ScholarlyCommons FY17 Report

ScholarlyCommons (http://repository.upenn.edu/) has served as the University of Pennsylvania's open access institutional repository since 2004. The repository has allowed for faculty and others to share scholarly publications publicly and without paywalls. For the first ten years, it saw slow but steady growth in the deposit of a relatively small number of publications, as well as dissertations and theses. Since 2014, ScholarlyCommons has grown dramatically to include a wide variety of collections from Penn community members.

ScholarlyCommons saw an array of positive developments in FY17. The number of materials posted to the repository hewed closely to the strong upward growth trends of the past three years, new staffing and training workflows were developed, and a slate of services were introduced that make it easier for the Penn community to capitalize on our offerings. This report will provide an overview of collection statistics, staffing, and other key developments in FY17. This report will also provide a general snapshot of overall repository growth.

Quick Repository Facts (*as of October 17, 2017*)

- 32,9071 items available
- 13,328,144 total downloads[1]
- 4,057 works added in 2016 (2,172 so far in 2017)

Table of Contents

Improvements to Services and Staffing

This year saw significant changes to ScholarlyCommons staffing and services. The arrival of the Digital Scholarly Publishing Librarian, Kenny Whitebloom, in September 2016 allowed for the development of a coherent slate of ScholarlyCommons services, as well as a refined series of workflows and processes. These developments, which built off previous efforts by the former repository manager, allowed for

[1] There has been some scuttlebutt on the DigitalCommons listserv about a systemic drop in download over the course of FY18. We've noticed a slight dip as well, but it's unclear at this stage whether or not it's the result of a technical change to the platform or another issue.

increased outreach to faculty, centers, and other research outfits across campus and resulted in the largest number of faculty uploads in a single year.

ScholarlyCommons Services
In an effort to make outreach and marketing more coherent to the Penn community, ScholarlyCommons staff developed a more service-oriented approach in FY17. This refined approach included the development of collection-based and mediated deposit services (Faculty Assisted Submission, described below), both of which are now described in detail on a LibGuide and the ScholarlyCommons repository website.

The collection-based services that ScholarlyCommons offers include:

- Journal publishing
- Digital projects
- Dissertations, theses, and capstones
- Conferences and events

For all of these collection-based services, we offer the same degree of consultation and training support throughout the project lifecycle. This year our team developed a number of internal project management documents that have made tracking and following through on these types of consultations more systematic. We shared these workflows with the broader IR community in July 2017.

Faculty Assisted Submission
Faculty Assisted Submission (FAS) is a mediated deposit service offered to all Penn faculty in which we import a faculty's CV into a spreadsheet, determine who the publisher is for each of their works, and then figure out whether or not those publishers allow us to deposit the faculty's work(s) into an institutional repository like ScholarlyCommons. Based on that research, we then deposit what we can in ScholarlyCommons and/or reach out to the faculty member for copies of their works for which we need a non-final version. This process can be time-intensive and requires specialized training for student employees around copyright, open access, and publishing.

Mediated deposit underwent an overhaul this year, as we rebranded "CV Deposit" into the aforementioned "Faculty Assisted Submission" in an effort to make the service more intelligible to faculty. We also created a comprehensive information page devoted to the service, complete with FAQs and a drag-and-drop upload page for faculty to easily upload their CVs/list of publications for review. Over the course of FY17 (and into FY18), the amount of time spent processing these CVs has decreased significantly, reflecting our improved training processes and permissions workflows.

Notable FY17 Faculty Assisted Submission statistics include:

- 1,846 citations reviewed for copyright permissions.
- 21 faculty CVs reviewed (not including Wharton).
- 59 days on average to complete a CV containing 84 citations, the average CV length.

- *(Note: The amount of time steadily decreased over the course of the year and into FY18, as workflows were refined and student training was improved and codified. For instance, in FY18 the time to complete an average length CV was less than half of the time in FY17.)*

Faculty responses to FAS have been largely positive, especially from those that have provided us with postprints or other non-final versions of their papers (which allows us to deposit most of their material). We encourage FAS users to complete a short survey at the end of the process, and while we've only received two responses to date, both of them have given the service a 5/5 rating. Informal conversations with participating faculty complement these survey responses.

Publisher Policy Database

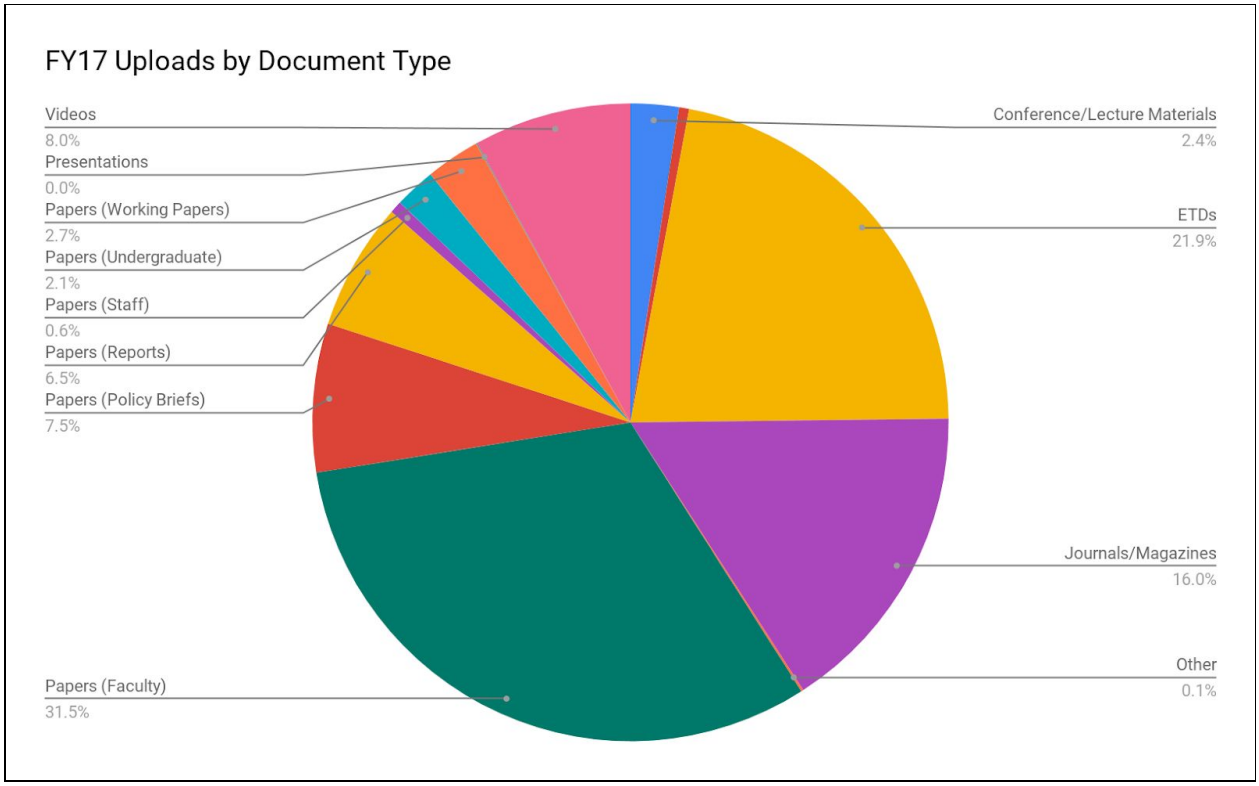
This FAS work is further aided by the development of our Publisher Policy Database, an expanding database of information about journal publishers and their permissions policies specifically pertaining to archiving in institutional repositories. Soft-launched in Spring 2016 and still under active construction, the Publisher Policy Database is designed to capture journal publisher policies, requirements, and other important details about self-archiving articles in non-profit institutional repositories. The goal of this database is to leverage our extensive in-house knowledge and day-to-day experience reviewing publisher policies to provide a more responsive, timely alternative to SHERPA-RoMEO, a public resource for publisher policies that isn't specifically tailored to IR depositing and often contains out of date information. We unveiled the Publisher Policy Database to the Digital Commons community in July following a presentation. The response was uniformly positive and we intend to build off that response as we further improve and populate the database.

Student Employee Training and Supervision

FY17 saw improvements in student employee training, as we developed a systemic, multi-part workflow for teaching copyright basics, permissions, and uploading/editing in the repository. We currently have 4 student employees working 10-20 hours per week (during the summer they work closer to 35 hours per week). This workflow, which has been refined over the course of FY17 and FY18, has allowed us to fully train and equip student employees for independent permissions review in 15-20 hours, depending on the student's abilities and schedule. Additionally, our improved permissions spreadsheet, BamBam project management tools, and Publisher Policy Database have made the permissions process even more efficient and standardized.

FY17 Collection Growth

In FY17, repository staff and administrators published 2,267 objects to the repository.[2] The three major document types uploaded to the repository in FY17 were faculty papers (31.5%), Electronic, Theses, and Dissertations (ETDs) (21.9%, which does not include 129 abstract-only records in which the full-text of the ETD is not included in the record), and works published in a journal structure (16.0%). Notably, more than 1,000 papers written by Wharton faculty were reviewed and uploaded to the repository as part of an ongoing project with the business school; these papers are awaiting final approval and release by Wharton admins. As of October 2017, the total amount of content uploaded to the Wharton site, and awaiting rollout, stands at more than 2,100 papers.

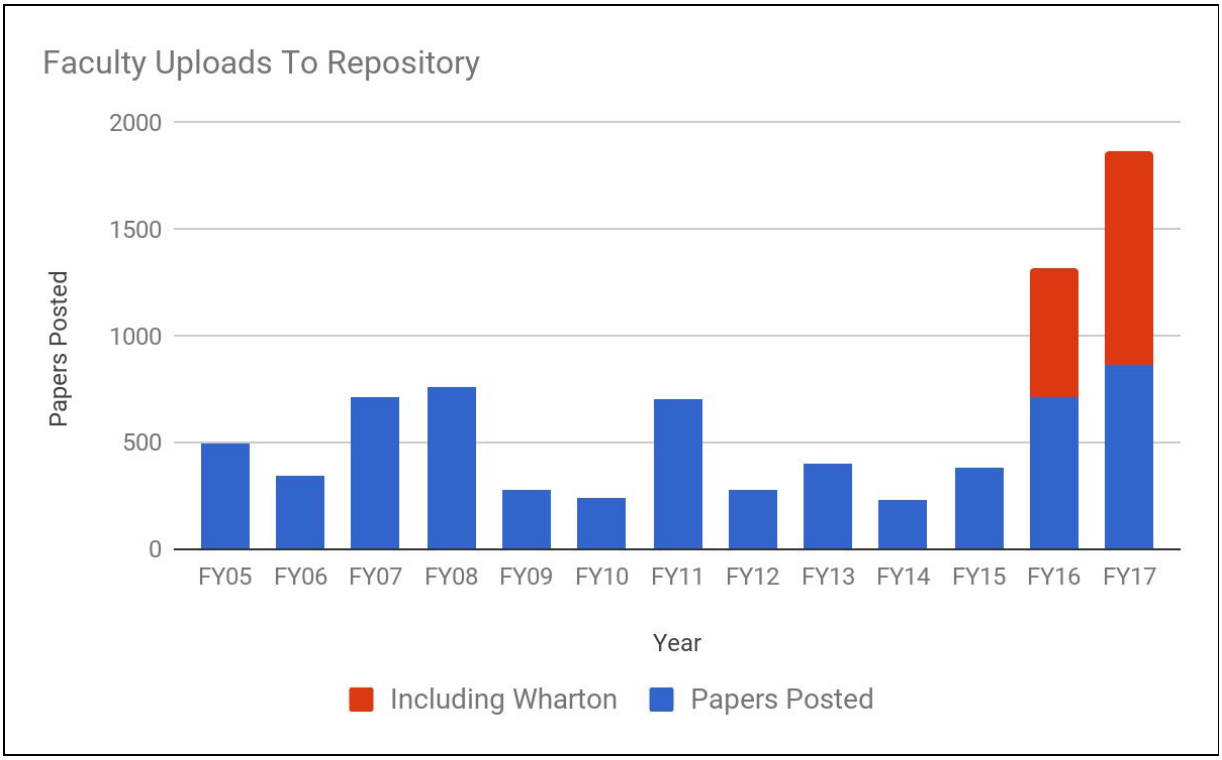


A few key takeaways on collection growth in FY17:[3]

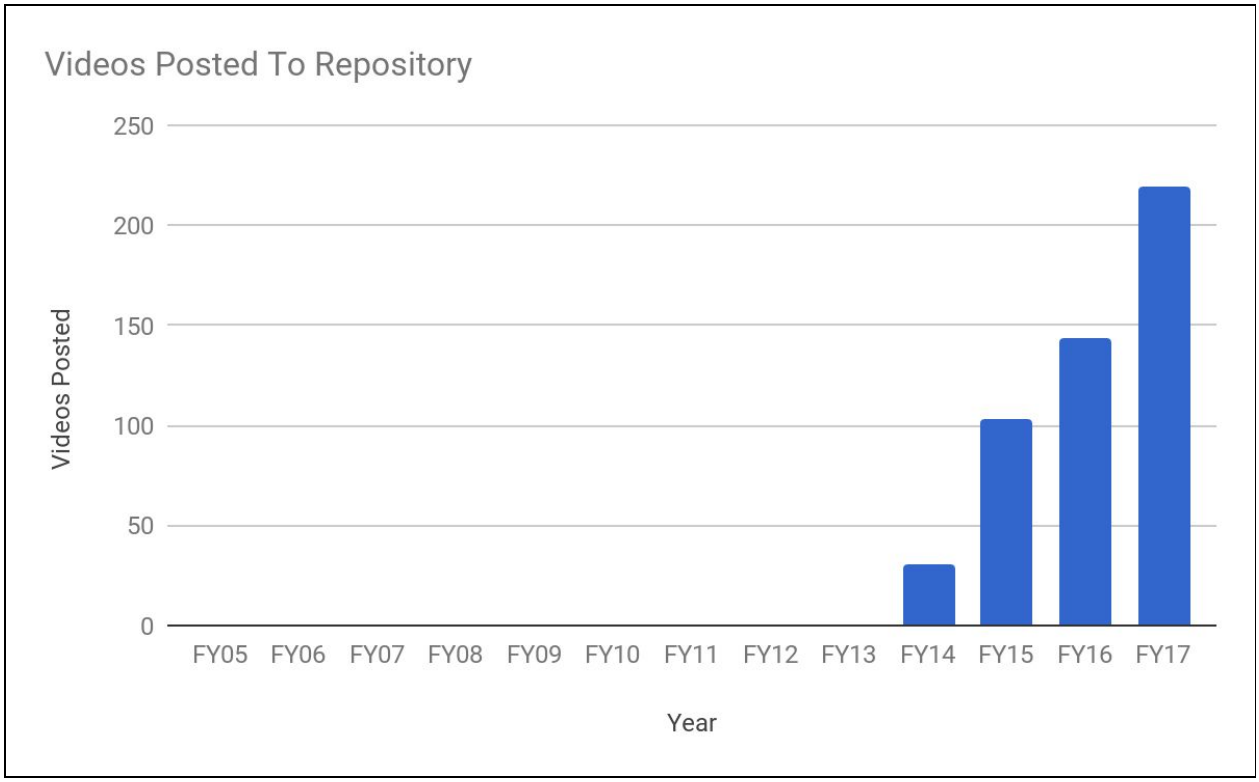
- **Faculty Papers (journal articles, book chapters, and other previously published material)** comprised 31.5% of all documents posted to the repository, the greatest amount in a single year (862). This growth in faculty uploads is a direct result of increased staffing and rebranding around our mediated deposit service, Faculty Assisted Submission, which is described in greater detail above. This growth is notable given the student employee time devoted to the ongoing

[2] This figure excludes ProQuest Dissertation abstracts, which are metadata-only records that link to content hosted on ProQuest Dissertations, and handful of other metadata-only records.
[3] Publicly Accessible Penn Dissertations were not batch uploaded in FY17, which deflates the amount of materials posted.

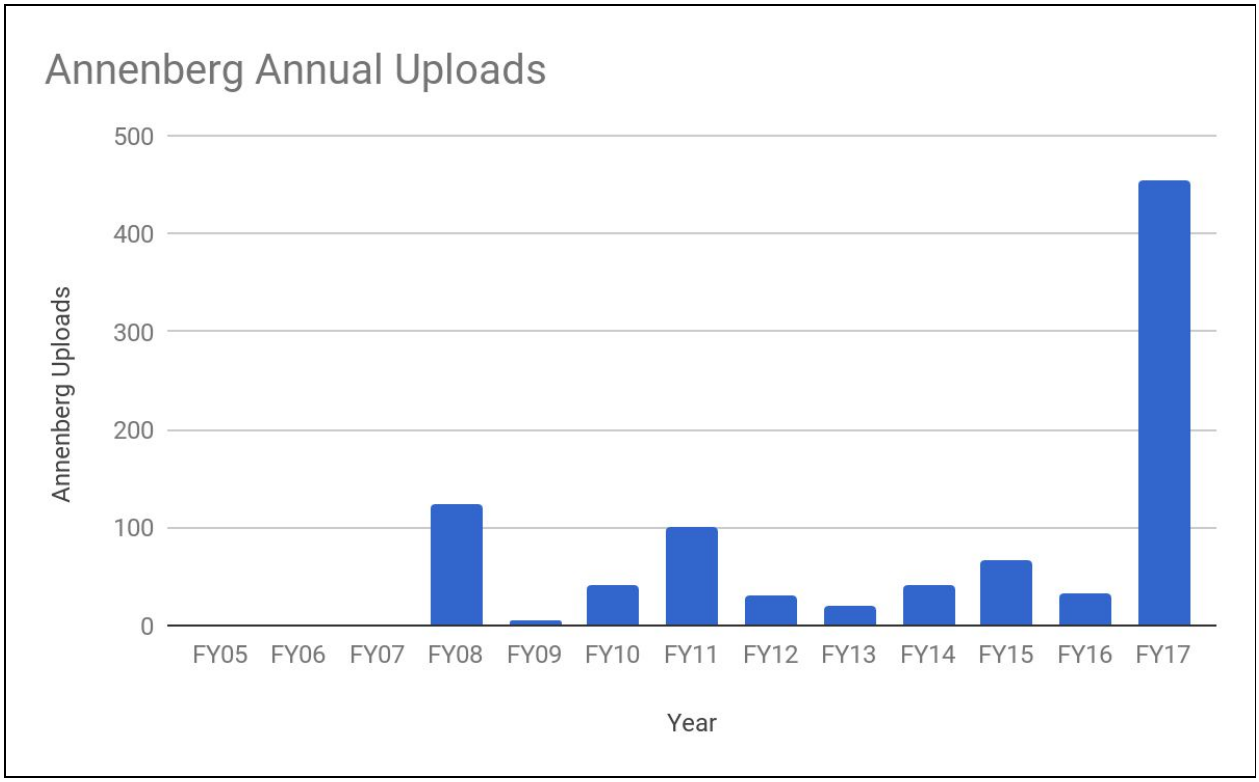
Wharton faculty permissions project during which we reviewed and posted approximately 1,000 papers.



- **Reports and Briefs** produced by research centers, units, or other Penn-affiliated groups comprised 14.0% of all uploads this year (combined). The uptick in this category of materials is a result of increased outreach to research centers, such as the Social Impact of the Arts Project and Penn Wharton Public Policy Initiative.
- **Works posted to journal structures** comprised 16.0% of all content added to the site, a decrease from last year but consistent with a three-year upwards trend of journal structure creation and publishing. This continued growth is a result of concerted marketing/outreach efforts on the part of ScholarlyCommons staff and the rollout of free journal structure web design by bepress in Spring 2014.
- **Video and Multimedia** is a small but growing slice of the content uploaded to the repository (219 videos published, 8.0% of content published). This growth is attributed to a couple of MOOC collections with the Online Learning Initiative. We also collaborated with Arthur Kiron (Penn Libraries and Jewish Studies Program) to upload a series of oral histories with members of the Jewish havurah counterculture movement of the 1960s. These videos, which are to be published exclusively on ScholarlyCommons, will be public in FY18. Furthermore, ongoing exploratory conversations and research with the Penn Immersive group have opened up the door to the creation of a suite of collections devoted to the publication and storage of VR, AR, and 3D files. As the libraries continues to enter into the immersive tech space, this type of interdisciplinary multimedia will present interesting opportunities for the repository in whatever form(s) it takes.

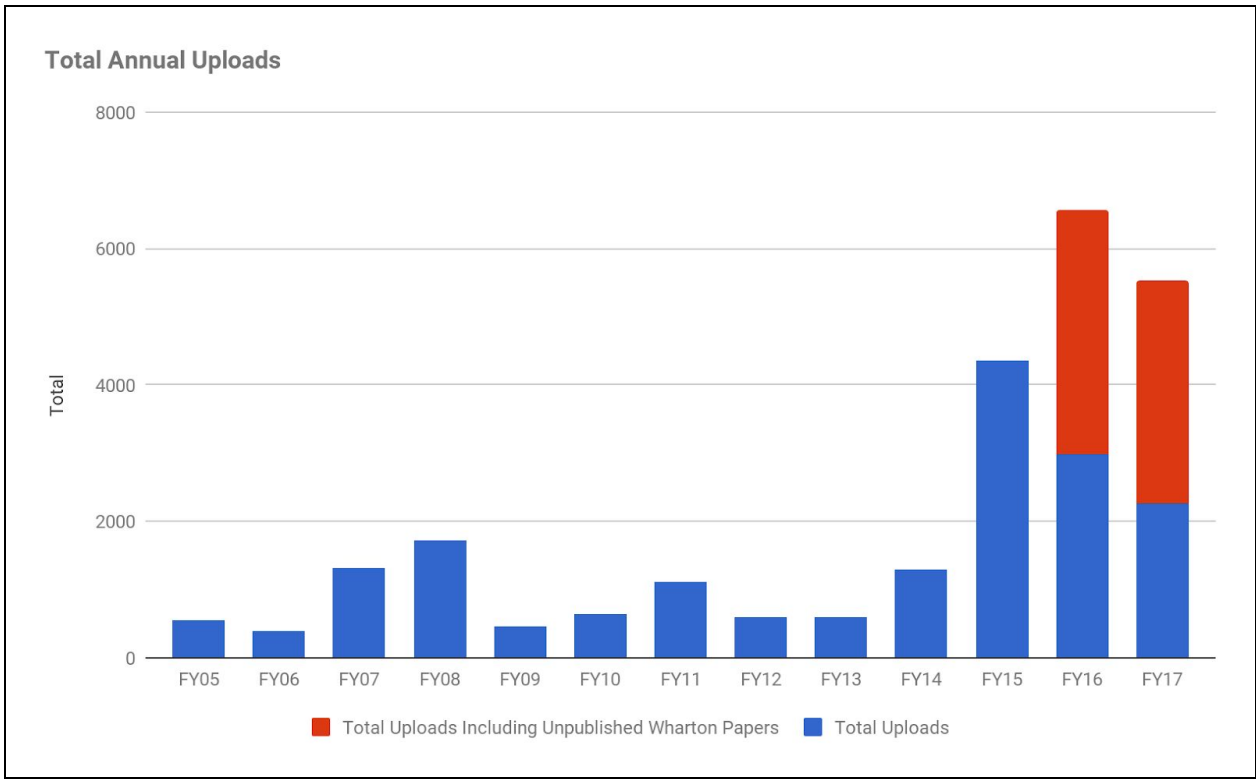


- **Annenberg School for Communication** has seen solid growth in its departmental papers series due to our special workflow in which we run permissions on faculty citations and they upload papers the actual papers, including postprints. This particular workflow, which takes some of the workload off our shoulders, has freed our staff up to run more permissions, more quickly. We have used a modified version of this workflow with other projects.


Annenberg Annual Uploads
500
400
300
200
100
0
Annenberg Uploads
FY05
FY06
FY07
FY08
FY09
FY10
FY11
FY12
FY13
FY14
FY15
FY16
FY17
Year

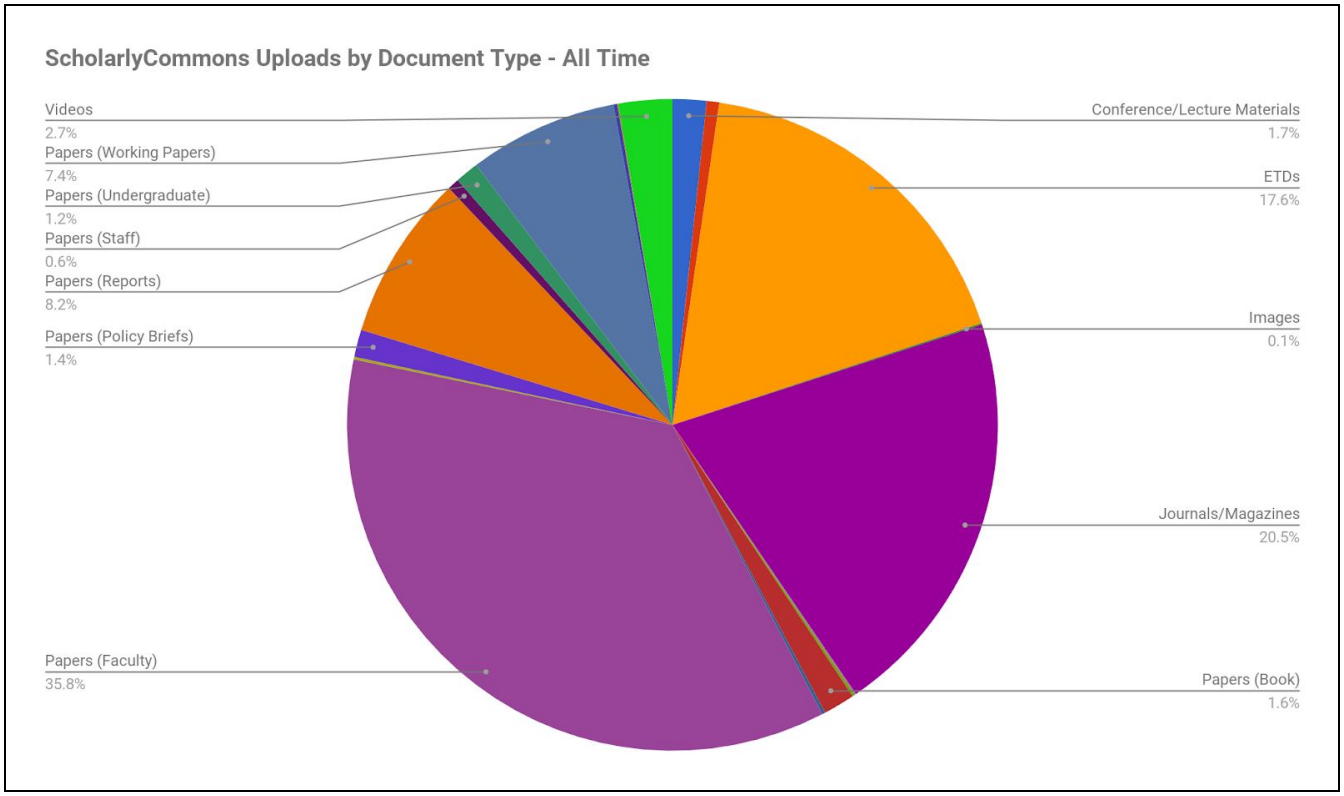
Overall Repository Composition

Since its launch in February 2005, ScholarlyCommons has evolved to serve as a publishing platform for a wide range of different item types and users. This growth has accelerated significantly beginning in FY15.



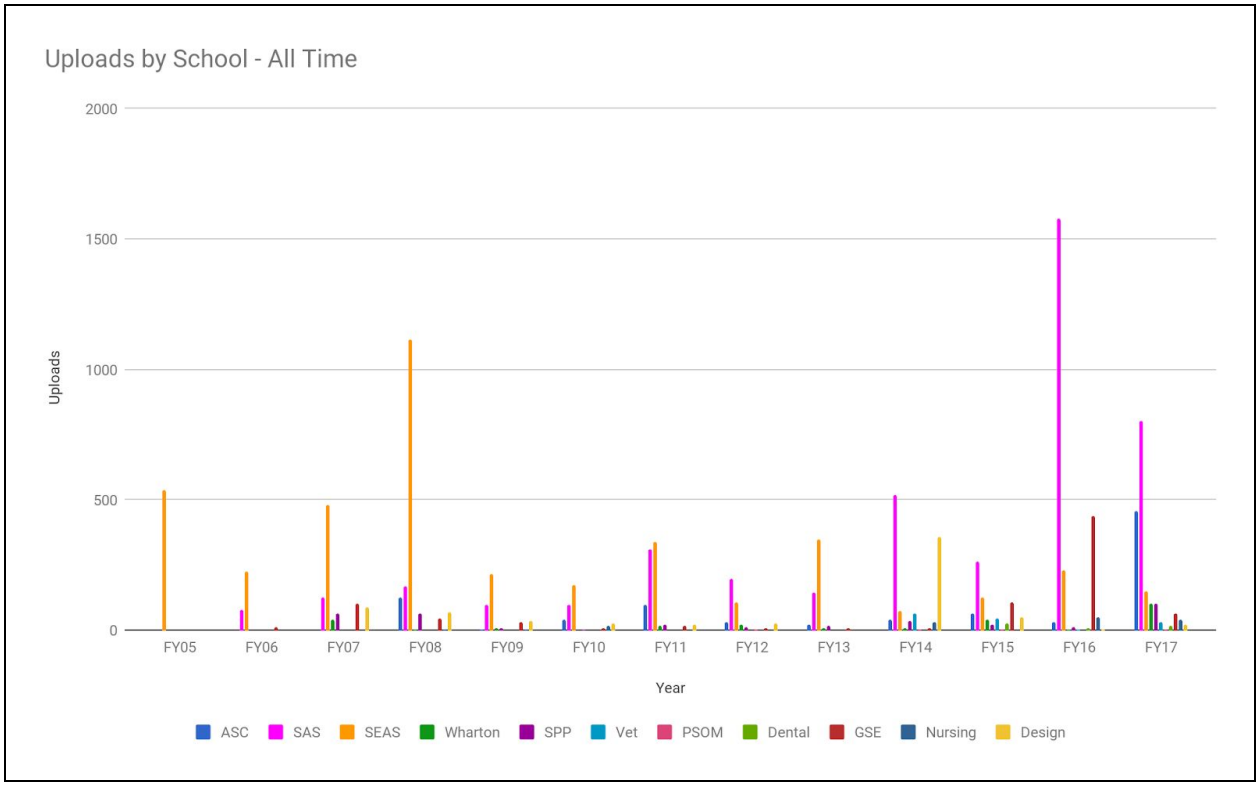
Reflecting our positioning of the repository as a platform for faculty to store and disseminate their scholarship, 35.8% of ScholarlyCommons' materials are faculty papers (including the unpublished Wharton papers that number increases to approximately 42.6%). Full-text ETDs comprise under a quarter of the repository (17.6%)[4], though this figure will increase as the Publicly Accessible Penn Dissertations collection grows and the institutional upload workflows are more firmly entrenched. Working papers, reports, and briefs together make up 17.0% of the repository's holdings and appear to be an area of stable growth. Conference and event collections, which make up just 1.7% of the repository, are rarely requested by parties outside of the library and are not currently a focus for outreach efforts.

[4] This figure does not include 13,270 abstract-only ETD records from the Master of Applied Positive Psychology program and ProQuest Dissertations. When included in the overall total, ETDs jump up to nearly half of the repository's holdings.



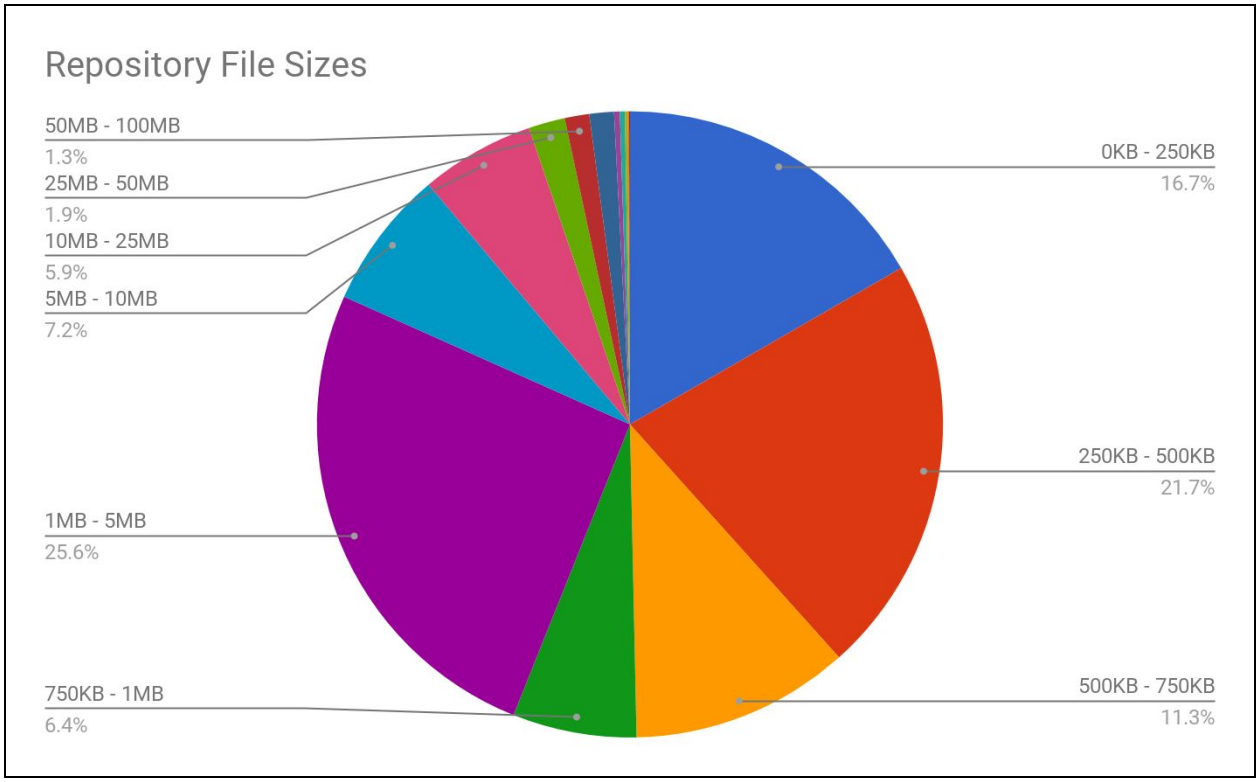
Journals, which comprise 20.5% of the repository, present a compelling example of repository use that cuts across academic disciplines and user groups (students, graduates, faculty, and staff). ScholarlyCommons' built-in peer review structures, free custom web design, and permanent URLs have proven to be powerful draws for admins interested in hosting their publication(s) on the repository. Mirroring and ultimately expanding the suite of functionality offered to current and prospective journal users is important for future development of a library publishing program.

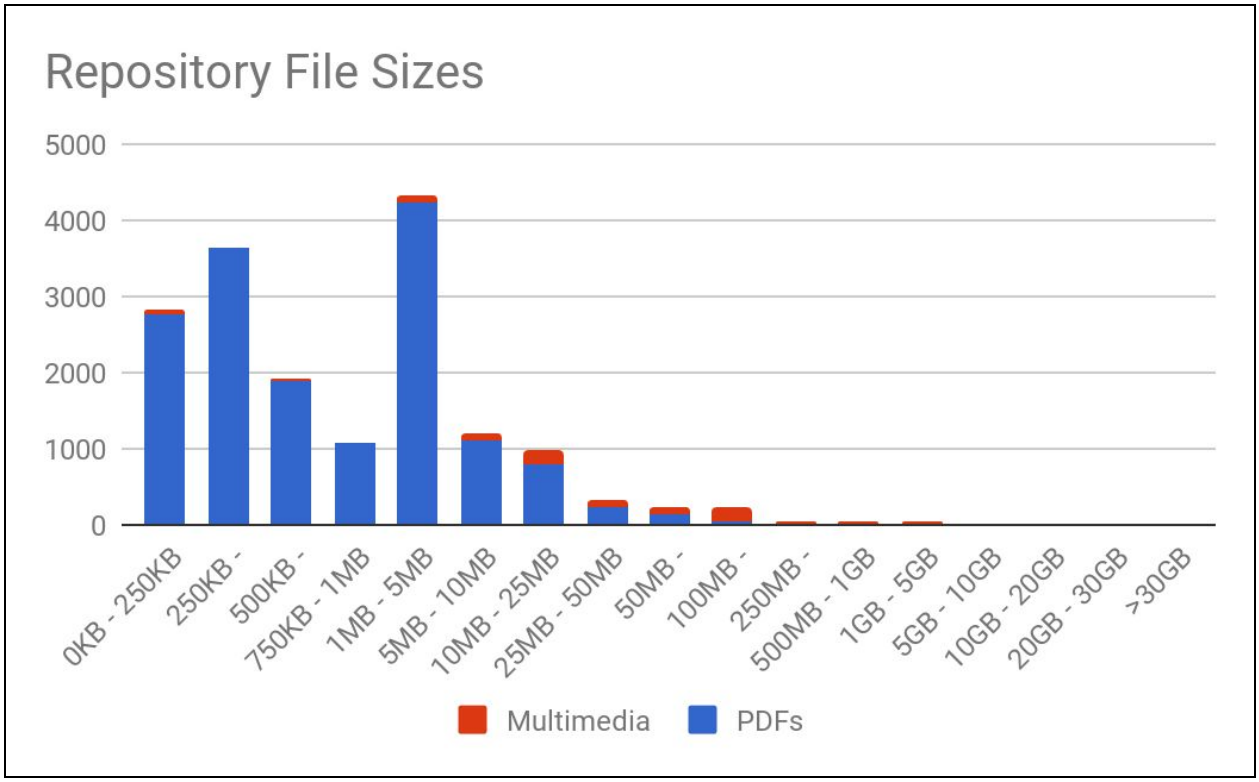
School uploads have become more evenly distributed over the past few years, as evidenced in the chart below. FY17 in particular saw a healthy distribution of uploads across the participating schools.



In terms of file sizes within the repository, the vast majority (89.9%) are less than 10MB.[5] The largest items, unsurprisingly, are multimedia files, though even those tend to be smaller than 500MB. The largest items in the repository are the unpublished Jewish Counterculture Oral History videos which range in size from 12GB - 25GB.

[5] Based on report generated on October 17, 2017. This figure does not include supplemental content.


Repository File Sizes
50MB - 100MB
1.3%
25MB - 50MB
1.9%
10MB - 25MB
5.9%
5MB - 10MB
7.2%
0KB - 250KB
16.7%
250KB - 500KB
21.7%
1MB - 5MB
25.6%
750KB - 1MB
6.4%
500KB - 750KB
11.3%


Repository File Sizes
5000
4000
3000
2000
1000
0
0KB - 250KB
250KB -
500KB -
750KB - 1MB
1MB - 5MB
5MB - 10MB
10MB - 25MB
25MB - 50MB
50MB -
100MB -
250MB -
500MB - 1GB
1GB - 5GB
5GB - 10GB
10GB - 20GB
20GB - 30GB
>30GB
Multimedia
PDFs

Lastly, and perhaps unsurprisingly, the overwhelming majority of the materials in ScholarlyCommons, including supplementary content, are PDFs (92%). The second most popular file type is multimedia, comprising 4.3% of the repository across eight different file formats.

File Format	Total	Percentage of Repository
Multimedia (.mp3, .mp4, .avi, .flv, .m4v, .wmv, .mov, .webm)	776	4.3%
Tabular Data (.xls, .xlsm, .xlsx, .csv)	81	0.4%
Image (.tiff, .jpg, .png, .gif, .img)	75	0.4%
Presentation (.ppt, .pptx, .ppsx)	91	0.5%
Documents (.txt, .doc, .docx, .epub, .html, .rtf)	394	2.2%
PDF	16,626	92%
Other (.zip, .ps, .obj)	22	0.1%

Collection and Service Gaps

For all of its strengths, ScholarlyCommons is not a great fit for every project that comes our way. These collection or object types include:

- **Large datasets or files**: we receive a number of emails each semester from faculty seeking a platform to store their large datasets or files. In addition to upload reliability issues with files larger than 1GB, ScholarlyCommons, built on an extensible Dublin Core metadata schema, is often not positioned well to describe these materials for a technical audience (unless we work closely with the admins to create a specialized collection).
- **Large multimedia files**: similar to large datasets, ScholarlyCommons can accommodate large video or multimedia files, but the upload process is not wholly reliable and the viewing experience is limited at best. (*Note: a streaming media pilot, which is forthcoming from bepress, will potentially alleviate the streaming problem.*)
- **Exhibitions and curated collections:** visual collections that require a modern, flexible front-end (à la Omeka or even WordPress) can be shoehorned into a ScholarlyCommons' image gallery, but there is no way to curate those materials into a coherent narrative structure or display. Furthermore, ScholarlyCommons' lack of an API or other interoperable technology makes it burdensome to use images in ScholarlyCommons on other platforms without duplicating efforts and bifurcating impact.
- **Image and other media incorporation**: it is currently overly complex to embed or include an associated image in a record, as some applied engineering disciplines like to do in order to highlight/foreground their research outputs (*see Kod*Lab*). With bepress, you can create a separate gallery structure and then embed those images in the item's record, but this is difficult and time-consuming for the average administrator. This problem also corresponds to a general difficulty of creating flexible relationships between collections in ScholarlyCommons.
- **Websites and blog content**: the preservation of entire websites is virtually impossible in ScholarlyCommons apart from the bundling of content into a .zip file.
- **Other technical functionality**: ScholarlyCommons has a number of other technical limitations, including: difficulty in re-imaging, gathering, or organizing an existing collection; inability to generate author-based RSS feeds apart from SelectedWorks; and inability to view author pages apart from preloaded search results or SelectedWorks pages, which duplicate content and exist semi-separately from the repository.

Conclusion

ScholarlyCommons experienced a range of positive developments in FY17. The introduction of new staffing allowed for increased attention towards services and project management, which resulted in the rollout of the rebranded mediated deposit service, Faculty Assisted Submission. Collection development continued in line with growth over the past few years, demonstrating a healthy level of departmental representation and participation. The turn towards a platform-agnostic, service-based approach augurs well for the future, as the scholarly communication team begins the process of exploring alternatives to bepress now and into 2018. In addition to this exploratory work, the ScholarlyCommons team in the months ahead plans to work more intensely on populating the Publisher Policy Database, developing outreach and educational materials around issues related to scholarly communication, and further refining and improving our workflows for permissions and deposit.