First steps in Identifying Academic Migration using Memento and Quasi-Canonicalization

Mat Kelly
Drexel University
Philadelphia, PA, USA
mkelly@drexel.edu

Deanna Zarrillo
Drexel University
Philadelphia, PA, USA
dz364@drexel.edu

Christopher Jackson
University of Tennessee
Knoxville, TN, USA
cjacks75@vols.utk.edu

Erjia Yan
Drexel University
Philadelphia, PA, USA
ey86@drexel.edu

ABSTRACT

Academic research faculty change their institutional association over their careers. This association may be documented on the web through the web sites of the faculty member's institutional department. The effects of disassociation are detrimental to historically black colleges and universities (HBCUs), but the degree of "brain drain" is difficult to evaluate without first obtaining evidence of these faculty member's movement over time. This work describes a preliminary effort to utilize web archives to identify faculty that resided at HBCUs in the past in an attempt to further examine the effects of brain drain. We utilize a Memento aggregator along with a manual data selection procedure to first identify the target URI-Ms. We then perform a recursive procedure to supplement TimeMaps with a more comprehensive picture of HBCU department web pages over time. This association and quasi-canonicalization procedure is a first step in identifying the effects of brain draft by utilizing the past Web.

KEYWORDS

web archives, memento, hbcu, canonicalization

1 MOTIVATION

College and universities have historically had a web presence. Their sites are typically structured by college, department, or other sub-units to ensure relevant information about the respective unit is effectively represented. The sub-units' sites often list those employed by the unit, for example, faculty, staff, and research assistants. The faculty employed by the unit changes in time. Thus, identifying how faculty have transitioned between sub-units or among higher education units is possible by consulting historical representations of the sub-units' sites.

The degree to which up-to-date and representative information is present on academic web sites may be relative to a variety of factors including the units' size, staffing, resources, etc. This, too, has changed in time as colleges and universities have recognized the importance of having an online presence.

Our focus is on faculty representation online at historically black colleges and universities (HBCUs). We leveraged the units' live web presence as a basis to identify the degree to which the units' online presence have been preserved. We anticipate being able to supplement the captures (e.g., URI-Ms for Howard University's Pharmacology Department[1]) with previously associated yet currently disassociated archival identifiers (e.g., URI-Ms for the department's URI-R in the past[2]). Our ultimate goal is to more thoroughly identify faculty at these HBCUs and how they have moved between the aforementioned units and sub-units. This work serves as the initial data collection phase toward this goal of identifying "brain drain" from HBCUs.

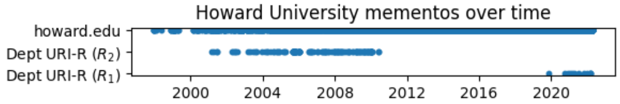

Figure 1: While an HBCU's homepage spans a time range, a department at the HBCU that exists on the live web (R_1) might have existed elsewhere in the past (R_2).

2 METHODOLOGY

Our initial objective was to build a web archive collection for analysis. Our base data set consisted of 35 URI-Rs identifying the homepages of the HBCUs. For archival queries, we utilized a local instance of the open source MemGator Memento aggregator[3] software with the default archival configuration.

We initially requested 35 TimeMaps for these URI-Rs (H) from the aggregator. Next, we manually collected URI-Rs for each subunit at each HBCU based on the live web ($\{D_f\}$). We again queried the aggregator for TimeMaps for each $D_i \in D_f$, which varied in quantity among the HBCUs. For example, Howard listed 51 sub-units on the live Web with varying scopes (e.g., colleges, departments, majors). Using the 35 TimeMaps in H, we downloaded the resource response of the oldest memento (M_0). We manually identified the historical departmental URI-Rs in each oldest memento for each $H_h M_0$ where $H_h \in H$. Following this, we identified the differences between the set of departments per each memento ($D_0 \in H_h M_0$) and the list of departments represented on the live web, where D_f need not be equivalent in magnitude to $H_h M_f$ for TimeMaps like R_1.

In upcoming work, we will programmatically map historical department URI-Rs to contemporary URI-Rs, noting patterns of deviation. Those that have concrete associations can have their URI-Ms for each URI-R variation in the same TimeMap (Figure 1). This association of a department over time is akin to canonicalization and will surface additional nuances in the relative structures encompassed within a TimeMap. We will then execute a recursive procedure to acquire the list of faculty at each department over time using approaches that leverage named entity recognition.

3 ACKNOWLEDGEMENTS

This work is supported by NSF SoS:DCI grant #2122525.

[1] Currently at R_1: https://medicine.howard.edu/graduate-programs/pharmacology

[2] Previously residing at R_2: http://med.howard.edu/pharmacology

[3] https://github.com/oduwsdl/memgator