Metadata versus Full-Text: Tracking Users' Electronic Theses and Dissertations (ETDs) Seeking Behavior

Daniel Gelaw Alemneh and Mark Edward Phillips

University of North Texas, Denton TX 76201, USA

Daniel.Alemneh@unt.edu and Mark.Phillips@unt.edu

iConference 2018
25th-28th March 2018
Sheffield, UK
Organised by The University of Sheffield and Northumbria University
TRANSFORMING DIGITAL WORLDS
The University Of Sheffield.
iSchools
ischools.org
northumbria
UNIVERSITY NEWCASTLE

Outline

- Background
  - ETD at UNT
  - Usage Statistics
- Characteristics of ETDs
  - ETDs Usage
- ETDs Discovery via Metadata Vs. Full-Text
  - Methods, Analysis, and Findings
- Summary
  - Future works

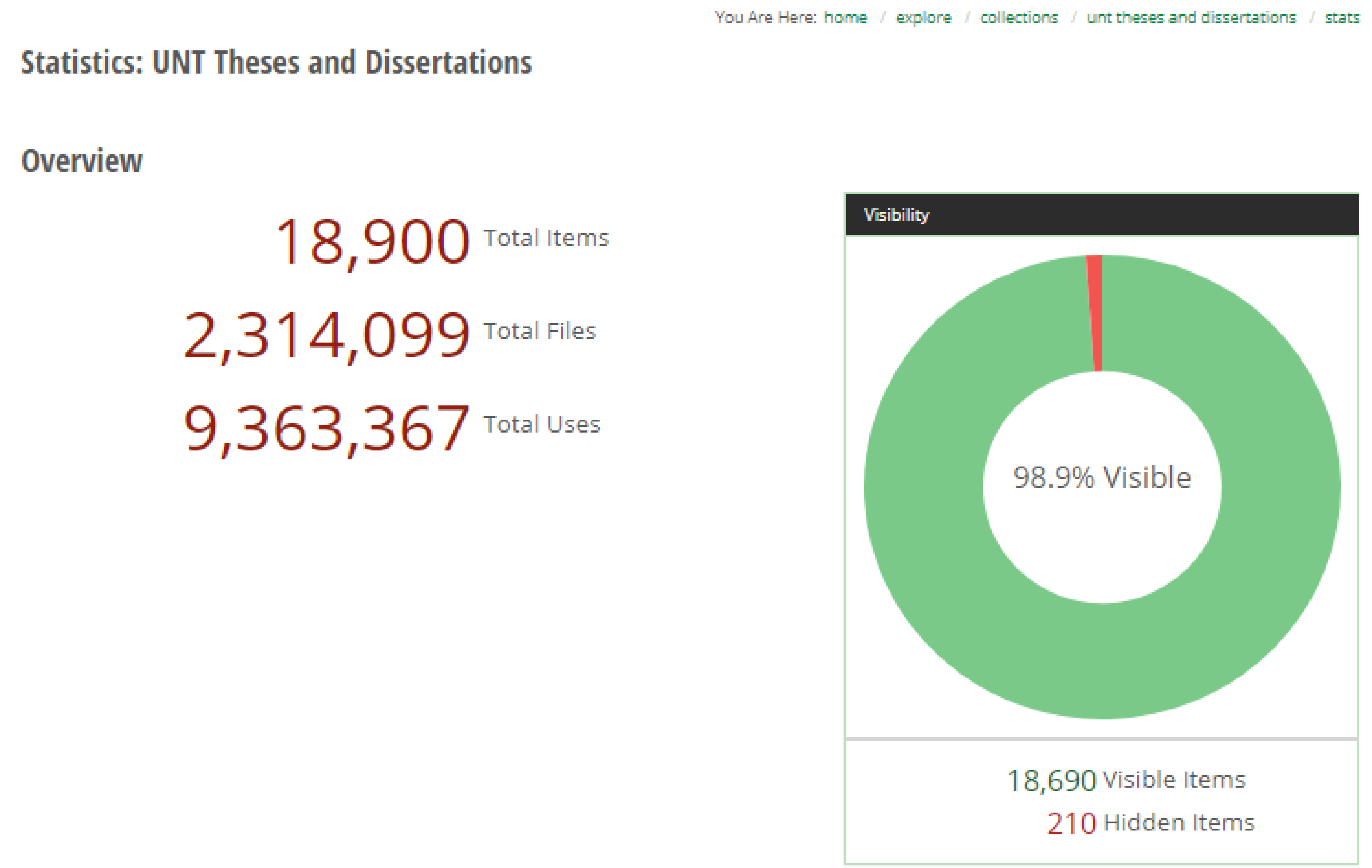
Background

ETD at UNT

- The University of North Texas (UNT) began accepting theses and dissertations in electronic format in 1999.
  - UNT is one of the early adopters of what was to become the ETD movement in higher education
  - One of the first three American universities to require ETDs for graduation.

UNT & ETD

- **At UNT, we have a combination of both born-digital ETDs and digitized analog these and dissertations.**
  - The ETD collection consists of more than 18,000 items and more than 2 mil-lion pages or files.
    - More than 12,000 analog theses and dissertations (from 1936 to 1999) were digitized
    - More than 6,000 born digital ETDs after fall-1999.
- **The UNT Libraries play an active role in facilitating access to UNT's ETDs**
  - Integrate Value added services into the ETDs
    - Multiple formats (PDF, JPG,)
    - Integrate Related contents (Datasets, videos, audios e.g. recitals)



You Are Here: home / explore / collections / unt theses and dissertations / stats

Statistics: UNT Theses and Dissertations

Overview

18,900 Total Items

2,314,099 Total Files

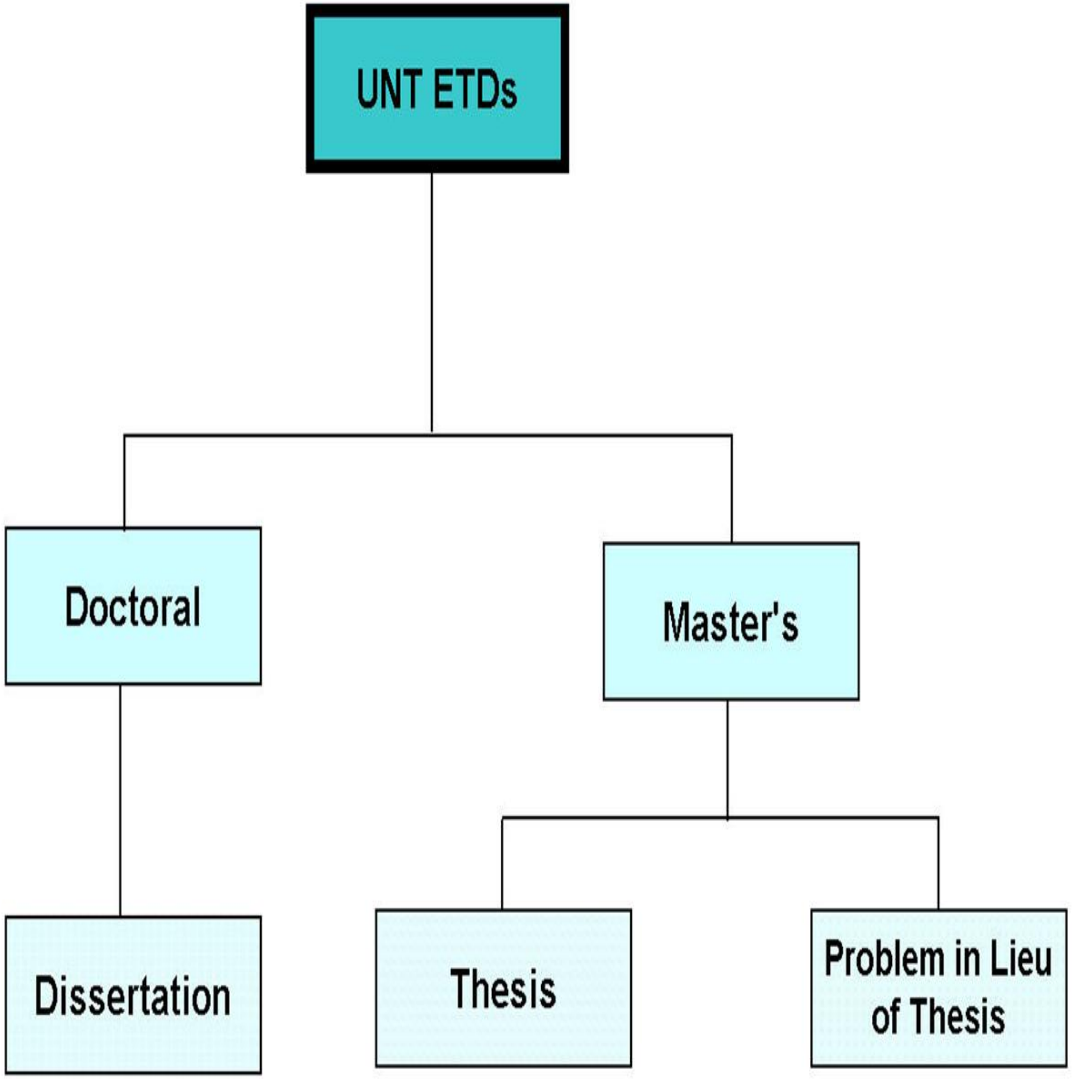
9,363,367 Total Uses

Visibility

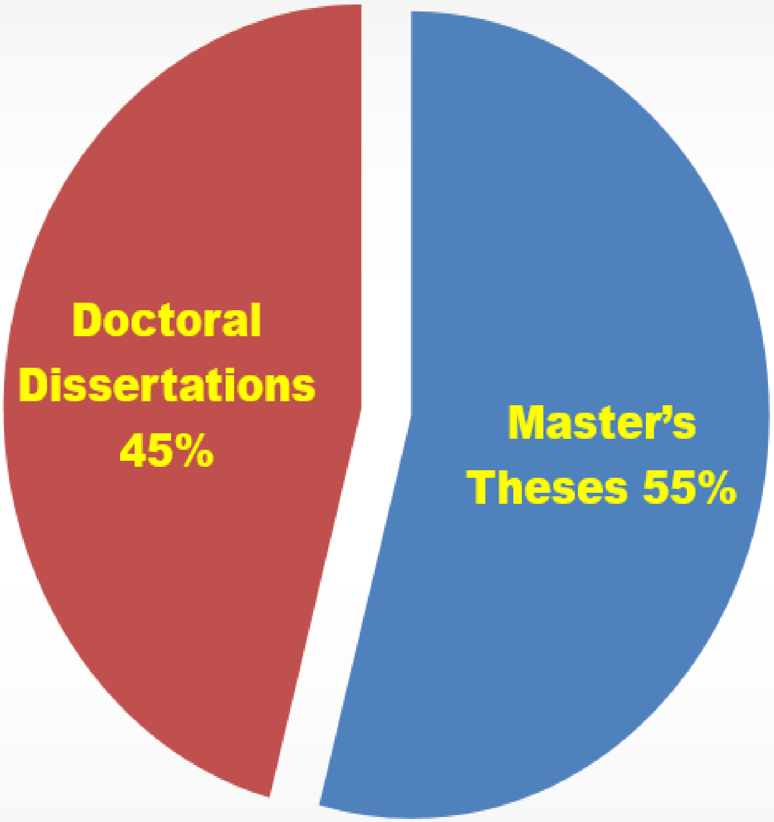
98.9% Visible

18,690 Visible Items

210 Hidden Items



UNT ETDs Size
As of March 2018



ETDs Usage

A green light to greatness.

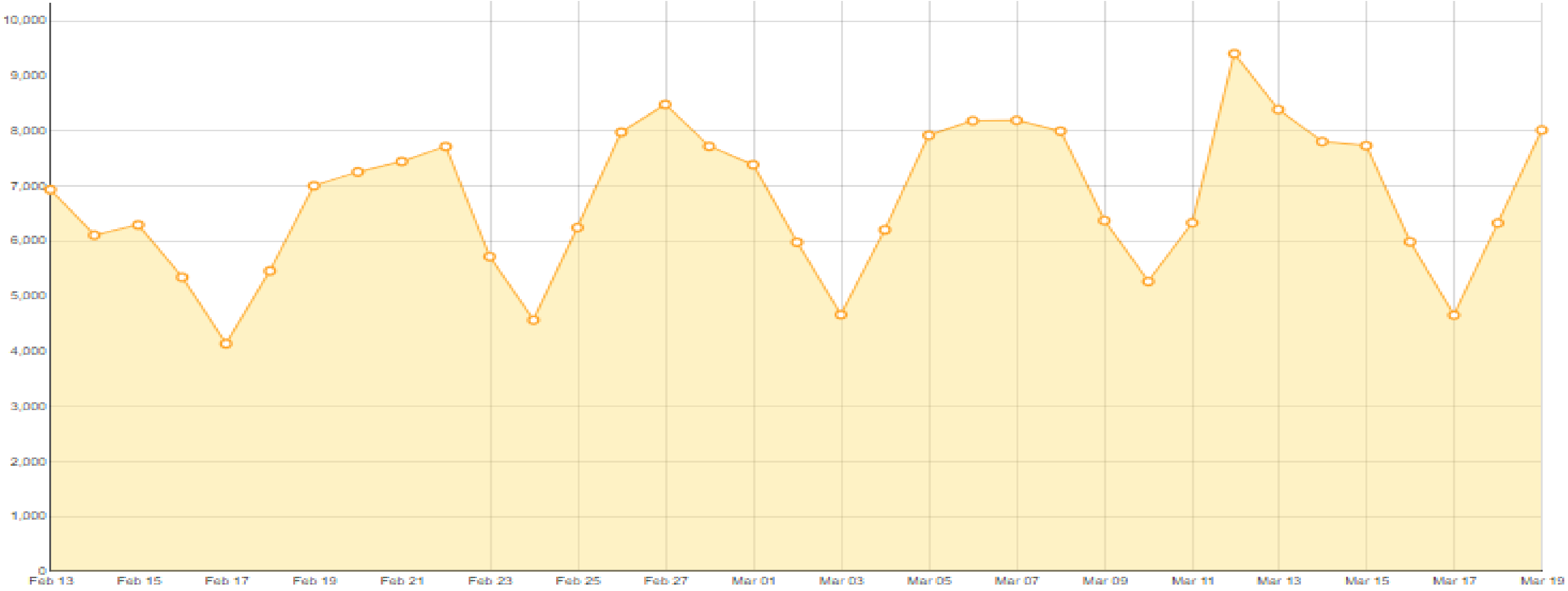
UNT

ETDs Usage Statistics

- UNT tracks the use of its ETD, as well as the use of other digital collections hosted by UNT Libraries.

- ETD Collection is heavily used, with daily use level ranging between 2,000 and over 6,000 uses from 200+ countries.

Usage



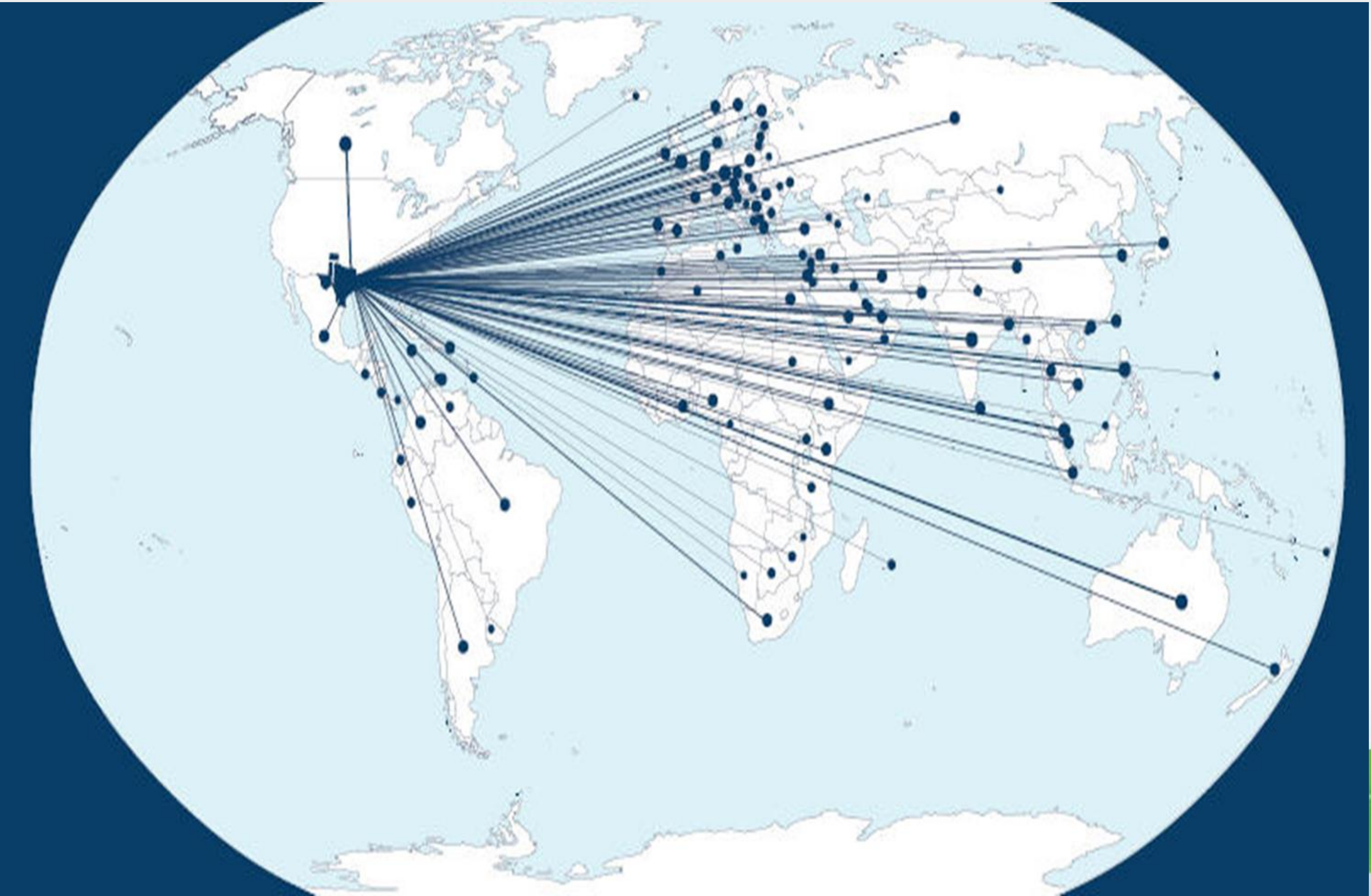
Recent usage

Usage by Month/Year

Year	January	February	March	April	May	June	July	August	September	October	Novem
2018	151,193	175,866	132,644								
2017	128,437	118,873	151,646	150,332	146,061	108,400	101,730	101,445	142,058	162,700	165,889
2016	138,837	164,215	176,924	167,523	141,502	130,470	105,206	119,758	154,088	190,269	159,359
2015	134,963	116,623	129,656	148,227	134,242	103,647	101,629	102,067	134,457	158,863	166,169
2014	91,109	97,216	108,147	115,545	107,376	100,094	256,453	104,855	100,832	112,120	112,932
2013	79,828	85,073	91,746	93,782	91,401	84,386	110,618	128,683	127,176	111,411	112,834
2012	52,627	64,297	62,338	62,427	58,394	48,135	50,941	53,428	60,254	80,274	80,226
2011	13,374	15,350	26,757	35,767	45,195	44,404	36,092	29,477	37,370	45,170	47,258
2010	9,641	18,573	23,620	24,678	17,369	13,500	11,500	10,598	14,524	13,330	14,216

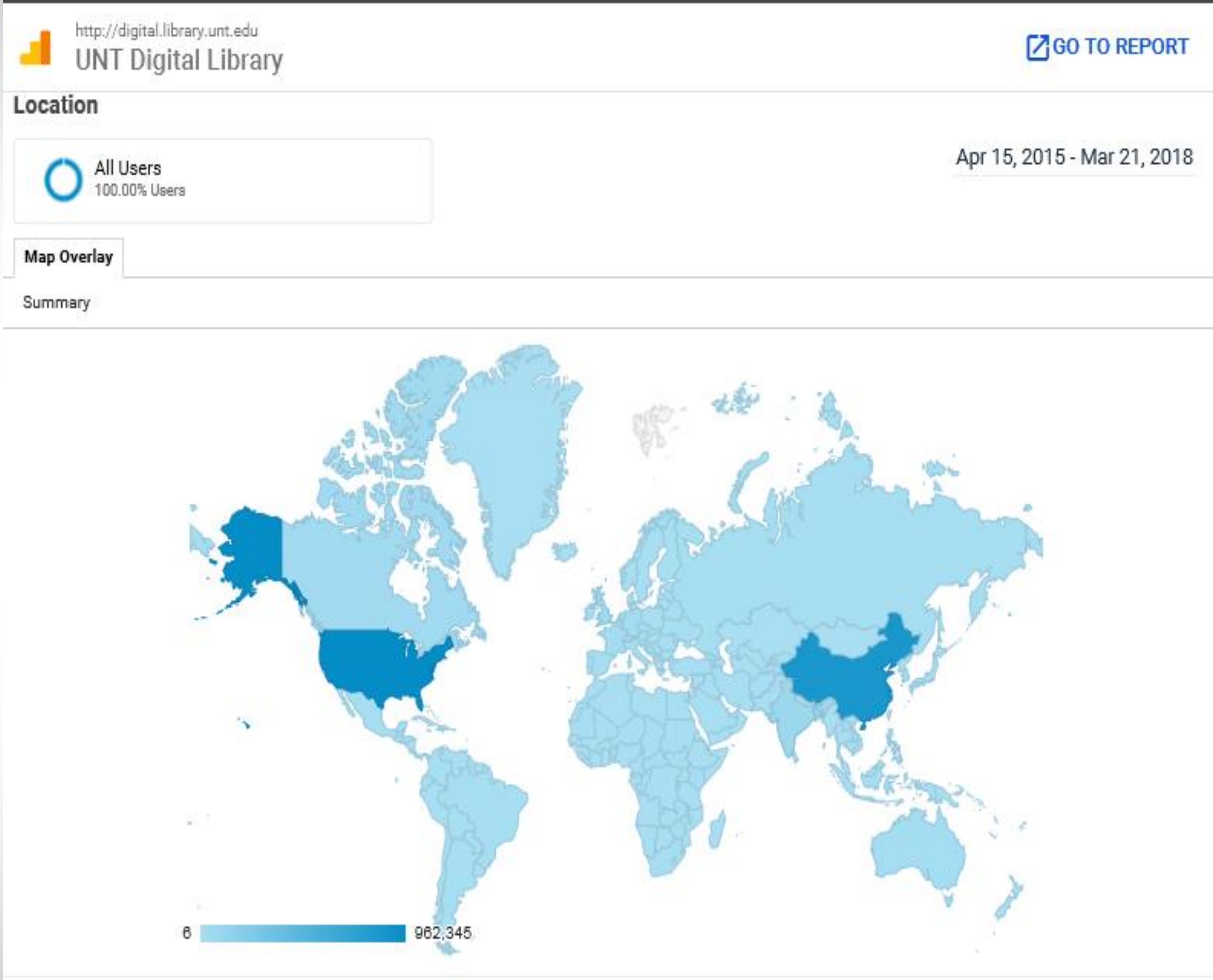
Visits from 215 Countries

http://digital.library.unt.edu/explore/collections/UNTETD/browse/



Top Countries used the UNT ETDs Collection

(April 2015 - March 2018)



http://digital.library.unt.edu

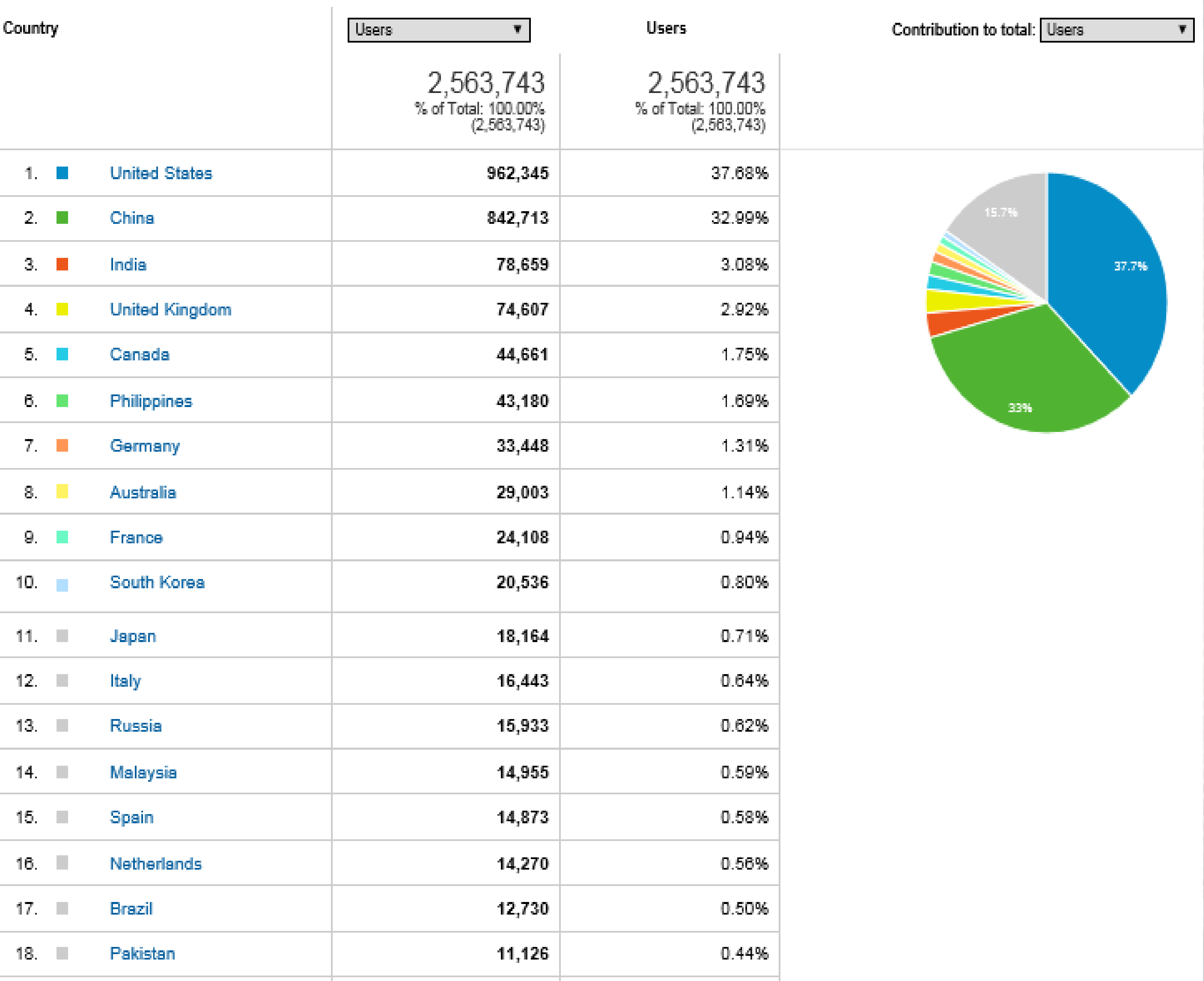
UNT Digital Library

Location

All Users
100.00% Users

Apr 15, 2015 - Mar 21, 2018

Top 18 Countries used the UNT ETDs Collection (April 2015 - March 2018)



Country | Users ▼ | Users | Contribution to total: Users ▼

		Country	Users	Users
			2,563,743 % of Total: 100.00% (2,563,743)	2,563,743 % of Total: 100.00% (2,563,743)
1.	🟦	United States	962,345	37.68%
2.	🟩	China	842,713	32.99%
3.	🟥	India	78,659	3.08%
4.	🟨	United Kingdom	74,607	2.92%
5.	🟦	Canada	44,661	1.75%
6.	🟩	Philippines	43,180	1.69%
7.	🟧	Germany	33,448	1.31%
8.	🟨	Australia	29,003	1.14%
9.	🟩	France	24,108	0.94%
10.	🟦	South Korea	20,536	0.80%
11.	⬜	Japan	18,164	0.71%
12.	⬜	Italy	16,443	0.64%
13.	⬜	Russia	15,933	0.62%
14.	⬜	Malaysia	14,955	0.59%
15.	⬜	Spain	14,873	0.58%
16.	⬜	Netherlands	14,270	0.56%
17.	⬜	Brazil	12,730	0.50%
18.	⬜	Pakistan	11,126	0.44%

Top Browsers and Operating Systems Accessed UNT's ETDs Collection (April 2015 - March 2018)

Browser		Users	New Users	Sessions	Bounce Rate	Pages / Session	Avg. Session Duration
Chrome		192,863	184,101	220,364	57.56%	2.52	0:02:08
Safari		64,930	62,742	73,431	63.97%	2.5	0:01:42
Firefox		41,733	39,474	45,728	58.92%	2.55	0:02:17
Internet Explorer		34,314	33,006	36,668	64.75%	2.13	0:01:41
Edge		13,911	13,507	15,920	58.70%	2.95	0:02:42
Opera Mini		6,183	6,135	6,718	69.38%	1.66	0:01:03
(not set)		4,887	4,887	2,687	98.25%	1	0:00:01
UC Browser		3,329	3,234	3,662	73.38%	1.53	0:00:58
Opera		2,402	2,318	2,949	60.09%	2.16	0:02:05
Android Browser		1,558	1,510	1,629	78.08%	1.34	0:01:10
Android Webview		1,367	1,355	1,546	68.43%	2.46	0:02:40
Safari (in-app)		1,284	1,248	1,391	60.68%	2.96	0:01:10
Samsung Internet		939	916	1,082	57.12%	2.22	0:01:46
Amazon Silk		713	701	761	71.88%	1.45	0:02:30

ETDs Discovery - Via Metadata Vs. Full-Text

A green light to greatness.

UNT

ETDs Case Study

- To get a better sense of users discovery of digital resources, we decided to assesse and see:
  - Whether users were arriving at our digital resources from searches that were answered by an items descriptive metadata or by parts of the full-text of the item.

- This study analyzed access to UNTs ETD Collection from two sides:
  - Searches that were answered by an items descriptive metadata
  - Users request met by parts of the full-text of the item.

Methodology

- For use analysis, we used Web server logs from the application server that provides access to the UNT Digital Library.
    - Before we extract the specific requests for ETDs in the UNT Digital Library, the data was obtained from a server log that contained 172 Million lines of requests
    - The log files were limited to discoveries of items in the UNT ETDs collection

- At the time this research was conducted, (Occurred between May 4, 2014 and January 24, 2015):
    - There were 11,873 unique ETDs available with metadata

Methodology

- The original raw dataset contained 172,115,682 lines during that timeframe, in the standard Extended Log File Format.
  - The resulting (two-column) intermediary dataset contained 84,837 item-query pairs

- Further limitations removed:
  - • Requests made by known robots,
  - • Requests without known search queries,

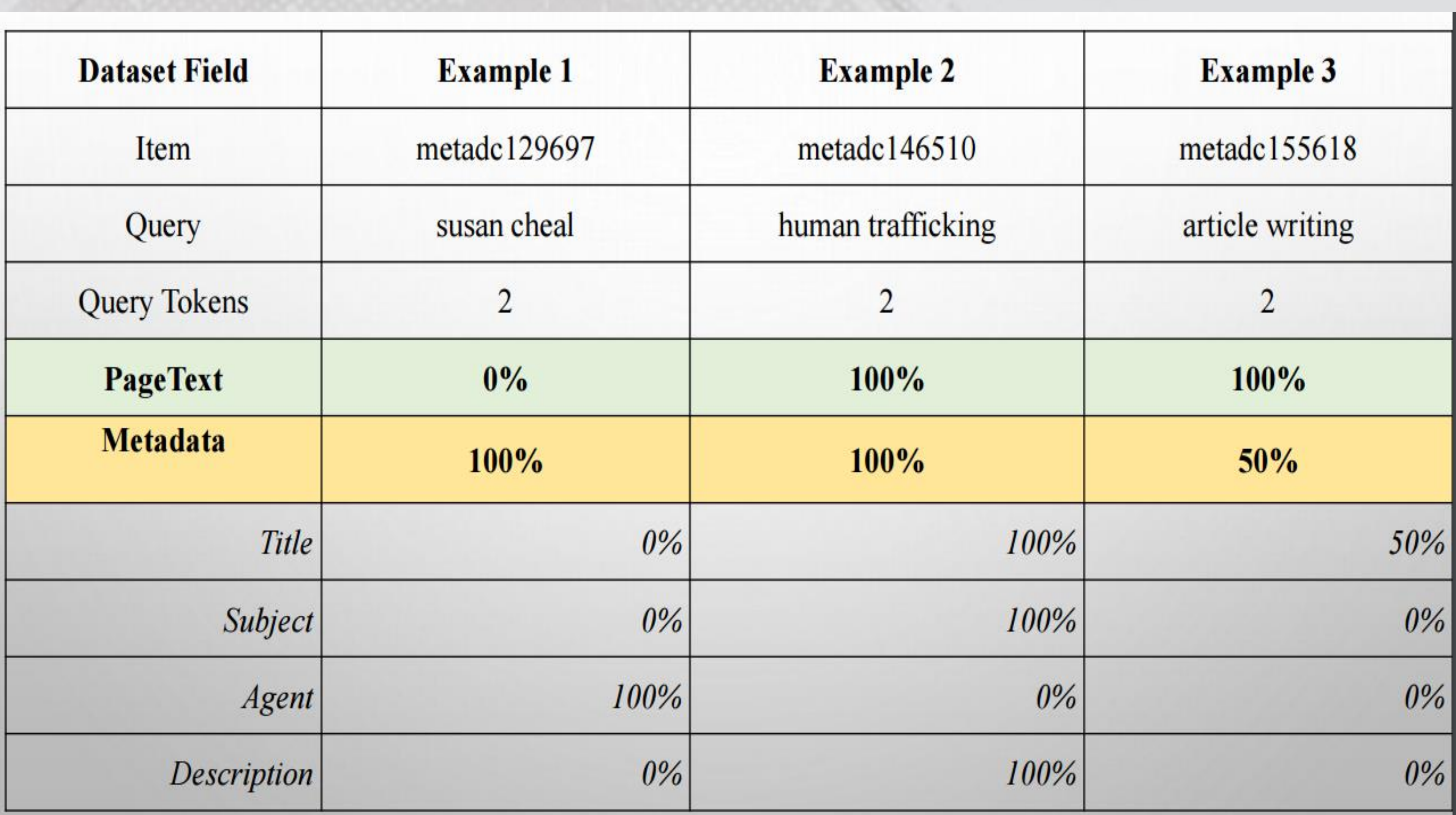
- Following further normalization, the dataset contained 43,420 unique query results;

Statistics for the Number of Tokens Per Query

N	Min.	Median	Max	Sum	Mean	Studdev
43420	1	2.5	31	104102	2.40	1.59

- **Queries varied in length and they were analyzed as individual words (or tokens) rather than phrases.**
  - This allowed for partial matches in a given field,
  - The distribution of tokens across queries ranged from 1 to 31 tokens

Example Dataset Entries for Three Search Queries

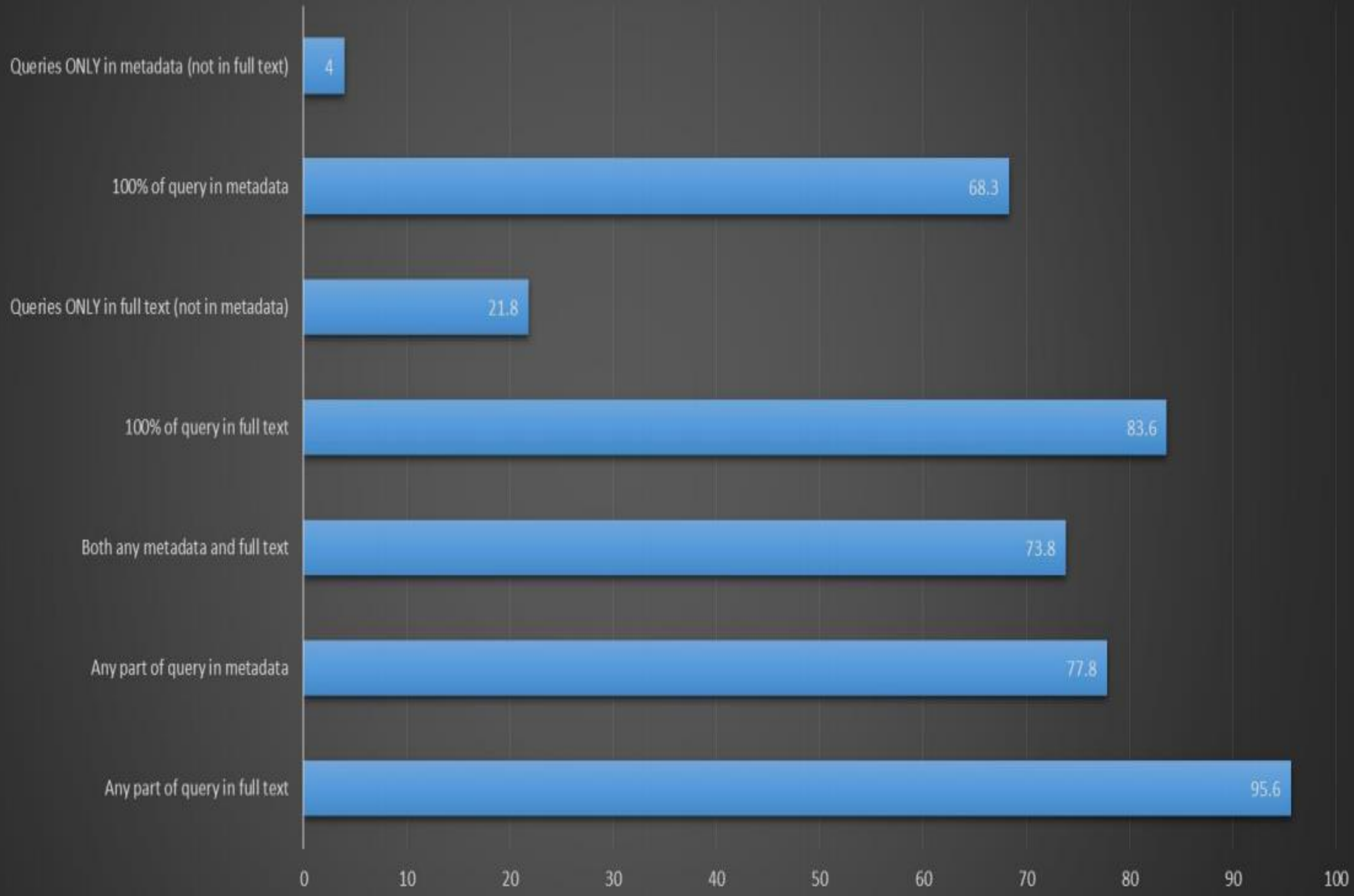


Dataset Field	Example 1	Example 2	Example 3
Item	metadc129697	metadc146510	metadc155618
Query	susan cheal	human trafficking	article writing
Query Tokens	2	2	2
PageText	**0%**	**100%**	**100%**
Metadata	**100%**	**100%**	**50%**
Title	*0%*	*100%*	*50%*
Subject	*0%*	*100%*	*0%*
Agent	*100%*	*0%*	*0%*
Description	*0%*	*100%*	*0%*

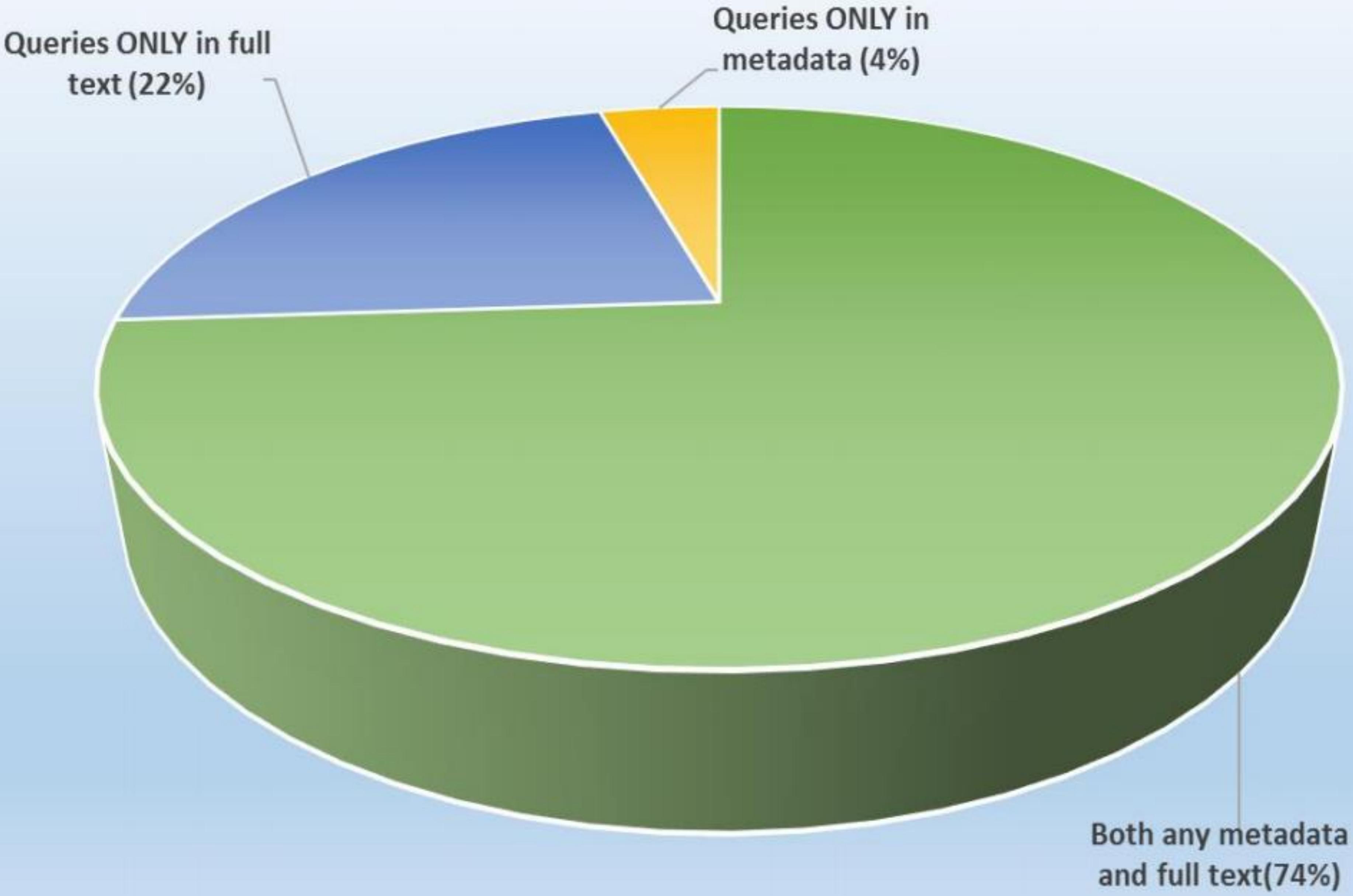
Record Discoveries on Matches in Metadata and Full-Text (N=43420)

Matches found in: Bottom of Form	Total No. of Queries Found: %	%
Any part of query in full text	41,519	95.6%
Any part of query in metadata	33,779	77.8%
Both any metadata and full text	32,056	73.8%
100% of query in full text (all tokens)	36,318	83.6%
Queries ONLY in full text (but not in metadata)	9463	21.8%
100% of query in metadata (all tokens)	29661	68.3%
Queries ONLY in metadata (but not in full text)	1723	4.0%

Percentage of Queries Found:

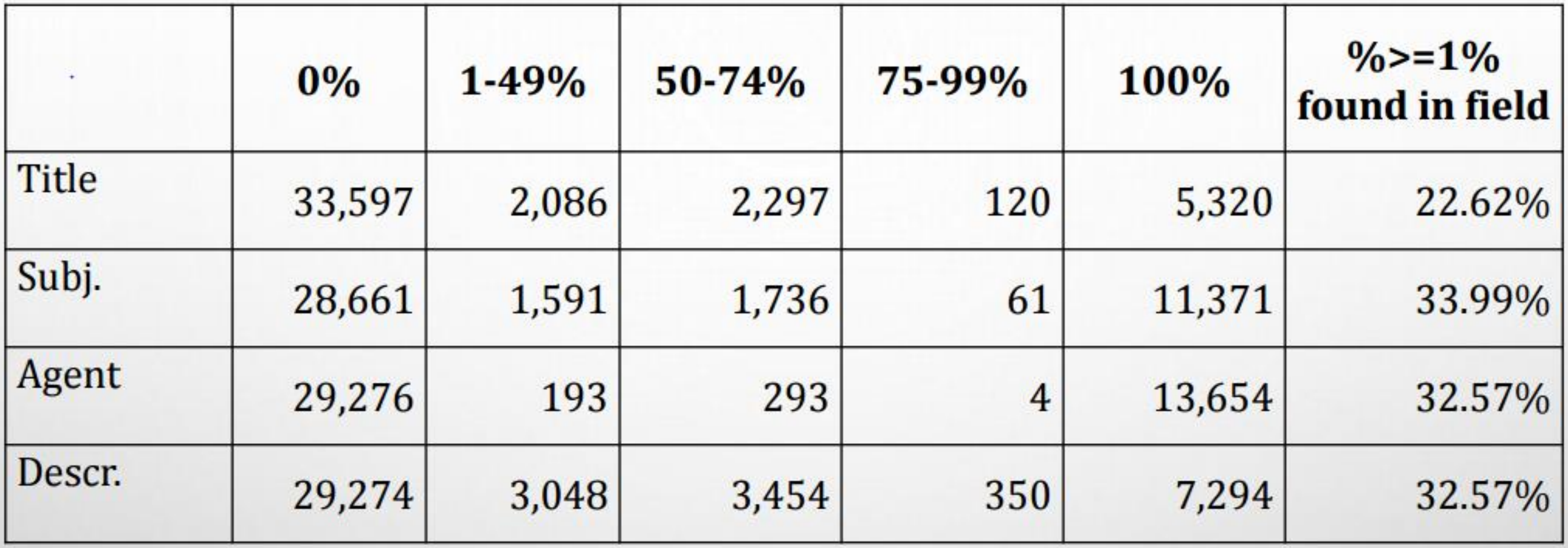


Percentage of Matches Queries Found:



RECORD DISCOVERIES PER FIELD BASED ON PERCENTAGE OF QUERY PRESENT IN FIELD. (N=43420)

At a more granular level, the following table shows record discoveries broken down by match percentages of each field, for the entire dataset. This shows the extent of the matches (partially for longer query strings) and the overlap across multiple fields.



	0%	1-49%	50-74%	75-99%	100%	%>=1% found in field
Title	33,597	2,086	2,297	120	5,320	22.62%
Subj.	28,661	1,591	1,736	61	11,371	33.99%
Agent	29,276	193	293	4	13,654	32.57%
Descr.	29,274	3,048	3,454	350	7,294	32.57%

Looking Ahead

Future Works

- Effective metadata and taxonomies add value and amplify the mostly interdisciplinary ETDs– allowing users to explore and delve deeper in multidimensional ways.

- The URLs referenced in a large corpus of ETDs may be present interesting insight into the subjects, disciplines and patterns in the ETD documents which warrants further investigation.

- Additionally an investigation into how specific disciplines or subject areas are referencing URLs in their ETDs would be helpful in identifying particularly high areas of URL linking versus lower levels.

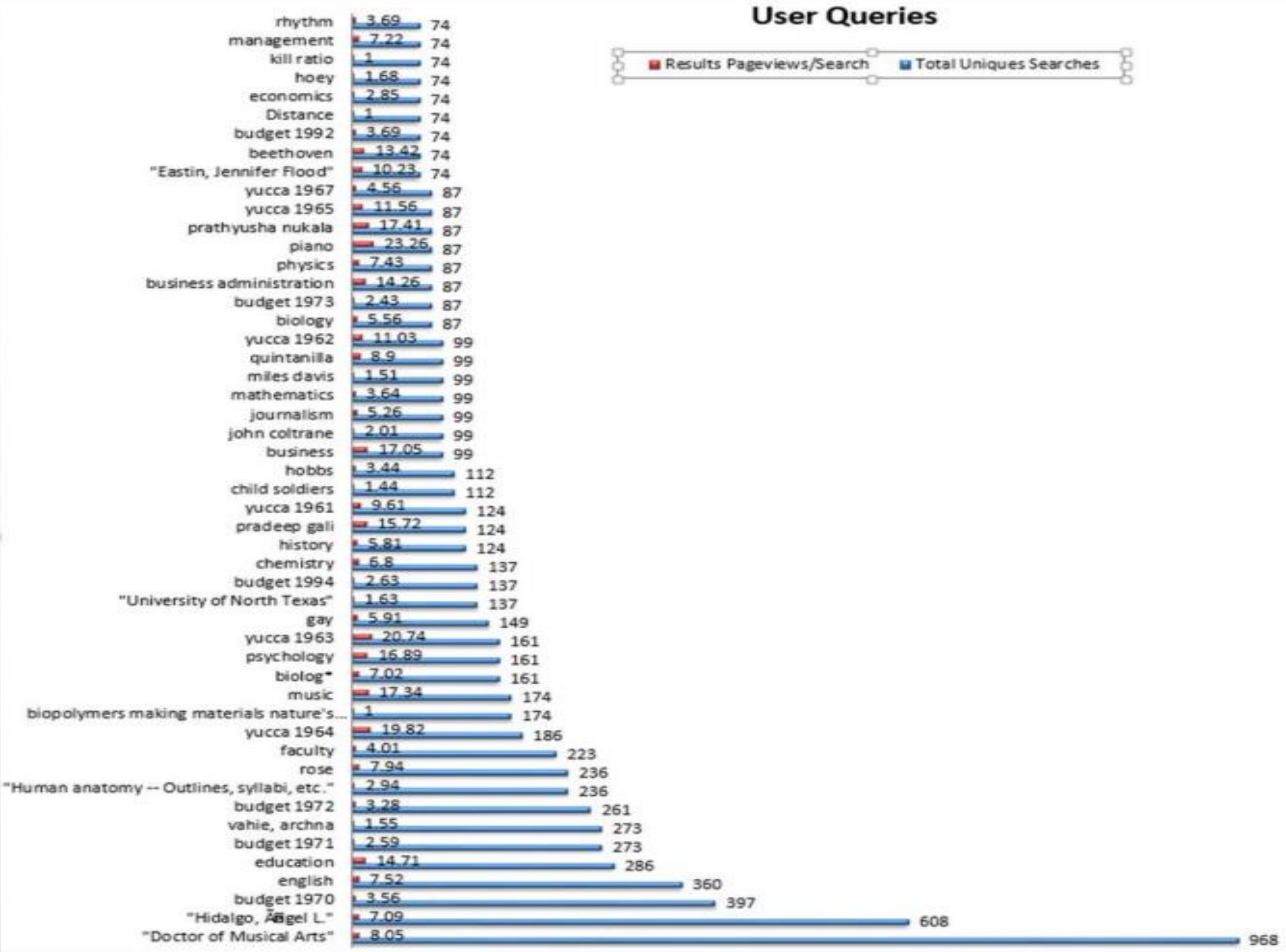
- An analysis of URL inclusion in ETDs across institutions would make a logical follow-on investigation that would show if higher level trends exist in ETDs.

Preliminary Analysis of Longitudinal Data For ETDs with URLs

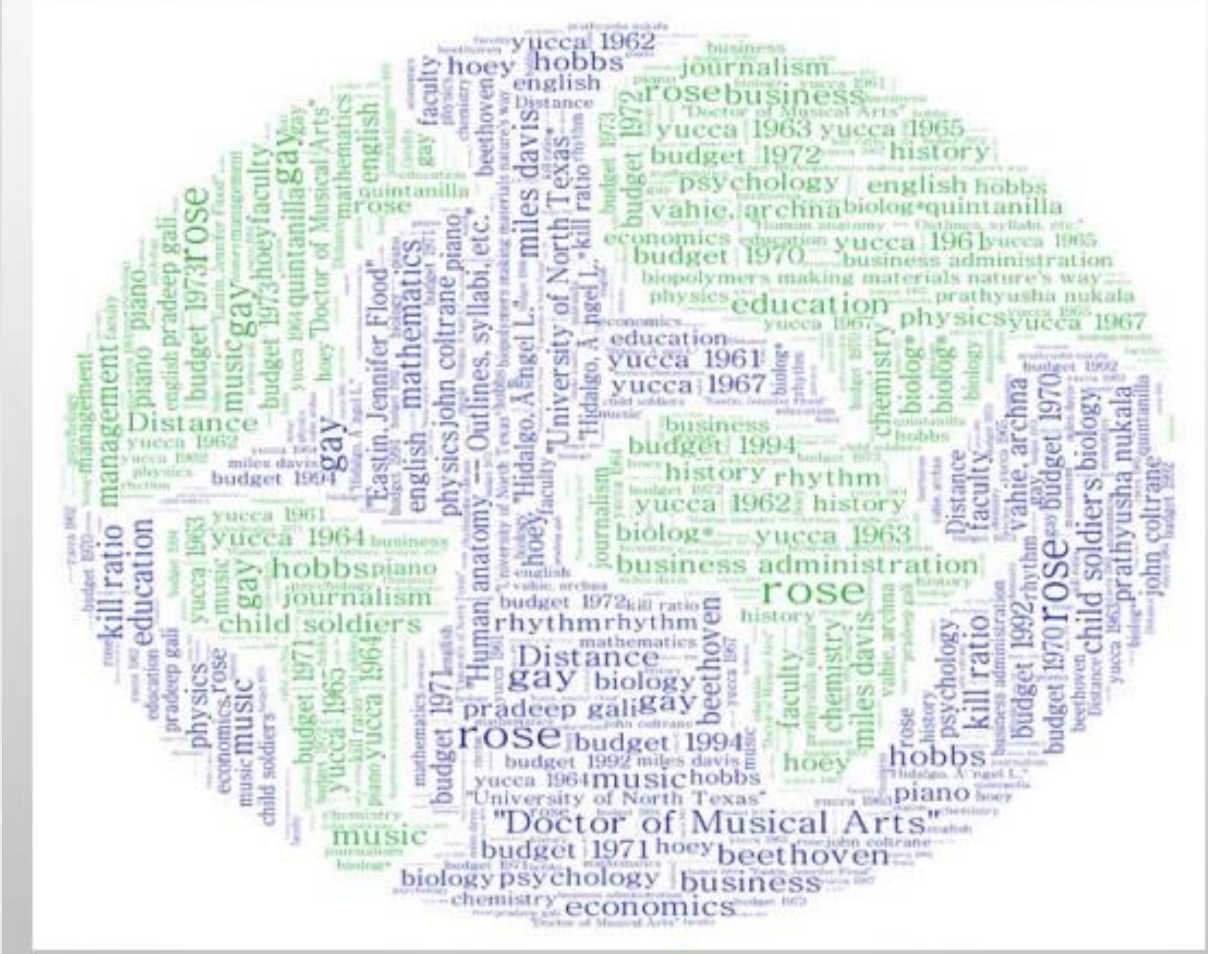
Year	Total No. of ETDs	No. of ETDs with URLs	% of ETDs with URLs	.com		.org		.edu		.gov		.net	
				#	%	#	%	#	%	#	%	#	%
1999	120	42	35.00%	14	42.9%	9	32.1%	11	39.3%	4	14.3%	4	14.3%
2000	315	127	40.32%	66	52.0%	70	55.1%	53	41.7%	41	32.3%	19	15.0%
2001	290	116	40.00%	63	54.3%	67	57.8%	57	49.1%	39	33.6%	20	17.2%
2002	298	146	48.99%	102	69.9%	73	50.0%	62	42.5%	47	32.2%	18	12.3%
2003	328	198	60.37%	133	67.2%	96	48.5%	84	42.4%	70	35.4%	24	12.1%
2004	304	181	59.54%	122	67.4%	89	49.2%	84	46.4%	66	36.5%	23	12.7%
2005	284	199	70.07%	141	70.9%	112	56.3%	100	50.3%	83	41.7%	43	21.6%
2006	326	235	72.09%	155	66.0%	143	60.9%	116	49.4%	98	41.7%	40	17.0%
2007	349	258	73.93%	182	70.5%	157	60.9%	122	47.3%	116	45.0%	35	16.6%
2008	336	242	72.02%	166	68.6%	140	57.9%	99	41.0%	91	37.6%	40	16.5%
2009	311	132	42.44%*	87	65.9%	83	62.9%	69	52.3%	50	37.9%	25	19.0%
2010	366	286	78.14%	199	69.6%	170	59.4%	127	44.4%	129	45.1%	50	17.5%
2011	418	335	80.14%	231	69.0%	215	64.2%	134	40.0%	146	43.6%	66	20.0%
2012	290	230	79.31%	153	66.5%	155	67.4%	94	40.9%	104	45.2%	38	16.5%

Summary

- Considering the diverse global ETDs users’ communities, effective retrieval depends not only on the subject terms assigned to describe an item, but on the search query terms entered by users as well.


User Queries
Results Pageviews/Search
Total Uniques Searches
rhythm 3.69 74
management 7.22 74
kill ratio 1 74
hoey 1.68 74
economics 2.85 74
Distance 1 74
budget 1992 3.69 74
beethoven 13.42 74
"Eastin, Jennifer Flood" 10.23 74
yucca 1967 4.56 87
yucca 1965 11.56 87
prathyusha nukala 17.41 87
piano 23.26 87
physics 7.43 87
business administration 14.26 87
budget 1973 2.43 87
biology 5.56 87
yucca 1962 11.03 99
quintanilla 8.9 99
miles davis 1.51 99
mathematics 3.64 99
journalism 5.26 99
john coltrane 2.01 99
business 17.05 99
hobbs 3.44 112
child soldiers 1.44 112
yucca 1961 9.61 124
pradeep gali 15.72 124
history 5.81 124
chemistry 6.8 137
budget 1994 2.63 137
"University of North Texas" 1.63 137
gay 5.91 149
yucca 1963 20.74 161
psychology 16.89 161
biolog* 7.02 161
music 17.34 174
biopolymers making materials nature's... 1 174
yucca 1964 19.82 186
faculty 4.01 223
rose 7.94 236
"Human anatomy -- Outlines, syllabi, etc." 2.94 236
budget 1972 3.28 261
vahie, archna 1.55 273
budget 1971 2.59 273
education 14.71 286
english 7.52 360
budget 1970 3.56 397
"Hidalgo, Ãgel L." 7.09 608
"Doctor of Musical Arts" 8.05 968

QUESTIONS?



Daniel.Alemneh@unt.edu

Mark.Phillips@unt.edu