Journal of Contemporary Archival Studies

Volume 2 Article 5

2015

# Quidditch, Zombies and the Cheese Club: A Case Study in Archiving Web Presence of Student Groups at New York University

Aleksandr Gelfand
*Unaffiliated*, ag3566@nyu.edu

# Quidditch, Zombies and the Cheese Club: A Case Study in Archiving Web Presence of Student Groups at New York University

**Cover Page Footnote**
Dedicated to Nancy Cricco (1953-2015) - Colleague, Mentor, and Friend.

# Quidditch, Zombies, and the Cheese Club: A Case Study in Archiving Web Presence of Student Groups at New York University

Colleges and universities have widely acknowledged participation in student groups, organizations whose activities foster socialization, as an essential element of the learning process. Even the smallest of colleges may boast their own unique sets of student groups intended for moral, academic, physical, or social development of their student bodies. Writing in *Varsity Letters: Documenting Modern Colleges and Universities*, a guide for documenting institutions of higher learning, Helen Willa Samuel has noted the difficulty in collecting information on such groups: many of the activities students engage in are not recorded in any form, and existing documentation is frequently created by university and college administrative offices.[1] Samuel recommends an early and active engagement by archivists with student groups in order to ensure the existence and survival of student-created records. For most college and university archivists, however, a conscious and continuous documentary effort, sometimes involving hundreds of clubs, is a difficult undertaking; the typical repository, if it attempts to collect student-life materials at all, engages in the process on an ad-hoc basis. At the same time, the migration of student groups onto the online environment in the mid- to late 1990s has seemingly added to that difficulty. However, the recent advent of easy-to-use web archiving tools, specifically Archive-It, has created new possibilities for documenting student life.[2] Although an initial expenditure of time and funding is required, the investment is worth it and will ensure that the institution preserves a unique portion of its history for future generations of researchers and alumni alike.

## Background

Founded in 1831, New York University (NYU) is the largest private university in the United States with over forty thousand students attending eighteen different schools and colleges at five major centers in Manhattan, as well as at sites located in Africa, Asia, Europe, and South America.[3] NYU is both a residential college, with thousands of students living in dormitories near the Manhattan and Brooklyn

[1] Helen Willa Samuels, *Varsity Letters: Documenting Modern Colleges and Universities* (Chicago: Society of American Archivists, and Metuchen, N.J.: Scarecrow Press, 1992), 76–78.
[2] Archive-It is a web archiving service made available by the Internet Archive. For more information, see https://www.archive-it.org/learn-more/.
[3] http://www.nyu.edu/about.html.

campuses, and a commuter school, with students living throughout all five of New York City's boroughs, as well as in some of the neighboring states.

Established in 1977, the university archives is the smallest of the three special collection repositories housed at NYU.[4] For a number of years the university archivist, Nancy Cricco, had been actively advocating for the establishment of a digital records program and the hiring of a digital archivist to deal with the issue of born-digital records that have increasingly become prevalent on campus. She decided to establish a web archiving program during the summer of 2013 as an important first step in tackling the acquisition of digital records. NYU was already a subscriber to both the California Digital Library's Web Archiving Service (WAS) and Internet Archive's Archive-It, two of the most popular web archiving services used by university archives.[5] Archive-It was selected because the platform has the ability to archive a greater variety of websites than WAS. Due to the dynamic nature of student groups and the fact that the archive already possessed a large analog collection of student group materials, the university archivist decided that this would be one of the first web-based collections added to the archive. The new digital collection was given the same name as its analog component, Record Group (RG) 19: Student Organizations and Publications. Rather than encompassing all schools and colleges of the university, some of which have more autonomy with their own archives, the scope of the collection was initially limited to fifteen entities.

**Table 1. Schools and Colleges Included in the Collecting Scope**

| | |
|---|---|
| College of Arts and Sciences | School of Continuing and Professional Studies |
| College of Dentistry | School of Law |
| College of Nursing | School of Medicine |
| Gallatin School of Individual Study | Silver School of Social Work |
| Graduate School of Arts and Science | Steinhardt School of Culture, Education, and Human Development |
| Leonard N. Stern School of Business | Tisch School of the Arts |
| Liberal Studies | Tandon School of Engineering (formerly NYU-Poly) |

[4] The others are Fales Library and Special Collections and the Tamiment Library and Robert F. Wagner Labor Archives.

[5] In early 2015, California Digital Library announced a partnership with the Internet Archive, transferring all of its WAS core infrastructural activities to Archive-It.

| | |
|---|---|
| Robert F. Wagner Graduate School of Public Service | |

## Getting Started

A search of contemporary archival literature for examples of other universities engaged in similar endeavors yielded meager results.[6] Browsing publicly available web collections created by other university archives did not produce good examples to imitate. Other institutions concentrated primarily on departmental and main institutional websites rather than on individual student group websites.[7] Of the repositories that engaged in collecting student groups, most did the work on a small scale, describing each website with a limited amount of metadata.

Before creating the web collection, I attempted to compile a thorough list of student organizations and their websites prior to their attempted capture. An initial list of approximately three hundred student clubs was procured from the website of the Center for Student Activities, Leadership, and Service (CSALS).[8] The center serves as a contact point between student groups and the university administration. However, upon further inspection, I discovered that a number of different lists were available through the center's site, some outdated and with varying degrees of duplication. The most complete and up-to-date inventory yielded five hundred different student groups classified by type.

### Table 2. Classification of NYU Clubs

| | |
|---|---|
| Academic | Literary/Publication |
| Art/Performance/Media | Political/Public Policy/Activist/Advocacy |

[6] One of the few exceptions was Christopher Prom and Ellen Swain, "From the College Democrats to the Falling Illini," *American Archivist* 70 (2007): 344–363. The article analyzes the web presence of student groups and the subsequent efforts of their capture at the University of Illinois at Urbana-Champaign in 2003–2004. The article provides useful advice on initial steps in starting a student organization web collection; however, due to the time span that has elapsed since the article's publication, I feel it has limited applicability.

[7] Browsing was done via the Archive-It portal at https://www.archive-it.org/explore?show=Organizations.

[8] http://www.nyu.edu/about/leadership-university-administration/office-of-the-president/office-of-the-provost/university-life/office-of-studentaffairs/office-of-student-activities.html.

| | |
|---|---|
| Community Service | Professional |
| Computer and Technology | Recreational |
| Cultural | Religious/Spiritual |
| Debate and Speech | Social |
| Fraternity | Sorority |
| Lesbian, Gay, Bisexual, Transgender, Questioning | Student Council/Government |

The majority of the clubs listed were available to students in all schools and colleges, classified as "all-square."[9] However, each of the fifteen colleges and schools included in the collection scope possessed their own student organizations, which were supported in part by their respective institutions and not listed on the CSALS website. In order to add them to the list, I was required to search the websites of each of the fifteen academic units. Furthermore, although many of the clubs listed on CSALS and other institutional websites included links to the club sites, the vast majority were found to be either broken or connected to long-abandoned sites.

Attempting to understand how student groups were being regulated, and to gain access to the latest listings, the university archivist and I reached out to the Student Center directly and were able to obtain a meeting. While the director and associate director of the center were new to the university, they provided us with an old master list of all club names and websites, which we used to supplement our existing list. In addition, we managed to discover that in prior years, clubs were managed on a locally built content management system that has since been discontinued and replaced with OrgSync, an online community management system that serves institutions of higher education.[10] The service allows the administration to easily manage and track student groups while creating a one-stop-shop for students looking for information on clubs or a comprehensive calendar of all club events.

At the time we spoke with the Student Center, NYU was in the process of migrating clubs from every school and college onto OrgSync, with over three hundred clubs in the system added at the time of our meeting. We found OrgSync a valuable resource and good first step for any archivist attempting to get an overview of the world of student life. However, the information available through

[9] The term "all-square" comes from the fact that a large number of the college's buildings are clustered around Washington Square Park in the Greenwich Village neighborhood of New York.
[10] http://www.orgsync.com/what_is_orgsync.

the service was limited, generally consisting of administrative, top-level information including a brief description of the club, its mission, advisor name, club officers, as well as a calendar of scheduled events. To document everyday student life from the students' point of view, we would need to capture websites created and managed by the students themselves, rather than by the university administration.

Eventually, the author assembled a list of 850 student organizations culled from Student Center master lists and individual NYU college and school sites. However, only a small segment of these lists contained any information about the clubs' web presence. In order to discover what was available online, it was necessary to perform a Google search on each of the student group names. Over the course of two weeks, I conducted a sporadic search lasting on average ten to twenty seconds per club. When a site was confirmed as belonging to the group in question, usually by reading the "about us" section, it was entered onto a spreadsheet specifically created for the task. Out of the 850 organizations, approximately 700 were found to possess accessible online content. Many of the remaining groups either had inaccessible, member-only sites or nothing was found. A brief consultation of old yearbooks from just one of the fifteen colleges and schools—the College of Arts and Sciences—from the late 1990s to early 2000 uncovered an additional three dozen defunct organizations. A search revealed ten of them still had websites, although these were long abandoned. The total end product of the search was approximately 1,500 websites whose dates of creation ranged from the mid-1990s to the present. In addition, during the search, materials related to student life but falling outside of the scope of the collection were discovered, leading to the creation of four additional collections under the same record group.[11]

The types of sites discovered ranged from basic, one-page informational sites listing club officers and brief mission statements to elaborate, multilevel websites that include slide shows, movies, and forums. The fluid and ephemeral nature of student groups was evident in the number of abandoned sites and the short length of time between site migrations. Clubs with five or six abandoned websites were not uncommon. The ease of creating a website from scratch has made the process of migration more rapid and pronounced. For example, the first discovered site of

[11] The four collections are (1) Student Blogs: blogs that are written by students or describe aspects of student life at NYU; (2) Student Organizations News: online newspaper articles or blog entries that are about student life at NYU; (3) Student Organizations Event Ephemera: announcements and advertisements of events that are to be held by student organizations; and (4) Online Communities: popular communities that have large student participation where members rarely or never meet face to face and whose administrator typically remains anonymous.

Kesher, a Reformed Judaism club, was created sometime between 2000 and 2003 and completely abandoned by 2004; another site was abandoned in 2008, and the one after that was abandoned in 2010. A fourth one also appears to be abandoned, although it was not possible to determine the dates of when the site was created or last used. A Twitter account was created in 2011 and abandoned later the same year. A Facebook page was created in 2012 and was the only active web presence for the club at the time.

This path from regular, static websites to mutable social media platforms like blogs, Facebook, Twitter, and YouTube proved typical for most of the student organizations.[12] These types of sites may be updated with rapidity and greater ease, leading many groups to abandon static websites. However, this is not always the case. Large and active student organizations, fraternities and sororities chief among them, were found to possess multiple active sites. Frequently, for these groups, a professionally designed site serves as the immediate face of the group and features its history, photo albums, and recruitment information. These sites are often updated irregularly, usually not more than once or twice a year. Orbiting the main site are a number of supplementary sites, including blogs, Facebook, Twitter, YouTube, Pinterest, Instagram, and other easily managed platforms. These are updated constantly and serve as a conduit for information among its members. For example, the sorority Alpha Sigma Tau was found to possess two blogs (abandoned), one professional site (active), one Tumblr blog (active), one Facebook page (active), and one Twitter account (active). Cultural, political, and performance clubs were also found to be very active; for example, the College Republicans, in addition to having a static main website, also have Facebook, Twitter, YouTube and OrgSync pages.

During this initial collecting phase of the project, no appraisal criterion was applied in the acquisition of student websites.[13] All of the approximately 1,500 websites initially discovered fell within the collecting scope, and even sites that had very little information available were captured to display the plurality and at times uniqueness or rarity of the platform used: for instance, MySpace, which was once ubiquitous but is infrequently used today.

## Metadata

[12] Although blogs are considered a part of social media, increased customization has made many of them indistinguishable from regular websites.

[13] Appraisal was applied to the four related collections created (see note 10). For instance, only popular Facebook pages were collected in the Online Communities as determined by the number of likes generated by the page.

Once the size of the new collection was apparent, the importance of good descriptive metadata became a must. The resulting hybrid metadata schema was developed in order to better convey the complexities of the collection (tables 3 and 4).[14] The schema was composed of Dublin Core, which is native to Archive-It; Describing Archives Content Standard (DACS), a content schema used by archivists; and the custom fields that we felt were needed. For instance, when we knew that a website was no longer being updated or if it disappeared entirely, a "last updated" field was created that indicated the date range of its last update. Collection-level metadata was more closely aligned with analog collection description with the hope of possible interoperability and reuse at some point in the future. The seed-level metadata was felt to be of a more malleable, less reusable nature and liable to future revisions.

**Table 3. Collection-Level Metadata: *Dublin Core **DACS ***Custom Fields**

| | |
|---|---|
| Title* | Contact Information** |
| Description* | Condition Governing Access and Use** |
| Subject* | Condition Governing Reproduction and Use** |
| Processed By** | Acquisition and Appraisal** |
| Preferred Citation** | Related Materials** |
| Date Range** | Status*** |
| Language of Material** | Technical Access** |

**Table 4. Seed-Level Metadata: *Dublin Core ***Custom Fields**

| | |
|---|---|
| Title* | Crawl Type*** |
| Creator* | Format* |
| Subject* | Language* |
| Description* | Status*** |
| Last Updated*** | Coverage* |
| Type* | |

[14] Some metadata fields are automatically populated by Archive-It and were thus not included. To see how this metadata is being displayed, please consult the collection at https://www.archive-it.org/collections/3771.

## Legal Issues

At a well-attended web archiving panel at the annual Society of American Archivists conference in New Orleans in 2013, a large number of participants expressed reservations regarding archiving content without getting permission from the content creator, while others stated that they were specifically instructed to avoid social media at all costs by their legal departments. Each repository had a different taste for risk when it came to web archiving based on what they were collecting and what their legal departments told them they could do. Thinking of how we were going to handle these issues, the NYU Archives decided right away that approaching each club for permission would not be a workable solution. Although approaching clubs via their advisors to ask permission was considered, in the end the sheer number of student organizations made the suggestion unworkable. Since membership of each group changes at a relatively rapid rate, we would need to constantly repeat the process, hoping that the student leadership of the respective club would get back to us themselves or via their advisor. We decided that publicly available student websites would be collected without asking the permission of their creators with a stipulation that we would remove the content if requested to do so. Social media content, which makes up approximately half of all publicly available student websites, presents an additional problem. Since social media represents the most dynamic and vibrant student content, we decided that we needed to collect this material, even if we had to restrict access to it for a period of time.

## Crawling the Collection

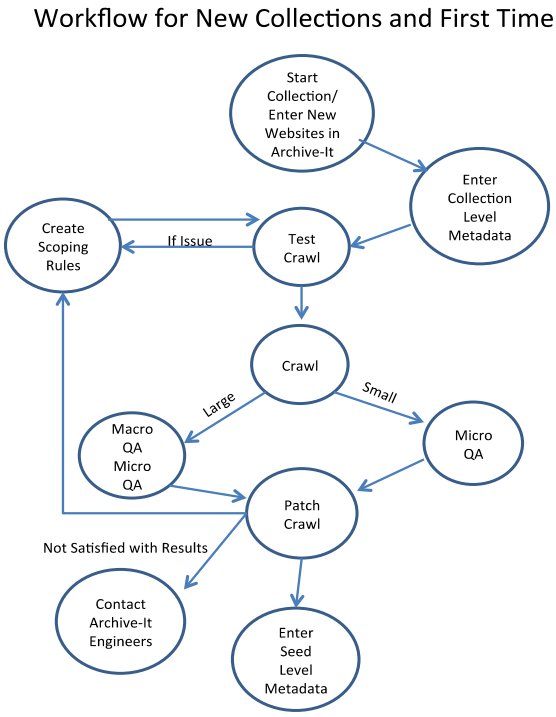
For those approaching web archiving for the first time, there is an undeniable learning curve, though Archive-It provides extensive, constantly updated documentation and video tutorials on the various aspects and intricacies of the platform. If all else fails, the folks at Archive-It are quick to assist via e-mail.[15] The first step in assessing the website to be crawled was to determine the type of crawl that would be best employed to collect the necessary content. Archive-It allows archivists to classify a seed into three different types, which will then impact how that seed will be crawled. The most common crawl is "default," where the crawler will attempt to capture the entire website. The next type is the "News/RSS feed," which will only capture the immediate webpage and any links contained within that page. The last type is "crawl one page only" and will only capture the immediate page (a snapshot) of the indicated page. Once the 1,500

[15] Special thanks to Scott Reed, Lori Donavan, Maria LaCalle, and Sylvie Rollason-Cass for their assistance with the numerous, sometimes inane questions they had to endure from me.

discovered websites were divided into the three categories, they were entered into Archive-It and actual crawling could begin. Due to the large number of websites, this was done in increments, allowing for better overall control of the process and the ability to divide the work between metadata entry and crawling (see workflow chart below).

Right away, we discovered that in the intervening weeks between the compilation of the list and the capture of the sites, more than three-dozen of the sites had disappeared completely before we had a chance to capture them, once more highlighting how unstable born-digital content tends to be. Ingesting the remaining sites proved to be no easy task. The messy, all-encompassing nature of the web frequently made it all but impossible to archive only the wanted websites, avoiding unnecessary materials. A balance needed to be reached whereby as much of the in-scope content was collected while limiting the collection of all unrelated materials. To do this, frequent test crawls were required that were then used to readjust the crawl and the process began all over again. Lack of patience and inadequate testing could easily lead to unintended circumstances and waste of resources. For instance, during the first three months of the project, I fell into two crawler traps resulting in many gigabytes of wasted data.[16] Once the seed was crawled, it had to undergo quality control to make sure that the quality of the capture was the best one possible. The last steps were to determine how frequently, if at all, the site was to be crawled in the future and to enter all pertinent metadata. We settled on three frequencies. Sites that were the most frequently updated were set up on a quarterly crawl (four times a year), less frequent were set to semi-annual crawl, and those that were updated infrequently, or were suspected but not confirmed of being dead, were set at an annual crawl frequency. Two years of site inactivity was considered as the average amount of time after which a site could be classified as dead and be deactivated. Once set, these crawls would need to be checked over the course of the year based on their crawl frequency and readjusted accordingly.

[16] A crawler trap, as defined by Archive-It, is "a set of web pages that creates an infinite number of urls (documents) for the crawler to find. This means that a crawl could keep running and finding 'new' urls that it had not encountered previously for an infinite amount of time. . . . The most common example of a crawler trap is a calendar page on a website. Some calendars will have automatically generated links to 'next month' or 'previous' month on each page, and the crawler may end up crawling calendar pages into the year 3456 or back to 1776, or further!" (https://webarchive.jira.com/wiki/display/ARIH/How+to+Identify+and+Avoid+Crawler+Traps).




## Lessons Learned

Two years after the project started, the collection has increased to almost two thousand websites, leading to a number of conclusions. We now realize that the collection grew far too fast and should have included a more selective appraisal policy at its inception. The web is a bottomless pit of content, and while a large number of websites were uncovered during the search process, this is still a fraction of the materials that are out there. Before starting any similar projects, archivists need to acknowledge that they are not going to be able to capture everything that falls within the collection scope, concentrating instead on a smaller, well-selected, and more manageable number of websites. Although a solid appraisal policy for the university archives is not yet in place, the websites added after the initial ingest have started to undergo much greater scrutiny prior to being accessioned. For instance, if a club already has content in the archive, the information on the website is scrutinized to see how much duplication there is before it is considered for accession. If there are already more than four websites belonging to a single club in the archive, chances are high that a new website will

not be added unless one of the sites is either deaccessioned or deactivated. Archivability needs to be an important aspect of any future appraisal, and we have started to take this into consideration prior to accessioning new websites. For instance the platforms of newly discovered websites are reviewed in depth; if they are found to be difficult to capture, then this becomes one of the considerations against their accession. Some of the initially captured websites have been found to be not very archive friendly and have been inadequately captured.[17] These types of websites are prime candidates for future removal, if we decide to deaccession some of the material from the collection.

The large variety and number of websites that make up this collection has led to the creation of hundreds of scoping rules that give instructions to the crawler when it is reviewing indicated sites. Because the rules can only be applied at the collection level, their sheer number ends up contradicting each other, since what may be right for one type of website may not be right for another. This causes a situation where rules constantly need to be amended or deactivated in order to try and get the best possible captures. By the end of 2015, a fully upgraded version of Archive-It (5.0) will hopefully allow users to create rules applicable to only the specified website, which should solve this problem.

The most labor-intensive portions of the project were metadata entry, learning to use Archive-It, and quality control. The first two facets gradually lessened as the project continued. Previously entered metadata was reused for clubs whose websites are already in the collection; batch metadata entry also made the process substantially easier. The only facet of the project that became more time consuming was quality control. With a constantly increasing number of websites in the collection, more time was needed to ensure that the capture was done properly, eventually taking up the vast majority of the working day. However, since doing quality control is a relatively easy, straightforward task, in the future this task may be relegated to a student worker or a paraprofessional, leaving the archivist to focus on other tasks.

## Conclusion

Web archiving is a constantly moving target, requiring ongoing monitoring to ensure the best possible results. This is even more true for student organizations, whose websites are much more fluid than institutional ones. The archivist

[17] For instance, ISSUU, a very popular publishing platform, was initially difficult to capture. It took the engineers at Archive-It months of testing to succeed. However, all previous captures of ISSUU content remain unusable.

attempting to preserve them needs to be in-tune with events on campus in order to be better able to understand and capture student life as it develops. Over the course of this project, NYU's main student newspaper, *Washington Square News*, and the main university blog, *NYULocal* (http://nyulocal.com), were invaluable sources of information on new clubs and developments in those already existing. Regardless of the challenges and issues that the project created, the results are worthwhile. Hundreds of sites that have since disappeared have been preserved and made available through the repository. As an added benefit, while attempting to archive Facebook pages, the university archives created a Facebook page of its own, and it is actively being used to promote the repository and its mission.[18]

For now, to my knowledge, the collection has been used mostly within the university archives; the primary effort thus far has been to get the web archiving program up and running. The next step is to be able to provide effective and efficient access to the collection. As part of that effort, we have been building a portal that lives on the repository website and explains how to navigate its web collections. The recently released partial update to Archive-It gives subscribers the option of using Google Analytics. Once we connect it to the account, we will be better able to gauge the use and interest of this and other collections.

Experience gained while archiving student groups was invaluable when the project expanded and we began capturing course descriptions, bulletins, and departmental websites. Record creators have also started to include their websites as part of the records given to the university archives. For instance, earlier in 2015, as part of an accession from an institute that was terminating its existence, the archive received a number of boxes of analog records, a few gigabytes of data, and a specific request to archive its website. Discussion of web content is now sometimes included in dialogue with record creators prior to the accession of their records. In at least one case, while discussing analog records, a leader of a student group indicated that their club website was about to be replaced. This led to the capture of the old site prior to its disappearance. These types of hybrid collections are only going to become more frequent in the future, and in order to succeed in their profession, archivists will need to become more involved in web archiving. Capturing web content of student organizations could be one way for university archivists to take up this challenge.

[18] The Facebook site https://www.facebook.com/ArchivesNYU is itself being archived in a subsequent collection.