Using Web Archives to Model Academic Migration and Identify Brain Drain

Mat Kelly, Deanna Zarrillo, and Erjia Yan
{mkelly, dz364, ey86}@drexel.edu
Drexel University, College of Computing & Informatics (CCI)
Philadelphia, PA

@machawk1

IIPC Web Archiving Conference (WAC) Online Day
May 3, 2023

The Project

- Identify academic mobility using web archive
- Historically Black College & Universities more prone to “brain drain”
- HBCU faculty often leave for PWIs
- HBCUs make substantial contributions to the preparation of Black professionals
  - 80% Black federal judges
  - 85% Black doctors
  - 50% Black engineers

Sources

- Web archives!
- Combine with Web of Science Data
- Combine with AARC data
  - via UWisconin-Madison

U.S. Focus → U.S. Web Archives?

- IA is popular as a sole source but we first looked to be more comprehensive
- Memento aggregation
- MemGator
  - Custom set of archives
  - Setup as own web service or host locally
- Ultimately, overwhelming majority of captures were from IA

High-Level Approach

- 35 HBCUs' homepages on live web H_i
- Variable # of departments at each HBCU on live web $H_i D_j$
- Work backward in time
- A contemporary basis was flawed
  - Departments are created, dissolved, merged, moved, etc. in time
- We still wanted to associate the same dept. at different URLs

Quasi-Canonicalization

- Typically used to associate URLs with subtle different that are the same site
  - www.example.com and example.com
  - https://example.com and http://example.com
  - example.com and example.com/index.php
- URI-R of departments' sites in time may be more dissimilar
  - e.g., http://www.howard.edu/CEACS/ and https://cea.howard.edu
- Examples
  - http://www.sun.com/java/ → https://www.oracle.com/java/
  - http://thefacebook.com → https://facebook.com

M. Kelly, D. Zarrillo, C. Jackson, and E. Yan, "First steps in Identifying Academic Migration using Memento and Quasi-Canonicalization," Presented at the ACM/IEEE JCDL 2022 Workshop on Web Archiving and Digital Libraries (WADL 2022), Cologne, Germany, June 20–24, 2022. (**PDF**)

Procedure Execution

- **35** HBCUs × **Y**-departments × **Z** data points
- Data rapidly grew
- Some departments' captures were sparse, the faculty pages even moreso
- Some faculty listings paginated

Procedure Execution…and the Problems that Arose

- Redirects caused temporal drift
- Temporal drift also naturally occurred from (home→dept→faculty page)
- Many broken links (some real 404s, some soft, some 5XX)
- Automated approach missed data if not widely adaptive

The (A) Solution: Manual Collection

- Unleashed the (Funded) students on the web archive
- Annual granularity
- Used their web browsers
- Redirect issue remained, Memento wouldn't solve it
- Scraping names/titles using automation was unreliable
  - We wanted to emphasize recall, not miss instances of a professor due to bad RegEx

Lessons So Far

- Past academic Web particularly prone to page movement in time
- Canonicalization/Association of drastically different URLs is critical
- Manual collection introduces further errors and need-to-clean
- Have identified some solid data points that highlight migration

Spin-Offs

- Ethical considerations of collecting individuals' past info
- Need for anonymization
- HBCU faculty recruiting from HBCU students?
- Professor name changes/time?

Conclusions

- Automated approaches lead to missing data
- Departments' URLs move in time, making tracking faculty difficult
- "Quasi-canonicalization" is necessary to associate departmental URI-Rs that may (have been) drastically different

E. Yan, **M. Kelly**, D. Zarrillo, J.He, C. Ni, and R. Palmer, "Examining the academic mobility at Historically Black Colleges and Universities in the U.S.," Accepted to be In Proceedings of the 20th International Conference of the International Society for Scientometrics and Informetrics (ISSI), Bloomington, Indiana, **July 2-5, 2023.**

D. Zarrillo, **M. Kelly**, C. Jackson, and E. Yan, "Collecting Diachronic Affiliation Data for Faculty at HBCUs Using Memento," In Proceedings of the **Association for Information Science and Technology**, Vol. 59, pp. 528–532, Pittsburgh, Pennsylvania, October 29–November 1, 2022.

M. Kelly, D. Zarrillo, C. Jackson, and E. Yan, "First steps in Identifying Academic Migration using Memento and Quasi-Canonicalization," Presented at the ACM/IEEE JCDL 2022 Workshop on Web Archiving and Digital Libraries (WADL 2022), Cologne, Germany, June 20–24, 2022.

More information on project:
https://hbcumobility.cci.drexel.edu/

Science of Science:
Discovery, Communication, and Impact
SoS:DCI Award #2122525