Good afternoon. I’m Gail McMillan the Director of Scholarly Communication at Virginia Tech’s University Libraries and a founding board member of the Networked Digital Library of Theses and Dissertations. I’m sorry I can’t be with you in Portugal, but thank you for coming to my presentation, “The ETD Lens on the Institutional Repository and the University.”

ETDs form the foundation of many institutional repositories (IRs) that manage and preserve community assets. Institutional Repositories give their institutions both a digital library and a showcase, but I wanted to know whether Institutional Repositories accurately reflect their home institutions. Assessing Institutional Repositories from the perspective of the institution’s scholarly output is not a common frame of reference, so I tested whether ETDs provide a good lens through which to assess whether the Institutional Repository is representative of the scholarship and activities of the home institution. This study was a foray into one aspect of how we might evaluate our universities’ digital repositories.

Populating the VT IR, VTechWorks

- Self-deposit
  - Faculty deposit directly
  - Faculty deposit through integrated systems, e.g., Course Elements
  - Graduate Students’ ETDs through local online system
  - Undergraduates meet course requirements
- VTechWorks staff deposit
  - Unsystematically, e.g., something they read about in the daily VT News
  - Project-based, e.g., OA articles funded by the Open Access Subvention Fund
- SWORD: auto harvesting from some publishers
  - BioMed Central, Hindawi, MOPH

Content in most Institutional Repositories has not been developed like library collections have been developed, that is with librarians who are subject specialists with budgets assigned to represent the university’s research and teaching areas. Instead, Institutional Repositories depend on a variety of people’s time to locate and deposit works. Virginia Tech’s seven-year-old repository, VTechWorks has been populated using a variety of methods.

- There is self-deposit by members of the university community—by choice or by edict.
ETDs, Institutional Repositories and the University

- For example, our faculty deposit directly into the Institutional Repository through integrated systems like Symplectic Elements, which Virginia Tech uses as an electronic faculty activity data system to collect and manage information about research and scholarship, creative works, teaching, extension, outreach, and service activities.
- Graduate students deposit ETDs approved by the Graduate School through a local online system, and
- Some students also deposit as a course requirement.

- VTechWorks staff select and deposit both
  - Unsystematically, for example, based on something they read about in the daily online VT News reports. And,
  - Project-based: such as depositing articles funded by the library’s Open Access Subvention Fund.

- VTechWorks also receives automated harvests from publishers such as BioMed Central, Hindawi, and MDPI who use the SWORD protocol.

VTechWorks hosted more than 71,000 works in April 2019 when I collected the data for this study. About 96% of the content was publicly available and about 85% were textual works.

Among the wealth of articles written about Institutional Repositories, there were very few articles about IR content development. Assessing Institutional Repositories from the perspective of the institution’s scholarly output as I suggest here is an as yet unused frame of reference. Articles about assessing Institutional Repositories tend to cover topics such as usability, preservation, marketing, findability, and intellectual property issues, but they do not attempt to assess the relationship of the content to the research and teaching foci of the home institutions as an indicator of the value of the Institutional Repositories.

Inspiration


My study was triggered in part by these two articles and a presentation that addressed questions about diversity within higher education institutions. At the 2017 CNI fall membership meeting, Amanda Rust from Northeastern University Library in Boston, Massachusetts, presented “Design for Diversity.” She reported on a grant funded project that focused on ways in which information systems embody and reinforce
cultural norms such as data models that enforce strict gender binaries and she addressed designing systems that take into account diverse cultural materials.

Sam Winn, a colleague and Special Collections librarian at Virginia Tech, made several salient points in her article, “The Hubris of Neutrality in Archives.” She wrote that

- Professional practices have upheld, and even facilitated, the institutionalized dehumanization of [among others] trans and gender-nonconforming people… [p1]
- She pointed out that …archivists contribute to the omission or erasure of historically marginalized groups in the archives. [p2]
- And that …a “radically inclusive historical record” will not happen by accident. [p2]

Rebekah Scoggins, librarian at Leander University in Greenwood, South Carolina, published in *College and Research Libraries News* her report on “Broadening Your Library’s Collection: Implementing a LGBTQIA Collection Development Project.” It struck me as a well-aimed study but one that was limited because it only considered the traditional library collection, that is purchased books, serials, and other media, but not the content of the Institutional Repository. She determined that her library was not meeting the needs of its users because it was out-of-date and incomplete.

Rebeka’s article in particular made me ask, “how would we know if our repository, VTechWorks, lacked diverse cultural materials and was contributing to the omission of historically marginalized groups?” Because of these influential works and the lack of articles about Institutional Repository content development, I chose to conduct a study that might help me determine if my repository was contributing to the omission of the works of marginalized people or providing an inclusive record. I decided to test it on a microcosm, that is a community that encapsulates in miniature the characteristics of the whole. I analyzed the microcosm of LGBTQ works in my repository, VTechWorks, and in Virginia Tech’s ETDs in particular, and compared the findings to the output of the university, including Virginia Tech’s faculty and academic units.
From several academic and community resources I compiled a vocabulary of 161 LGBTQ terms and phrases. These resources included *A Progressive’s Style Guide* [from SumOfUs], the University of Massachusetts at Amherst’s Stonewall Center, and Virginia Tech’s Safe Zone. I excluded some terms because their meanings have changed radically over time such as gay and queer. Some terms were not included because the Repository software, DSpace, uses the Solr search platform, which allows “fussy” matches. That is, Solr retrieves close matches such as sexual when the search term was sexuality. Solr also “stems,” that is, it expands words with common endings to include plural, past tenses, and the like. I used the common search practice of putting quotes around phrases. (By the way, in order to reduce wordiness, when I refer to “terms,” I mean both terms and phrases.) This list of terms may not be bias free or comprehensive, but it served as a starting place for inquiry into the LGBTQ microcosm at Virginia Tech.

161 terms/phrases: 5 collections

<table>
<thead>
<tr>
<th># items</th>
<th>collection</th>
<th>terms matched</th>
<th>hits</th>
</tr>
</thead>
<tbody>
<tr>
<td>71,734</td>
<td>VTechWorks</td>
<td>129</td>
<td>21,455</td>
</tr>
<tr>
<td>32,557</td>
<td>ETDs</td>
<td>115</td>
<td>9,017</td>
</tr>
<tr>
<td>3,870</td>
<td>Faculty Research</td>
<td>40</td>
<td>858</td>
</tr>
<tr>
<td>14,285</td>
<td>Colleges</td>
<td>89</td>
<td>406</td>
</tr>
<tr>
<td></td>
<td>Virginia Tech vt.edu</td>
<td>109</td>
<td>84,793</td>
</tr>
</tbody>
</table>

To refine my investigation, to help understand who is doing the scholarship and research in the LGBTQ microcosm, and to give context to the data, I searched the 161 terms in five collections: the university at large through its website, vt.edu, VTechWorks at large as well as within its collections of ETDs, peer-reviewed faculty publications, and academic units’ web publications. We call these units colleges at Virginia Tech so in this presentation I refer to the Colleges Collection. Searching the terms in VTechWorks, produced 21,455 hits on 129 terms in 71,734 works. I searched the terms in the collection of >32,000 ETDs; 115 terms got about 9,000 hits. The faculty research collection got 858 hits on 40 terms in nearly
4,000 peer-reviewed articles, and the Colleges Collection got 400 hits on almost 90 terms in about 14,000 works.

Searching VT at large, 109 of the terms got nearly 85,000 hits. I did not search from the university’s homepage because indexing of Virginia Tech’s websites includes VTechWorks. Instead, I searched the terms in the Google.com search box which allowed me to exclude the VTechWorks site when searching vt.edu [At https://www.google.com/ I used this search strategy: [term] site:vt.edu -site:vtechworks.lib.vt.edu -site:theses.lib.vt.edu].

Because these five collections have radically different sizes and ages, I converted the number of hits for each term in each collection into a percentage of hits within its collection. Then I compared the percentage of hits within and across collections which provided the lens of the graduate students through their ETDs, the faculty’s use of the terms in their peer-reviewed publications, and the academic units’ and university-wide use in works often aimed at the general public and alumni.

Creating spreadsheets for the search results from each collection allowed me to sort the data in various ways. This table shows highest percentage of hits sorted by hits in the ETD collection. As you can see, the term “gender” got nearly 50% of the hits in the ETD, faculty research, and colleges collections. In the ETD collection that left the remaining 114 terms with hits between (three point five percent) 3.5% and 0.01%.

I initially thought it would be interesting to compare faculty and graduate students’ non-use of the terms also. But, you may recall that of the 161 terms I searched, 115 appeared in ETDs but only 40 appeared...
in the Faculty Research Collection. There were 75 unused terms, which was too many to be meaningful for this study.

This chart shows that graduate students used some terms much more frequently than did faculty. But, faculty only used two terms more than Graduate Students, gender bias and gender expression.

This is when I wish I was a linguist or had collaborated with one in order to determine if there is some correlation among the terms used more frequently by one type of author, that is graduate students vs. faculty.

This is my last table and it shows search results for the top 20 hits in the university-wide collection. There were 4 notable disparities in the terms used more frequently at vt.edu than in the ETD Collection: sexual orientation (13.8% vs 2.9%), gender identity (7.8% vs 2.0%), and gender expression (7.0% vs 0.2%). These terms appear more in public-facing web documents than in graduate student’s academic research. Only one term appeared with considerably more frequency in ETDs: gender (49% vs 39%), that is “gender” was used more in graduate students’ research than in the academic units’ public facing university web sites. After analyzing terms that appeared in ETDs but not in VT at large and vice versa, the disparities were too small to be noteworthy for this study.

In conclusion: Are they similar enough?

If it is a valid method

- Does the IR reflect the work being done at the institution?
  - 94% of the terms in VT are in VTechWorks.
  - 94% of the LGBTQ terms in VTechWorks have <2% difference in the percentage of hits.

- Do ETDs reflect the work being done at the institution?
  - 95% of the terms in VT appeared in ETDs.
  - 92% if the terms have <1% difference in the percentage of hits.
The goal of this study was to investigate whether the institutional repository represents the scholarship and activities of its home institution, and whether this can be determined by studying a microcosm of the Repository, and particularly its ETD Collection, and comparing it with the same microcosm in the university. I used a controlled vocabulary to analyze the LGBTQ microcosm in the university website, vt.edu, and VTechWorks, particularly its ETD collection. Using a microcosm to look at graduate students’ ETDs and the university website provided the opportunity to reflect on the contribution of ETDs to the development of institutional repositories through a manageable data set. The percentage of search hits on the selected terms at the university’s website was the measure of the LGBTQ domain and contextualized the data from ETDs as well as the whole Repository.

Since 94% of the LGBTQ terms appearing in the university website through a Google search of vt.edu also appeared in the VTechWorks, and 94% of the terms had a <2% difference in the percentage of hits, I say that the institutional repository as a whole reflects quite well the scholarship and activities at the university.

Of the 109 terms that ETDs in VTechWorks and VT have in common, 99 vary in their occurrence by less than one percent. That is, 91% of the terms have the nearly equal percentage of hits in each collection. Is this also a high enough correlation to say that the ETD lens quite accurately reflects the scholarship and activities at the university, at least in the LGBTQ domain? I think so.

It will be necessary, of course, to study additional microcosms in order to better understand whether this is a high enough correlation to determine if the Institutional Repository and more particularly its ETDs are truly representative of the scholarship and activities at the university. I have started to compile a vocabulary to study the microcosm of Indigenous Peoples and to refine my search strategy to always use Google search instead of the Solr search built into the DSpace repository platform.

I would be very grateful if conference participants would send me their comments on the study methodology and the criteria I used to assess whether the Institutional Repository accurately reflects the scholarship and activities of its home institution. If anyone is interested in conducting similar studies and comparing across institutions, please let me know. My email address is gailmac@vt.edu. Thank you!