

Article

Understanding Connections: Examining Digital Library and Institutional Repository Use Overlap

Mark E. Phillips , Pamela Andrews *  and Ana Krahmer * 

University of North Texas, Denton, TX 76203, USA; mark.phillips@unt.edu

* Correspondence: pamela.andrews@unt.edu (P.A.); ana.krahmer@unt.edu (A.K.)

Received: 15 March 2019; Accepted: 28 May 2019; Published: 8 June 2019



Abstract: The University of North Texas Libraries' Digital Collections are situated as a unified whole within their preservation infrastructure, with three separate user interfaces serving the content to different audiences. These separate interfaces are: The UNT Digital Library (DL), The Portal to Texas History, and The Gateway to Oklahoma History. Situated within each interface are collections, and hosted within these collections are digital objects. One collection, the UNT Scholarly Works Repository, specifically serves UNT's research and creative contributions and functions as the Institutional repository (IR) for the University of North Texas. Because UNT Scholarly works is seated as a collection amongst other collections, users can access faculty research, not just out of an interest in research from specific faculty members, but also as it ties into the user's broader understanding of a given topic. With flexible infrastructure and metadata schema that connect collections beneath the umbrella of the wider preservation infrastructure, the UNT DL employs full-text searching and interlinked metadata to strengthen and make visible the connections between objects in different collections. This paper examined how users navigated between other collections within the UNT IR, as well as within the UNT DL. Through this examination, we observed patterns between how users navigated between objects, understood which collections may have related to one another, examined why some unique items were used more than others, and viewed the average number of items used within a session.

Keywords: institutional repositories; user interfaces; usage; digital libraries

1. Introduction

At the 2017 National Digital Newspaper Program meeting, sponsored by the National Endowment for the Humanities (NEH) and the Library of Congress, a recurring conversation concerned how users interacted with digital collections. Attendees expressed interest in observing how a user's individual usage session with a digital collection might include interactions with different types of digital objects, with distinct collections tied together through facets, or with objects from entirely different subject matters. This discussion resulted in a communal acceptance that enabling access to dissimilar but related (through metadata) object types was positive for the user experience. However, meeting participants also noted how and if users actually did crossover to different digital collection areas had been rarely been measured.

The UNT Libraries' Digital Collections encompasses three different interfaces: The UNT Digital Library, The Portal to Texas History, and The Gateway to Oklahoma History. Each of these interfaces contains a number of collections holding digital objects. By placing these collections alongside other collections within one of our three interfaces, users have the potential to discover meaningful connections to other collections and digital objects. For example, research projects that utilize newspapers may also benefit from photographs, maps, audio or video. Arranging metadata facets enables users to create a constellation of digital objects across diverse resource types and collections,

which increases usage of these objects and enriches the overall research process. While enabling user research in different ways is a primary motivator that informs the infrastructure of the UNT Libraries' digital interfaces, we recognized the need to study the effectiveness of this concept on both a local and profession-wide scale. The goal of this study was to examine usage data from a specific interface, The UNT Digital Library, to answer three questions:

1. To what extent do users discover and use items from different collections?
2. To what extent do users discover and use items that are of different types?
3. To what extent do users discover and use items that are from different contributing partners?

As previously mentioned, all of the collections housed within a single digital interface can be located from a unified search box on each system. This allows a user to discover resources of all types from different collections housed within that interface. These collections are the primary way we group resources and provide additional context to aid users in their research. The higher level interfaces that group digital objects and collections into a unified view are separated to enable different user groups to interact with content. The Portal to Texas History is a repository that works with over 400 partner institutions around the state of Texas to provide access to cultural heritage objects freely to the public. The UNT Digital Library provides access to content created or collected by UNT and The Gateway to Oklahoma History encapsulates digital resources from the Oklahoma Historical Society that are collaboratively hosted by the UNT Libraries. This separation of different interfaces for different high-level content and users allows for us to tailor the interfaces slightly for the users of that interface.

This paper seeks to gather quantitative data to observe usage patterns between different item types, collection themes, and contributor characteristics for one digital library collection access interface.

2. Review of Literature

Anecdotal stories about user interactions across digital library collections are extensive in digital libraries, but limited peer-reviewed research is available that directly examines usage data as it correlates to user behavior across mixed collection and object interfaces. This made compiling a review of extant literature difficult for this project.

David Weinberger summarizes the problem of trying to wrap our metaphorical arms around so much digital information and how people access it, explaining that such complex systems are “database-based science” [1] (p. 128) that are so complicated we get little opportunity to understand them entirely, but we can at least know how these systems work without fully understanding them. This ties directly to why digital library designers seek to understand access behaviors and patterns in the digital library interface.

Zhu and Freeman employed content analysis theory to develop what they termed “a user interaction framework,” [2] (p. 1) to analyze how users engaged Open Government Data Portals (OGD). The user interaction framework necessitated the following components:

- Access: Data organization, searchability, restriction fee, license multiple languages, machine processability, open formats, and permanent URI
- Trust: Completeness, currentness, availability of a data policy, and granularity of data relevancy
- Understand: User support, application showcase, documentation, and metadata
- Engage-integrate: Availability of analytics, availability of API, availability of citation format, personalization, download, online manipulation, online visualization, and comparative data sets
- Participate: Proactive engagement, shareability, participation, and user feedback [2] (p. 1)

This framework examined the above capabilities in reference to a single digital collection, and the UNT Digital Library had these access features for all collections and individual digital objects.

The UNT Libraries' Digital Collections interfaces has similarities to what the Europeana project has adopted. Europeana is based on a model of “cultural commons,” with the goal of employing linked data to break out of silos and allow users to create their own research context, based on a model of

sharing and searching across varied digital contributions from European countries [3] (p. 67). This type of design, with large sets of diverse materials searchable across different contributors and collections, becomes a complex system of navigation interactions, contextualized by the users rather than by the interface technology.

Blumer, Hügi, and Schneider examined the impact of faceted navigation on users, in direct, full, usability studies [4]. Their findings about the usefulness and recommendations about how to tailor facets informed how we examined our own log data to understand user interactions across collections because user interactions at the UNT Digital Library are heavily influenced by the faceted navigation built into the interface.

Meera, Manjunath, and Kaddipujar employed faceted navigation toward connecting digitized archival objects in the Raman Research Institution in Bangalore, India [5]. Their primary goals centered around promoting their physical and digital collections, represented by a variety of different media and subject types, digitized across sixteen communities [5] (p. 311). The repository designers implemented a Dreamweaver-designed linked list of specific fields to connect objects across a DSpace repository, basing this type of list on the model of scientific research databases to enhance object connectivity and access to multiple resources. These goals, though on a much smaller scale, were not unlike the goals of facets employed in the UNT Digital Library.

3. Essential Background

The University of North Texas Libraries' Digital Collections is not hosted on an existing software platform (e.g., DSpace/ContentDM). Rather, these collections are hosted using a set of locally developed software tools that provide digital preservation services and content delivery services to users. These two systems respectively are named Coda, and Aubrey. The Coda repository provides digital object auditing, replication, and fixity services and Aubrey provides search, retrieval, and metadata editing interfaces. The ability to locally develop interfaces for users that take advantage of common design patterns found in other digital library platforms, as well as being able to break from old patterns as needed, has allowed UNT Libraries to develop new interfaces and new ways for users to interact with digital resources.

In addition to a locally developed system architecture, the digital resources are described using a unified metadata element set called UNTL. The element set extends the standard 15 Dublin Core metadata elements with elements such as: *Degree*, *citation*, *collection*, *partner*, *note*, *primarySource*, and a field for storing information about the metadata records called *meta*. Together these metadata elements are used to describe all resources held in the UNT Libraries Digital Collections. Like many other information retrieval systems, metadata values are linked so that users can easily navigate from one resource to another. This system employs a metadata field *relation* to note when items are related, such as when a presentation has a corresponding script or handout. While these items have their own record, we ensure that they are closely linked together. Items from a given conference or other series also share a common series or serial title field to acknowledge this relationship. Implementing metadata fields in this way, in conjunction with a unified full-text search interface, increases the likelihood of a user crossing over collections to locate items immediately relevant to the current object they are viewing.

While a collection provides grouping functionality and additional context to resource, UNT has also implemented another metadata field for grouping resources, called "partner." The partner designation recognizes the contributing source from which the object originates. The partner is typically the organization for which UNT digitized and hosts materials, or the organization that holds ownership of the materials. This varies depending upon whether the organization is external or internal to UNT. For institutional repository items, the partner corresponds to the UNT College with which the author is affiliated. For cultural heritage materials from an external source, the partner may correspond to the external institution holding those items, such as the Barbara C. Jordan Archives at Texas Southern University.

4. Methods

We employed a quantitative research model and gathered user logs and statistics to examine a total of 1,379,439,042 lines of Apache access logs from 2017. From these access logs, we defined a session as all interactions that returned item content within a 30 min window by a single IP address. Sessions lasting longer than 30 min were divided into multiple sessions at each thirty-minute mark. We also removed lines originating from known bots or crawlers. We wrote custom Python scripts to parse, filter, and group entries in the log files into user session. This resulted in 10,427,111 user sessions, where at least one item was accessed. After extracting user sessions, we aggregated metadata records for each item in the session. From each item's metadata, we could see what collections, resource types, and partners they were associated with.

To examine this data, we posed three questions about how users work between digital objects and collections:

- Q1. To what extent do users discover and use items from different collections?
- Q2. To what extent do users discover and use items that are different types?
- Q3. To what extent do users discover and use items that are from different contributing partners?

To begin an investigation of usage across collections, we decided to examine how usage data demonstrated access movement from the UNT Scholarly Works Repository collection to and between other objects and collections. Using this one collection as a comparison point allowed us to see how its items may have come into contact with others and to better understand how its items may have supplemented the understanding of a topic, based on the search result sets. At the University of North Texas, content traditionally associated with an institutional repository is frequently distributed across multiple collections: The UNT Scholarly Works collection, which contains faculty and staff authored works; UNT Theses and Dissertations; UNT Undergraduate Student Works; UNT Graduate Student Works; and the UNT Data Repository. While these collections distinguish between the status of authors and types of resources, having them connected under the umbrella of the UNT Digital Library allows for users to gain access to any material that might fall under the purview of the institutional repository without having to immediately distinguish between the authors' status or type of object.

While we had item-based usage that could be aggregated at the collection or partner level, these statistics did not show how collections were being used together.

5. Results

5.1. Items per Session

To begin, we had to separate single-item interactions, where users accessed only one item and then moved away from the site entirely, from multi-item sessions. As our goal was to examine usage across the digital library, we found that 86% of these sessions only accessed a single item in total, without further usage occurring at all. This left 14% of the sessions interacting with multiple items, a total of 1,447,967 sessions, in which users may have potentially viewed items across collections, partners, and types. From this set, there were 19 sessions that utilized over 1000 items, which may have been the result of an unidentified script or harvester. Table 1 displays the descriptive statistics gathered on interactions within unique user sessions.

Table 1. Descriptive statistics on items accessed within a user session.

N	Min	Median	Max	Mean	Stdev
10,427,111	1	1	1828	1.53	4.735

5.2. Duration of Sessions

We defined a “session” as any interaction with item content from a single IP address within a 30 min window. We calculated the time from the first HTTP request in a session until the last HTTP request in that session. This allowed us to give each session a duration. We found that 82% of these sessions only lasted up to 59 s in length, as seen in Table 2. Of sessions lasting less than one minute, 69% had a duration of 0 s. These were typically sessions that used items through an embedded link or a pdf viewed directly from another site, such as Google, Twitter, or another webpage. These embedded or direct links did not generate a duration because there existed only a single HTTP request in the Apache access logs.

Table 2. Sessions with a duration of less than one minute.

Duration	Sessions	Percent of Sessions under 1 min
0 s	5,892,556	69%
1–9 s	1,476,112	17%
10–19 s	478,262	6%
20–29 s	257,916	3%
30–39 s	181,326	2%
40–49 s	140,492	2%
50–59 s	112,889	1%

To begin looking at connections within sessions, we focused on the UNT Scholarly Works Repository, which contained 5187 items by the end of the 2017 calendar year. We found 253,369 user sessions that accessed items from this collection. Of these, 88%, or 223,168 sessions, interacted with only a single item, leaving us with a set of 30,201 sessions that used more than one item for analysis. As the overall dataset for the UNT Digital Library showed automated forms of access with over 1000 items accessed in a session, we determined a specific range for items accessed that might reasonably indicate a human user. 95% of the sessions used between 2 and 11 items. This filtered the working dataset down to 28,638 user sessions that accessed between 2 and 11 items in total within a thirty-minute window from a single IP address, with one of those items belonging to the UNT Scholarly Works collection.

5.3. Cross-Partner Usage

Within the UNT Scholarly Works collection, the partner indicated the contributing college, department, or center on campus where the author of the resources was affiliated. Although a given document in the UNT Scholarly Works collection may have authors from multiple colleges, we typically listed the Partner as the College affiliated with the primary UNT author, as this field did not permit multiple partners to be listed within an item. This ensured a one-to-one relationship between the partner and the object, and any additional partners accessed within the session required the user to have visited another partner’s set of associated items. As Figure 1 illustrates, we found that 66.7% of the sessions that accessed 2 and 11 items also accessed items from more than one partner.

Most of the sessions viewed either one or two unique partners. However, 237 sessions used documents from five unique partners. For instance, someone using an article contributed by the College of Education may have also viewed an item from the College of Engineering in the same session.

5.4. Cross-Type Usage

As the institutional repository, the UNT Scholarly Works collection primarily contains articles, with presentations coming in second. Most of the scholarship produced by UNT faculty members takes the form of journal articles, book chapters, or monographs, but the collection also contains posters, artwork, papers, reports, and texts. These were represented as resource types, and each digital object was generally associated with a single type. Figure 2 illustrates the number of unique resource types

accessed within the user sessions accessing between 2 and 11 items. 76% of the sessions included items of different types.

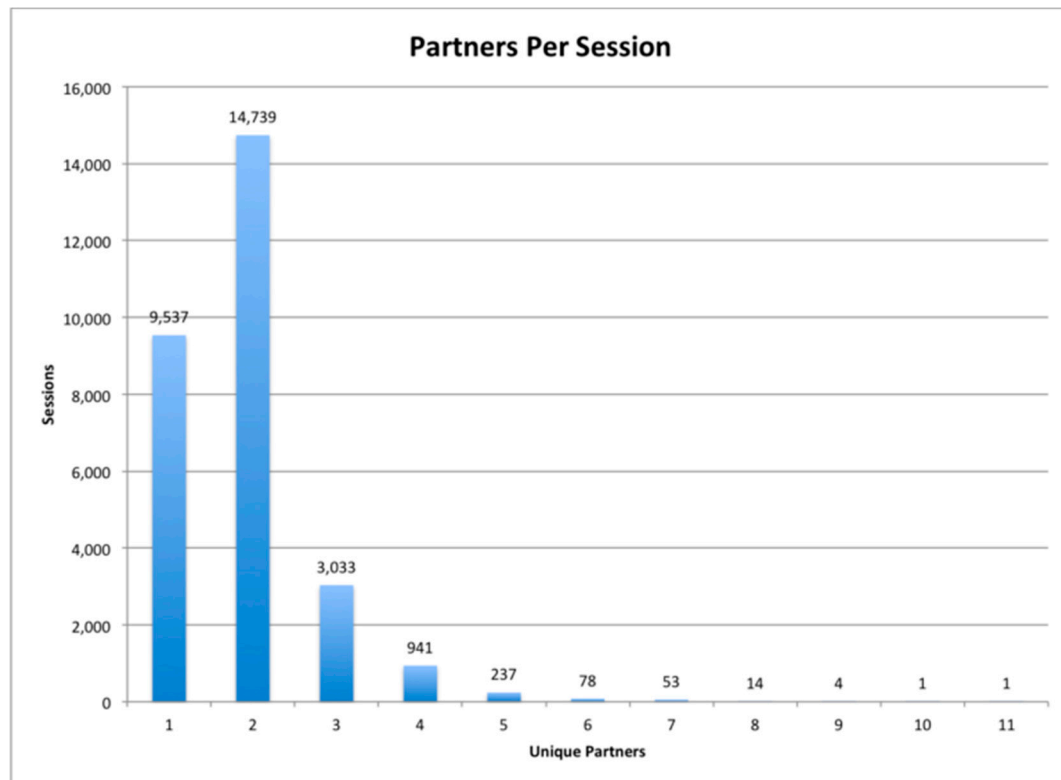


Figure 1. Total number of unique partners listed within a user session.

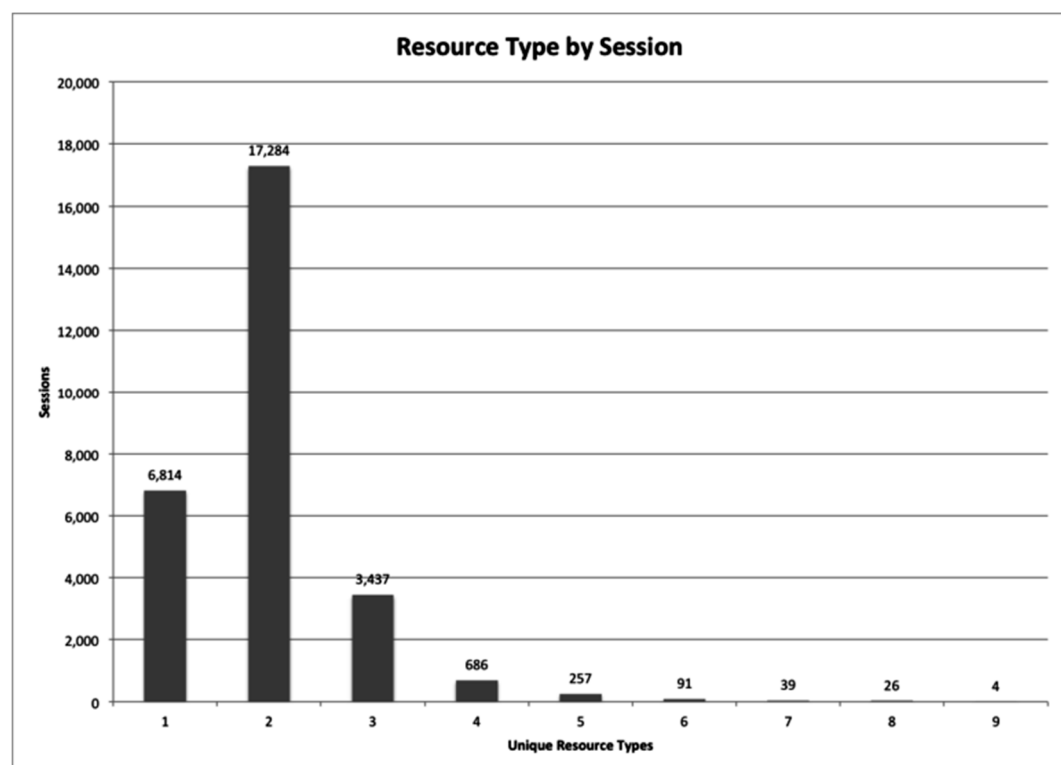


Figure 2. Total number of unique resource types listed within a user session.

The use of multiple resource types further pointed to the need to acknowledge that not all scholarship resembled a single object. There may have been a predominance of articles, but users were still accessing and finding value in the other forms of scholarship available.

5.5. Cross-Collection Usage

Despite the previous areas having a one-to-one relationship between the object and its associated partner or resource type, collections were a bit more complicated. As collections were housed within the larger interface, it was possible for a digital object to be associated with more than one collection. For instance, if one collection housed conference proceedings, and those proceedings contained a paper by a UNT faculty member, the paper would appear in both the conference proceedings collection and the UNT Scholarly Works collection. Figure 3 shows the number of unique collections listed within a user session who accessed between 2 and 11 items, with 75% of those sessions viewing items from two or more collection combinations.

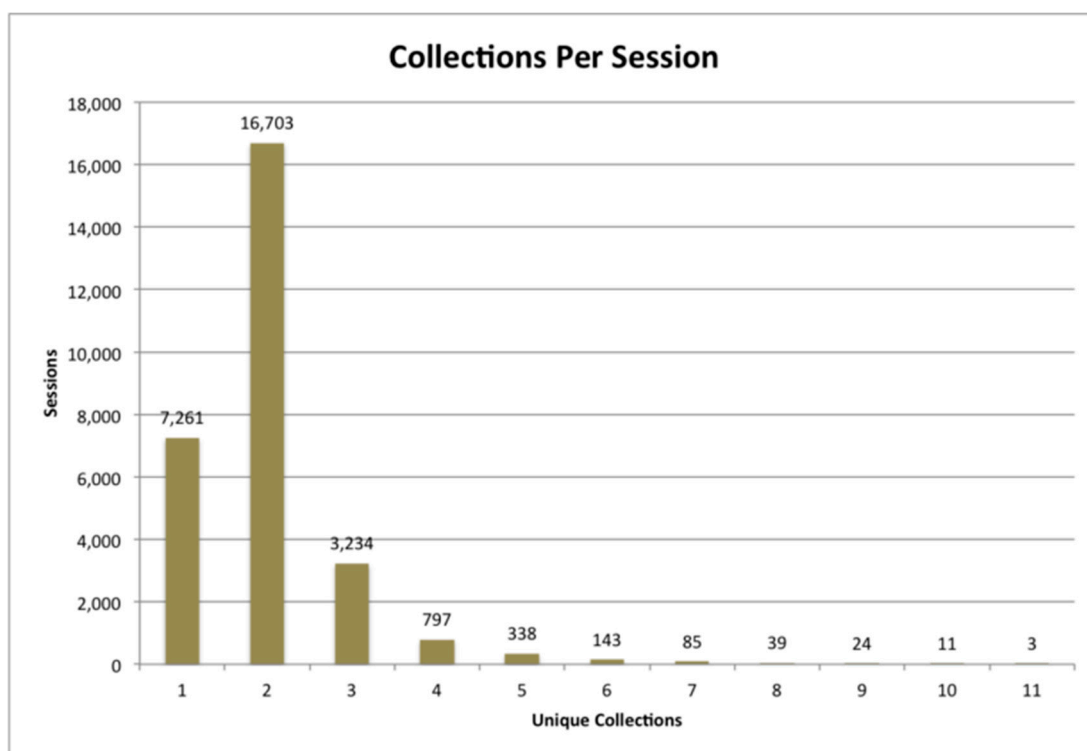


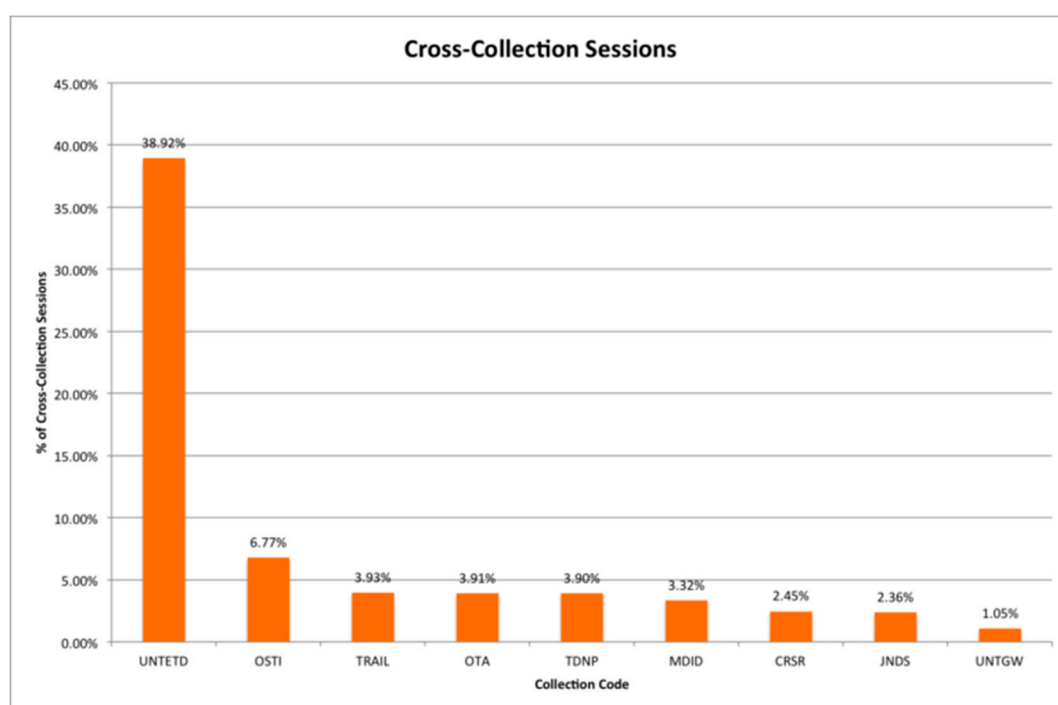
Figure 3. Total number of unique collections listed within a user session.

When finding items from collections, users could enter the interface through multiple venues. Users may have found items from different collections using the main search from either the UNT Digital Library or UNT Libraries web pages; Google may have landed them on a collection page; or a search using multiple tabs to run multiple searches within the system may have had search results collapsed within a single user session, as we had arranged the access logs for this study. Examining the combinations of collections accessed, we may have guessed a user's search narrative, or not, when they were using collections. Table 3 presents a paired matrix that shows which collections were visited together within a session. See Appendix A for Collection Codes and Names from the UNT Digital Collections.

Table 3. Paired matrix of collection combinations within a user session.

	UNTSW	UNTETD	OSTI	TRAIL	OTA	TDNP	MDID	CRSR	JNDS	UNTWG
UNTSW	0	11,147	1938	1126	1121	1118	952	703	676	302
UNTETD	11,147	0	323	258	895	80	230	175	59	63
OSTI	1938	323	0	165	9	48	3	91	28	8
TRAIL	1126	258	165	0	19	44	16	63	11	14
OTA	1121	895	9	19	0	2	60	15	5	4
TDNP	1118	80	48	44	2	0	0	17	4	12
MDID	952	230	3	16	60	0	0	3	2	1
CRSR	703	175	91	63	15	17	3	0	8	15
JNDS	676	59	28	11	5	4	2	8	0	0
UNTWG	302	63	8	14	4	12	1	15	0	0

This view of user sessions required that all sessions interacted with at least one item from the UNT Scholarly Works collection. This meant that in Table 3, the 165 user sessions that accessed items from the TRAIL and OSTI collections also viewed at least one item from the UNT Scholarly Works Collection. From this table, we could answer questions such as how often someone might use items from the UNTSW (UNT Scholarly Works), UNTETD (Theses and Dissertations), and TDNP (Texas Digital Newspaper Program) collections, which was 80 times. This analysis was more striking when looking at the percentage of sessions that combined these collections, as shown in Figure 4.

**Figure 4.** User sessions accessing items from UNT Scholarly Works and an item from another collection.

Of these collection combinations, 39% of these user sessions viewed items from the UNT Scholarly Works collection and the UNT Theses and Dissertations Collection. Although we might have known from anecdotal data that these collections were accessed frequently together, we now have better evidence from usage data to understand these connections.

6. Discussion

Although we did not track the search terms employed or goals of users, we could examine this usage data to see that users were indeed crossing into other groups of items, whether from a different partner, collection, or resource type. What we could gather from these crossings was that:

- Users viewing multiple items were finding items from more than one discipline: Crossover interactions between UNTSW and the other collections indicated access to multiple items within individual sessions.
- Users viewing multiple items were finding more than one type of item: The high level of crossover between Scholarly Works and the ETD collection demonstrated user behavior of commonly accessing multiple item types within single sessions.
- Users viewing multiple items were finding items from more than one collection: As exemplified by the interactions between the UNTSW, UNTETD, and TDNP interactions, users were accessing materials across collections and disciplines.

This data revealed four insights that will help UNT improve access to its digital collections:

- Development for Metadata Fields: This data showed how users made crossover connections, which helps UNT's Digital Libraries Division plan for more thoughtful metadata to plan metadata navigation to enable better collection and object crossover access.
- Informing User Experience: Interface improvement to the faceted navigation options was informed by this usage data to enable better user experiences with inter-collection, inter-partner, and varied object navigation.
- Outlier Overlap: The high overlap between use of UNT Scholarly Works and UNT Theses and Dissertations may point to students seeking work of faculty alongside the work of recent graduating laboratory members. This may also point to use of the repository as a tool to identify possible mentors, both from faculty and peer mentors among those recently graduated.
- Future Directions: As access to scholarship continues to grow through support, usability, and availability of different resource types, crossover connections may point to the need for increased preservation of gray literature, or more explicit instruction on how to turn a dissertation into an article.

7. Limitations

While this study offered quantitative data about usage interactions from one specific collection, UNT Scholarly Works, to other objects, its scope was deliberately narrow, to simply see usage and movement numbers. Eventually, we will pair this data with user feedback data, but due to the current field-wide gap in quantitative analysis of usage patterns, a need exists to address that gap prior to pairing the results into a wider research study.

Although the data-gathering method for this research was unique to the specific Scholarly Works Collection, the questions posed are replicable within other systems where researchers seek to observe usage behavior, starting with a controlled point and moving out to different object types through faceted navigation.

8. Conclusions

While this data represented crossovers solely from the UNT Scholarly Works Collection, we could employ this method as a way to examine how these connections occur from other collections, such as the Texas Digital Newspaper Program in The Portal to Texas History, which houses millions of pages of Texas newspapers, but for which we have only anecdotal evidence that suggests that users are interested in more resources beyond newspapers such as photographs and maps. Future research will help us see how users interact with these newspaper resources and whether they utilize resources from a wide range of titles. We could now implement this framework in a variety of situations in the future to better understand how users are interacting with collections.

Author Contributions: M.E.P. conducted data-gathering and prepared the Discussion section; P.A. outlined the initial draft and provided final formatting for this paper; A.K. prepared the Literature review, Limitations, and Conclusions sections and provided copy-editing.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Collection Codes and Names from the UNT Libraries Digital Collections

At UNT, collections were assigned with codes to organize items within the Digital Collections. Individual codes are unique across all access interfaces. A full listing of collection codes can be found here (<https://digital2.library.unt.edu/vocabularies/collections/>).

Code	Collection Name
UNTSW	UNT Scholarly Works
UNTETD	UNT Electronic Theses & Dissertations
OSTI	Office of Scientific & Technical Information Technical Reports
TRAIL	Technical Report Archive and Image Library
OTA	Office of Technology Assessment
TDNP	Texas Digital Newspaper Program
MDID	Marfa, Diversity in the Desert
CRSR	Congressional Research Service Reports
JNDS	Journal of Near-Death Studies
UNTDG	UNT Student Graduate Works

References

1. Weinberger, D. *Too Big to Know: Rethinking Knowledge Now That the Facts Aren't the Facts, Experts Are Everywhere, and the Smartest Person in the Room Is the Room*; Basic: New York, NY, USA, 2012.
2. Zhu, X.; Freeman, M.A. An evaluation of U.S. municipal open data portals: A user interaction framework. *J. Assoc. Inf. Sci. Technol.* **2019**, *70*, 27–37. [[CrossRef](#)]
3. Concordia, C.; Gradmann, S.; Siebinga, S. Not just another portal, not just another digital library: A portrait of Europeana as an application program interface. *IFLA J.* **2010**, *36*, 61–69. [[CrossRef](#)]
4. Blumer, E.; Hügi, J.; Schneider, R. The usability issues of faceted navigation in digital libraries. *Ital. J. Libr. Arch. Inf. Sci.* **2014**, *5*, 85–100. [[CrossRef](#)]
5. Meera, B.M.; Manjunath, M.; Kaddipujar, M. Facets of digital data dissemination: Value addition through “imprints collection”. *Libr. HI Tech.* **2013**, *31*, 308–322. [[CrossRef](#)]



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).