers published in the humanities, social sciences, education, journalism or social work. Many of these CVs were reviewed as part of a work from home project during COVID-19, in which workers were encouraged to choose a department to review based on their personal interests. Happily, this approach generated content from under-represented subject areas which can be utilized in future projects.

Race

Authors of scholarship included in the Content Liberation projects primarily identify as white, which tracks with the demographics of UNC-CH tenure and tenure-track faculty. Out of the 28,000 articles which were loaded from the 1foldr report into CDR, only 198 were authored by a faculty member on the BIPOC faculty list. None of the 71 Highly Cited Researchers appeared 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% of tenured doctoral scientists and engineers.[13]

Results for the CV review project were more encouraging. Out of the 154 authors on the BIPOC faculty list, 36 had their CVs reviewed, representing 23% of the overall BIPOC list. The CV Review results may be

[10] 1Foldr (2021). https://www.1science.com/1foldr/
[11] See Suber (2017). Why Is Open Access Moving So Slowly in the Humanities? https://blog.apaonline.org/2017/06/08/open-access-in-the-humanities-part-2/
[12] https://recognition.webofscience.com/awards/highly-cited/2020/methodology/
[13] https://ncses.nsf.gov/pubs/nsf21321/report/academic-careers#tenure-and-academic-positions

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 lists trended towards the sciences.

Gender

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. 54% of the names in the 1foldr report were male-predicted, which is slightly less than the 58.9% of tenure and tenure-track male faculty at UNC-CH. The Highly Cited Researchers project had the starkest disparities, as 56 researchers (78.8%) used male pronouns and 13 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.[14] Given that the 1folder 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 18 more researchers used male pronouns than researchers using female pronouns.

Research Question 2: Do CDR's initiatives and content sources ameliorate or replicate existing inequities in the academy?

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 high priority for preservation and hoped that contacting prominent researchers might lead to an increased awareness of the CDR among faculty. While we did preserve high impact research, contacting prominent faculty did not lead to an increase in self-deposit. We 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 our faculty list. This tracks with findings from the literature, which determined that articles with female first authors were cited less than male first authors.[15] It is expected that fewer female authors would be included on Clarivate's Highly Cited Researchers list.

Despite these findings, it is useful to note that the Highly Cited Researchers project was successful in driving traffic to the CDR and opening research done by the few women on the list. In 2020, 43 out of the 100 most downloaded articles from the CDR were part of the Highly Cited Researchers project. Ten of these articles were written by women – including the top three most downloaded articles. It seems

---

[14] Larivière, V., Ni, C., Gingras, Y., Cronin, B. and Sugimoto, C.R. (2013). Bibliometrics: Global gender disparities in science. Nature. 504:7479. 211-213. https://dx.doi.org/10.1038/504211a

[15] Lariviere, V., Ni, C., Gingras, Y., Cronin, B. and Sugimoto, C.R. (2013). *Bibliometrics: Global gender disparities in science.* Nature. 504(7479). https://www.nature.com/news/bibliometrics-global-gender-disparities-in-science-1.14321

that scholarship written by highly cited women is useful for the research community and this can inform our priorities going forward.

After the Highly Cited Researchers project was completed, the Open Access Implementation Team turned to the 1foldr report. The goal of the 1foldr report was to quickly change the focus of the CDR from a repository focused on student papers to a faculty scholarship repository. The 1foldr report did achieve that goal, but the DEI analysis shows that it overwhelmingly contained work done by white men in the sciences. Work done by white men in the sciences is now the focus of the CDR and while we were able to achieve the overall goal of the project, the implications deserve further reflection.

Out of all the Content Liberation projects, the CV review project was the most equitable, although it was not perfect. The CV review project opened more content by women, BIPOC researchers and in humanities and social sciences than the other two projects. However, they still represent a small amount of the research being done at UNC-CH.

Research Question 3: How can CDR bring attention to the work of marginalized groups?

The CV Review project is the most flexible and targeted of the three Content Liberation projects and can be adapted to help ameliorate the inequities in the CDR. In fact, CV Review has already been used in a targeted manner. During the COVID-19 pandemic, I created a sub-project which collected papers from coronavirus researchers at UNC-CH and deposited them in the CDR. Similarly, the 36 completed CV reviews from BIPOC researchers can be checked, processed and deposited in the CDR. The rest of the BIPOC faculty list could also be reviewed and deposited. Furthermore, I could write blog posts and engage the Libraries' social media department to bring attention to the newly deposited work. The same process could be utilized for women's CVs and CVs of researchers in the social sciences and humanities. This plan would not fully represent the work of under-represented communities and subject areas, but it is a start.

Conclusions and Future Work

As shown above, the Content Liberation projects have succeeded in adding more scholarly articles to the CDR. However, the articles are mostly in science and medicine areas and are written by white men. While these subject areas and demographics represent the makeup of UNC-CH, it is necessary to reflect whether it should be the focus of the CDR. Indeed, keeping the focus of the CDR as-is would not be aligned with the UNC Libraries Reckoning Initiative Framework, which charges us to "... implement practices and policies that sustain equity, opportunity, and inclusive excellence."[16] Our next steps are to consider the approach that we will take to accomplish this charge.

The following initiatives will start the process to refocus the CDR:

- Research and reflect on ways to identify gender and race of CDR authors in an accurate and ethical manner.
- Continue review and deposit of CVs from the BIPOC faculty list.

[16] See UNC Libraries Reckoning Initiative Framework: https://library.unc.edu/reckoning/framework/

- Review CVs of researchers engaged in research on racial and gender equity.
- Revisit Highly Cited Researchers who use female pronouns and deposit all eligible scholarship, rather than only the highest cited scholarship.

In the discussion of Research Question 3, I noted that CV Review could be used in a targeted manner to bring attention to the work of marginalized groups. It should be noted that CV Review is a time-consuming and highly manual process. To be clear, the CDR will not see the same gains in deposit numbers as it did during the Highly Cited Researchers and 1foldr projects. This is not necessarily negative, since the CDR has experienced rapid growth over the past three years and a slower, more intentional approach may be desired.

Nevertheless, sustainability of the Content Liberation projects should be considered. Currently, there is only one full time librarian dedicated to the content initiatives of the CDR. Other library workers help with metadata, software development, user acceptance testing and accessibility initiatives, but they are not dedicated to content. To sustain growth, we will need more workers to help on content projects. In the fall, I will begin work with the Research and Instructional Services department to review CVs of BIPOC faculty in the humanities and social sciences. This initiative is a first step to sustainability of CV review going forward and I will evaluate the results in the coming months.

Subject area focus also may impact the growth of the CDR going forward. The CV review project consists of humanities and social sciences content. Humanities and social sciences researchers tend to not publish as many scholarly articles as science and medicine researchers, as their primary forms of scholarly communication rely on books, book chapters, conference presentations and others. Since UNC-CH's Open Access Policy applies only to scholarly articles, it is likely that efforts to collect large amounts of scholarship in the humanities and social sciences will be constrained. Going forward, the Open Access Implementation Team believes that opening fewer articles in key strategic areas is more beneficial to equity in CDR content than providing access to already open materials. Easier is not necessarily better.

We hope that 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. We will continue to assess our progress and publicly publish updates on a yearly basis, as we have done with the Content Liberation projects and the CDR platform updates.