# TOWARDS PRESERVING WEB-BASED STUDENT PUBLICATIONS AT CONCORDIA UNIVERSITY

**Sarah Lake**
*Concordia University*
*Canada*
*sarah.lake@concordia.ca*

**John Richan**
*Concordia University*
*Canada*
*john.richan@concordia.ca*

**Abstract – Student-run papers, journals, and magazines that were previously published in print form are now almost exclusively hosted on digital publishing platforms. How will this shift impact the longevity of student scholarship and institutional memory? This paper will present an ongoing project that aims to archive web-based student publications at Concordia University. We will discuss the rationale behind the project, our initial objectives and scope, and the challenges that we have encountered so far. We will conclude with a discussion of future opportunities for outreach and other envisioned pathways for collaboration.**

**Keywords – web archiving, student publications, university archives, academic libraries, Archive-It**

**Conference Topics – We're All in this Together**

## I. BACKGROUND

During the summer of 2018, the Concordia University Records Management and Archives (RMA) department started the process of web archiving within the greater context of its newly published Digital Preservation Program. With a departmental mandate directly tied to University records management activities, early efforts using Archive-It were focused on establishing and maintaining automated crawls within the Concordia domain. The objective of this work was two-fold: To preserve important information contained on Concordia websites and to respond to research needs originating from the community searching for information hosted on former University pages no longer accessible.

RMA web archiving has since evolved to respond to special interest topics, for example the Concordia COVID-19 Web Collection. With this evolution has also come the opportunity to collaborate with new partners and advocate for the importance of web archiving within the community.

One collecting gap that has been identified by RMA was archiving *online-only* student publications. These publications are characterized as being exclusively online journals or magazines managed by Concordia students showcasing undergraduate or graduate writing and work. Prior to the shift to online only, RMA collected extensively in this area and houses a large print collection of this type of material for research purposes. However, with the majority of this work now exclusively being published online, new technical and ethical questions have arisen for archivists and librarians.

In parallel to RMA's efforts, Concordia University Library began investigating the resource requirements of potential impact of developing its own web archiving program. A pilot project in 2021 uncovered collecting areas of interest, including websites related to Special Collections holdings and web-based scholarly output by Concordia researchers; and created a framework for an operational web archiving program at the library. With only one librarian with web archiving in their job description, ensuring the sustainability of the program is an ongoing challenge. In order to maximize the impact of the library's web collecting activities while balancing its limited resources, we

https://hdl.handle.net/2142/121113

continue to seek out opportunities to create impactful web collections that will benefit Concordia University and the broader research community. The project described in this paper is one example of such an opportunity.

## II. Planning the project

In 2021, we began to envision a collaboration between RMA and the library to collect web-based student-run publications and make them available using Archive-It. Concordia archivists and librarians recognized both the fragility and value of these websites, and believed that archiving them could have a meaningful impact. These publications contain unique scholarly output produced at the university and they hold important evidentiary value in documenting student life and culture. Archiving them would contribute to both RMA's mandate to preserve and provide access to the university's institutional memory and the library's mandate to preserve and provide access to Concordia research.

Except for two literary journals, none of the student publications we identified seem to exist in print form. This makes their contents even more at-risk, as web publications are inherently more fragile than their print counterparts and open access journals regularly disappear from the web [1]. Student-run publications are also particularly ephemeral due to their organizational infrastructure. They tend to be staffed by teams of student volunteers with high turnover, they operate with limited funding, and from what we can tell, most have no long-term stewardship or preservation plan in place. In the year since we started planning this project, we have already witnessed some of these websites disappear.

In 2022, we formed a joint working group which consisted of a small team of archivists, records managers and librarians involved in web archiving at the library and RMA. The working group developed the following plan for the project. We would begin by creating a seed list, i.e., a list of publications to capture and their URLs. We would then contact the editorial teams of these publications to notify them of the project, and begin to crawl the websites using Archive-It.

We expected running web crawls, reviewing results and troubleshooting issues to be the most time-consuming part of this project, especially since neither the library nor RMA has a single employee fully dedicated to web archiving. In order to expedite this process, we decided to enlist the help of a Library and Information Technician intern for spring 2023. Under the supervision of the project lead, the intern will run and review crawls and add descriptive metadata. Once we are satisfied with the results, our last step will be to make the collection publicly available on Concordia University's Archive-It page.

## III. Challenges

Creating an initial seed list proved to be more challenging than we had expected. These publications tend to be siloed in their respective faculty websites which makes them difficult to find. Without a centralized directory, we had to rely on browsing the different department web pages to find their associated student publications, which were often buried many sub-pages deep. Compiling a full inventory of all these publications would be quite onerous, so instead we aimed to create an initial seed list that would represent a sampling, rather than an exhaustive list.

Determining selection criteria for publications was another challenge. Some journals were exclusively managed by students, while others were managed by a combination of students and faculty. Some featured only student work, while others featured work by faculty and authors external to the university. Some websites pushed the boundaries of what we considered web-based publications and we were faced with difficult decisions: should we include a student-run conference website, a scholarly podcast, or a fanzine produced by a special interest club?

We decided to start with a narrow scope to keep this first phase as simple as possible, with the expectation that it could eventually be broadened to include a wider range of publications. We limited our seed list to 16 online journals and magazines that were self-described as exclusively student-run and featuring exclusively student work. We kept a spreadsheet of the websites that we had considered but ultimately scoped out, with appraisal notes explaining our decisions.

Our next step was to contact the editorial teams of our selected publications to inform them of the project and seek their collaboration. This proved to be challenging as well. Many institutions consider an opt-out rather than opt-in approach to be the only workable solution to web archiving, in part due to the typically low response rate from site owners which makes it difficult to obtain explicit permission to archive content [2]. A 2017 survey by the NDSA, Web Archiving in the United States, found that 70% of surveyed institutions capturing content do not seek permission or attempt to notify the content owner that their website is being archived [3].[1]

At the outset of this project, neither the library nor RMA had a formal web archiving policy in place, and the project team was hesitant to make the captured content available without the explicit consent of the website owners. The library is currently in the final stages of drafting a public-facing web archiving policy that will include opt-out and take-down procedures, and we plan to make the captured publications publicly available once this policy is approved. For now, we decided to notify the editorial teams of our selected publications by email, explaining our intention to crawl their site and inviting them to contact us if they have questions or if they wish to opt out. We drafted an email template and kept track of which publications had been notified using a spreadsheet.

As expected, we received few responses and struggled to make contact with the editorial teams. In the cases where we did make contact, we were faced with the challenge of managing the editors' expectations about how their content would be archived. The editors of one journal initially misunderstood the project as a backup service, where captures of their website would be replaced and refreshed periodically. When we explained that the aim of the project was to preserve the content in perpetuity, they chose to opt out, as they wanted to retain control over the archived content and to be able to edit the captured versions of their publication. This exchange highlighted the importance of clearly communicating the objectives and scope of the project to the editorial teams and allowing them to make informed decisions about their participation.

Moving forward, we anticipate technical challenges in crawling some of the publishing platforms that these publications are hosted on. For instance, two of the publications on our seed list are hosted on Issuu, a digital publishing platform that is not easily captured and rendered by Archive-It. As of the writing of this paper, Archive-It's help guide states that while they are working on improving their ability to both capture and replay Issuu publications, "at present, successfully archived Issuu publications will not fully replay" [4]. The wide range of hosting platforms, each with their own technical particularities, means that post-crawl quality assurance and troubleshooting could prove to be significantly time-consuming. We may even need to investigate providing alternative means of access to some captured content, such as that hosted on Issuu.

## IV. Conclusion

With the transition from archiving traditional print-based student publications to publications exclusively hosted online, new challenges and opportunities for archivists and librarians at Concordia University have emerged. These unique publications with no print equivalent are often at risk of going offline or undergoing major transformation rapidly. Factors related to the high turnover of students involved in these projects between academic years (sometimes more frequently) are central to the haphazard management of these websites. These sites regularly contain unique scholarly output, as well as a glimpse into student life and culture. In turn, this type of material holds high research value.

Within the context of this collaboration between RMA and Concordia Library we have presented some of the non-technical and technical challenges associated with archiving student publication websites. On the other hand, these challenges have presented archivists and librarians new opportunities to engage with non-traditional archives users. The discussions and reflections brought on by this project have inspired us to envision and implement new outreach initiatives in the Concordia community. For instance, the library has started to offer regular web archiving workshops to empower students and faculty to preserve their own web content using free and open-source tools. In the last number of years RMA has promoted its web archiving efforts through Concordia-based

[1] As of the writing of this paper, the results of the 2022 edition of this survey have yet to be published. If the trend from previous iterations of the survey maintains itself, this percentage will have increased since 2017.

articles, presentations, various blogs and social media channels.

Through these collaborative projects, continued advocacy and training it is hoped that web archiving at Concordia can foster a more engaged group of stakeholders in terms of preserving and making accessible this type of at-risk information.

## 1. REFERENCES

[1] M. Laakso, L. Matthias, and N. Jahn, “Open is not forever: A study of vanished open access journals,” *The Journal of the Association for Information Science and Technology*, vol. 72, no. 9, pp. 1099-1112, 2021. Available: https://doi.org/10.1002/asi.24460.

[2] C. Davis, “Archiving the Web: A Case Study from the University of Victoria,” *Code4Lib Journal,* 26, 2014. Available: https://journal.code4lib.org/articles/10015.

[3] M. Farrell, E. McCain, M. Praetzellis, G. Thomas, and Thomas, P. Walker, “Web Archiving in the United States - A 2017 Survey,” National Digital Stewardship Alliance (NDSA), 2018. Available: https://doi.org/10.17605/OSF.IO/3QH6N.

[4] M. Praetzellis, “Archiving Issuu and Scribd,” Archive-It Help Center. https://support.archive-it.org/hc/en-us/articles/208333043-Archiving-Issuu-and-Scribd#HowtoarchiveIssuupublications (accessed Feb. 28, 2023).