Measuring Genuine Use of Repository Content at Northeastern University

Sarah Sweeney, sj.sweeney@neu.edu
Northeastern University Libraries
repository.library.northeastern.edu

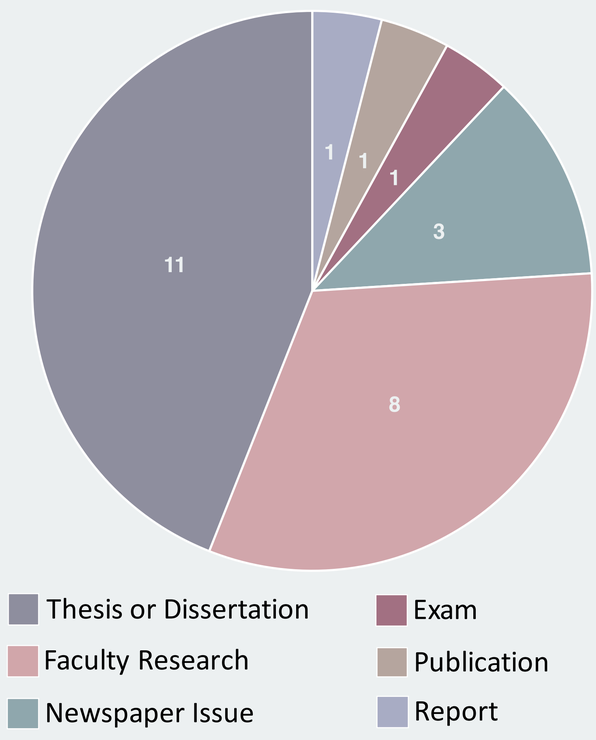
Usage Statistics in the Digital Repository Service

Repository usage statistics are utilized by content owners to measure the impact of repository materials and to measure the use of the repository as a whole. Given the value of these metrics, it is vital that we understand how repository statistics are gathered so we can sort genuine user interactions from automated traffic.

Early on in the Digital Repository Service development process we decided not to rely on Google Analytics to collect statistics. While Google Analytics provides valuable tracking, we cannot easily distinguish genuine user traffic from bots or crawlers. We decided to record and process our own statistics so we could isolate genuine use and ignore statistics generated by bots and other large automated consumers of our content.

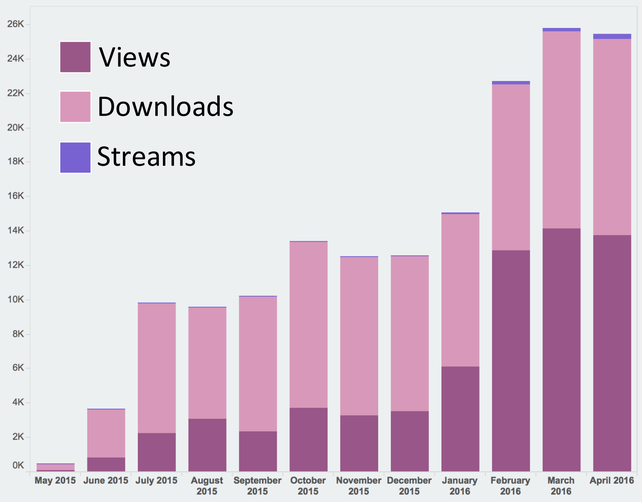
Popular DRS Genres

Top 25 viewed files by genre.



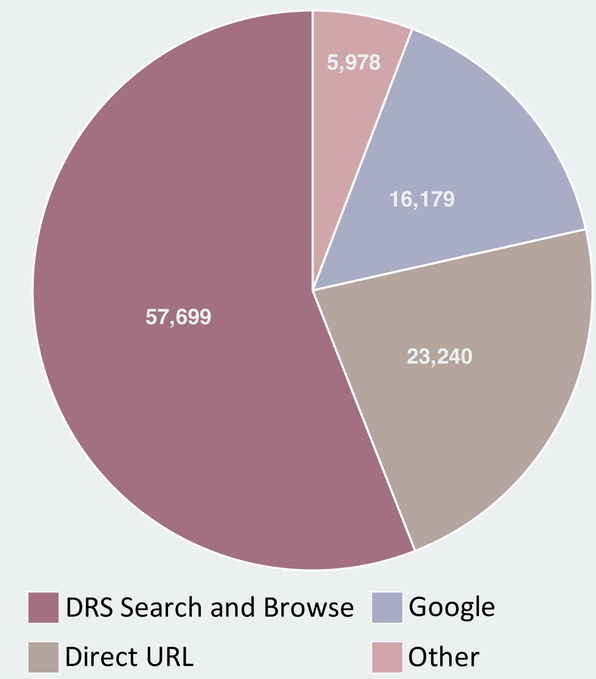
One Year of DRS Activity

Total views, downloads, and streams per month.



Top Five Referrers

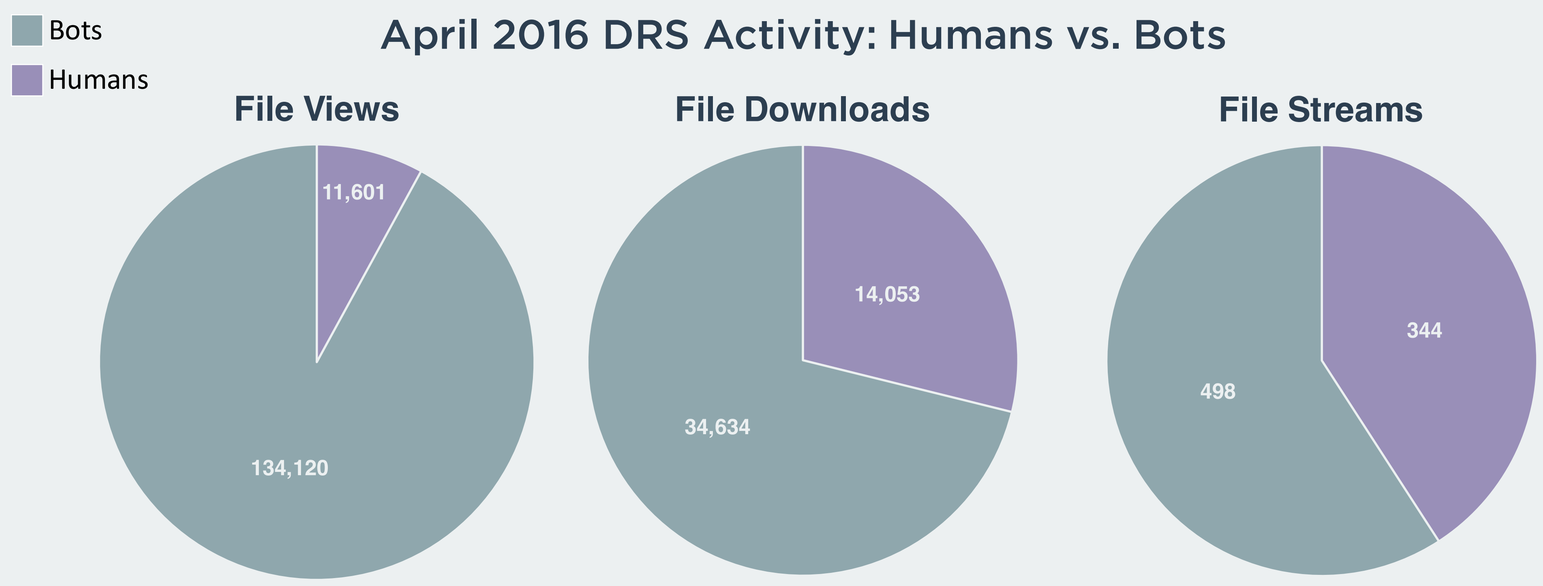
Top 5 referrers by number of page views



Not All Traffic is Equal

A seemingly endless number of bots exist to crawl publicly available repository content for harvesting and indexing. These bots help increase the discovery of repository content, but they can also greatly inflate usage statistics. Usage statistics are often gathered using third-party tools, like Google Analytics, which may or may not report their collection methods, and may not be aware of the difference between human and bot consumption. Although content owners tend to prefer higher numbers regardless of the consumer, we want to be able to defend the statistics we are gathering for our repository and declare them to be a genuine reflection of the use of our content by people, not bots.

April 2016 DRS Activity: Humans vs. Bots



Filtering

We designed a simple method for collecting and processing our usage statistics that allows us to filter out the non-human consumption of our content. Raw, unfiltered DRS usage statistics are stored in an impressions table in a SQL database. A nightly job processes this table by comparing the agent responsible for the impression against a list of keywords associated with known bots. Impressions made by agents that match any keyword are marked as FALSE and are filtered out of the statistics displayed to users in the interface.

The Impressions Table

Every file view, download, and stream is counted as an impression and recorded in the impressions table. Along with the type of impression, we also record the agent responsible for the impression, how the agent was referred to the file, the agent's IP address, and the date of the impression. Impression frequency is limited to one per file per IP address per hour, which helps reduce inflated numbers from repeated clicks or page refreshes.

ID	PID	session_id	action	ip_address	referrer	status	user_agent	public	created_at	updated_at	processed
184159	neu:190034	5f79a9934…	view	108.20.51…	direct	COMPLETE	Mozilla/5.0 (iPad; CPU OS…	TRUE	5/21/15 9:41	1/22/16 1:58	TRUE
184181	neu:190034	a745a871e…	view	129.10.107…	direct	COMPLETE	Mozilla/5.0 (Windows NT 6.1…	TRUE	5/21/15 12:32	1/22/16 1:58	TRUE
184182	neu:182058	a745a8715…	download	129.10.107…	direct	COMPLETE	Mozilla/5.0 (Windows NT 6.1…	TRUE	5/21/15 12:33	1/22/16 1:58	TRUE
184183	neu:1302	e5d85e81e…	view	129.10.106…	https://repository…	COMPLETE	Mozilla/5.0 (Windows NT 6.1…	TRUE	5/21/15 12:42	1/22/16 1:58	TRUE

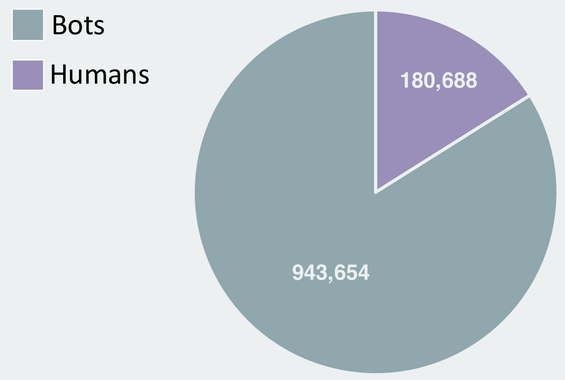
Significant Processing Values

- **"status"**: Files that are queued for download will be marked INCOMPLETE and are ignored until the download is finished.
- **"public"**: All impressions are initially set to TRUE. Once processed, impressions with agents on the bot list are marked as FALSE.
- **"processed"**: All impressions are initially set to FALSE. Once processed, the value is set to TRUE.

Ignored Agents

When the impressions table is processed, the agent responsible for each impression is compared against a list of common keywords associated with bots and crawlers, which is used to filter out agents from the impressions table. These keywords include: *archive*, *bot*, *crawl*, *curl*, *java*, *lynx*, *nutch*, *scrape*, *scrapy*, *slurp*, and *spider*

All DRS Impressions



Using the Data

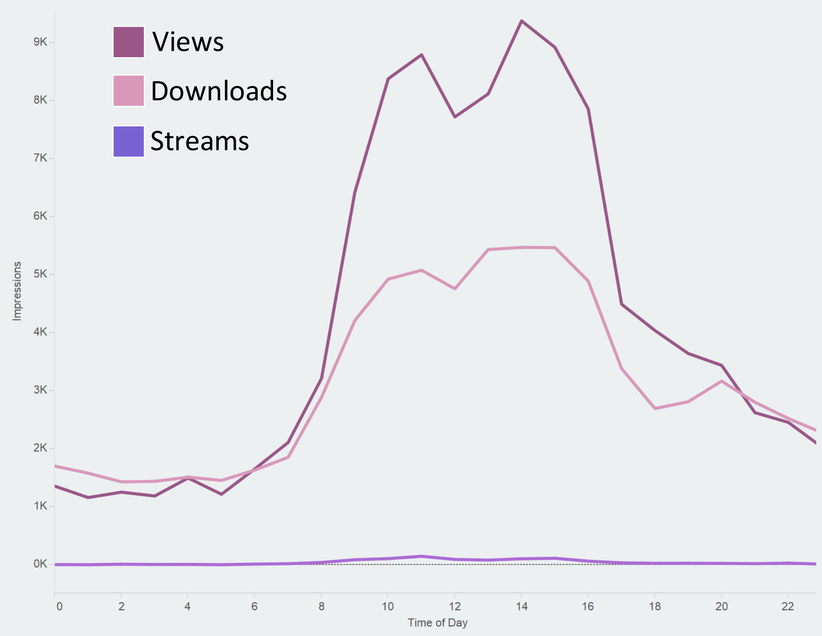
Categorizing our traffic and sharing our statistics gathering process has enabled us to more accurately track how the DRS is being used and more confidently defend our usage statistics as a reflection of genuine use of repository content. In 2016 we plan to improve our data gathering practices by:

- Inserting additional data points into the impressions table
- Operationalizing the process for adding new agents to the bot list
- Improving our statistical displays, including aggregated statistics and geographic visualizations

Improved Workflows

We also would like to use the data to improve our workflows. For example, the chart on the right can tell us the best time of day to schedule system deploys (between midnight and 6 am).

Measuring impact accurately is a difficult task, as is being able to confidently defend how measurements are recorded. Although there is important statistical value in recording automated bot traffic, measuring and reporting genuine repository use can improve our ability to communicate the true impact of repository content.



DRS *Digital Repository Service*
repository.library.northeastern.edu

Learn More

For more information about the DRS visit dsg.neu.edu/resources/drs
or github.com/NEU-Libraries/cerberus

Measuring Genuine Use of Repository Content at Northeastern University

Sarah Sweeney, sj.sweeney@neu.edu
Northeastern University Libraries
repository.library.northeastern.edu

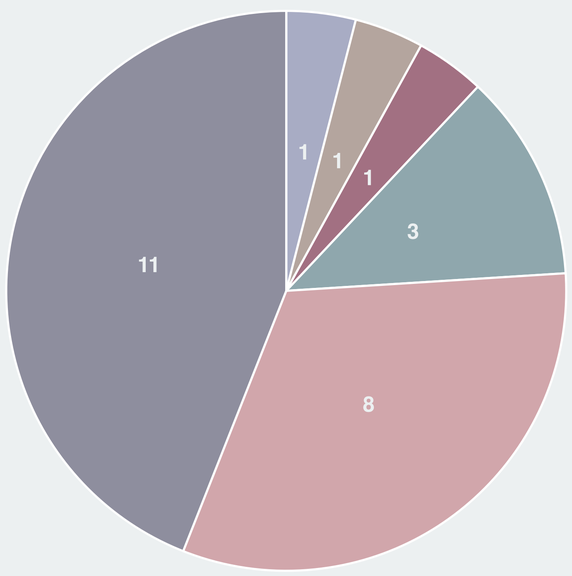
Usage Statistics in the Digital Repository Service

Repository usage statistics are utilized by content owners to measure the impact of repository materials and to measure the use of the repository as a whole. Given the value of these metrics, it is vital that we understand how repository statistics are gathered so we can sort genuine user interactions from automated traffic.

Early on in the Digital Repository Service development process we decided not to rely on Google Analytics to collect statistics. While Google Analytics provides valuable tracking, we cannot easily distinguish genuine user traffic from bots or crawlers. We decided to record and process our own statistics so we could isolate genuine use and ignore statistics generated by bots and other large automated consumers of our content.

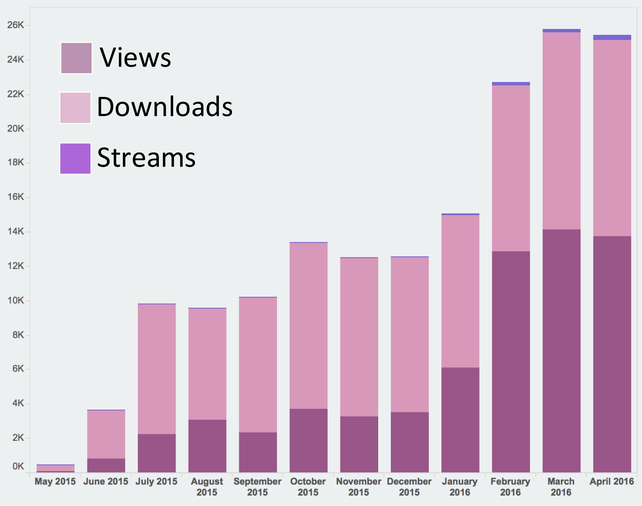
Popular DRS Genres

Top 25 viewed files by genre.



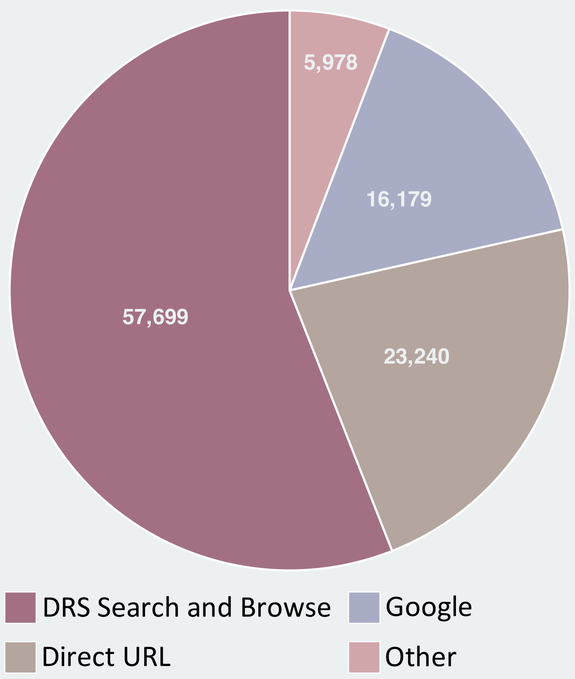
One Year of DRS Activity

Total views, downloads, and streams per month.



Top Five Referrers

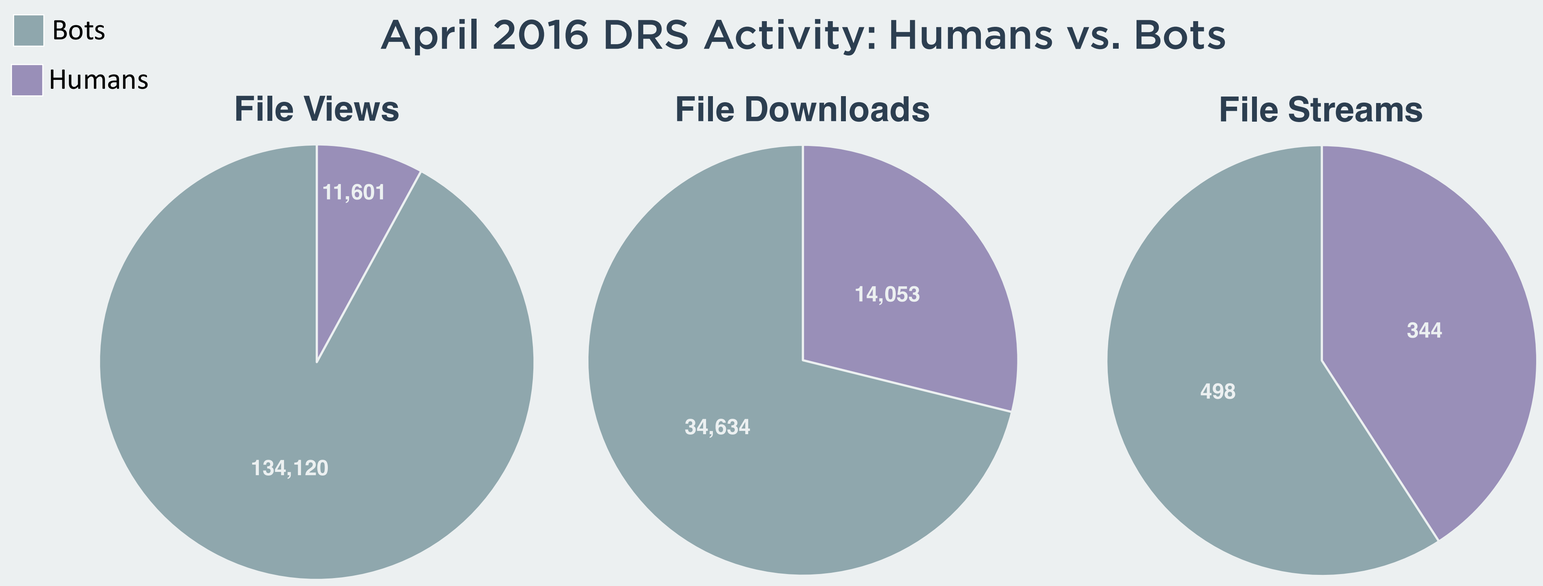
Top 5 referrers by number of page views



Not All Traffic is Equal

A seemingly endless number of bots exist to crawl publicly available repository content for harvesting and indexing. These bots help increase the discovery of repository content, but they can also greatly inflate usage statistics. Usage statistics are often gathered using third-party tools, like Google Analytics, which may or may not report their collection methods, and may not be aware of the difference between human and bot consumption. Although content owners tend to prefer higher numbers regardless of the consumer, we want to be able to defend the statistics we are gathering for our repository and declare them to be a genuine reflection of the use of our content by people, not bots.

April 2016 DRS Activity: Humans vs. Bots



Filtering

We designed a simple method for collecting and processing our usage statistics that allows us to filter out the non-human consumption of our content. Raw, unfiltered DRS usage statistics are stored in an impressions table in a SQL database. A nightly job processes this table by comparing the agent responsible for the impression against a list of keywords associated with known bots. Impressions made by agents that match any keyword are marked as FALSE and are filtered out of the statistics displayed to users in the interface.

The Impressions Table

Every file view, download, and stream is counted as an impression and recorded in the impressions table. Along with the type of impression, we also record the agent responsible for the impression, how the agent was referred to the file, the agent's IP address, and the date of the impression. Impression frequency is limited to one per file per IP address per hour, which helps reduce inflated numbers from repeated clicks or page refreshes.

ID	PID	session_id	action	ip_address	referrer	status	user_agent	public	created_at	updated_at	processed
184159	neu:190034	5f79a9934…	view	108.20.51…	direct	COMPLETE	Mozilla/5.0 (iPad; CPU OS…	TRUE	5/21/15 9:41	1/22/16 1:58	TRUE
184181	neu:190034	a745a871e…	view	129.10.107…	direct	COMPLETE	Mozilla/5.0 (Windows NT 6.1…	TRUE	5/21/15 12:32	1/22/16 1:58	TRUE
184182	neu:182058	a745a8715…	download	129.10.107…	direct	COMPLETE	Mozilla/5.0 (Windows NT 6.1…	TRUE	5/21/15 12:33	1/22/16 1:58	TRUE
184183	neu:1302	e5d85e81e…	view	129.10.106…	https://repository…	COMPLETE	Mozilla/5.0 (Windows NT 6.1…	TRUE	5/21/15 12:42	1/22/16 1:58	TRUE

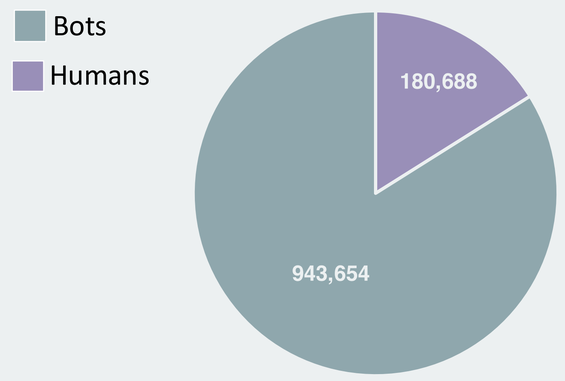
Significant Processing Values

- **"status"**: Files that are queued for download will be marked INCOMPLETE and are ignored until the download is finished.
- **"public"**: All impressions are initially set to TRUE. Once processed, impressions with agents on the bot list are marked as FALSE.
- **"processed"**: All impressions are initially set to FALSE. Once processed, the value is set to TRUE.

Ignored Agents

When the impressions table is processed, the agent responsible for each impression is compared against a list of common keywords associated with bots and crawlers, which is used to filter out agents from the impressions table. These keywords include: *archive*, *bot*, *crawl*, *curl*, *java*, *lynx*, *nutch*, *scrape*, *scrapy*, *slurp*, and *spider*

All DRS Impressions



Using the Data

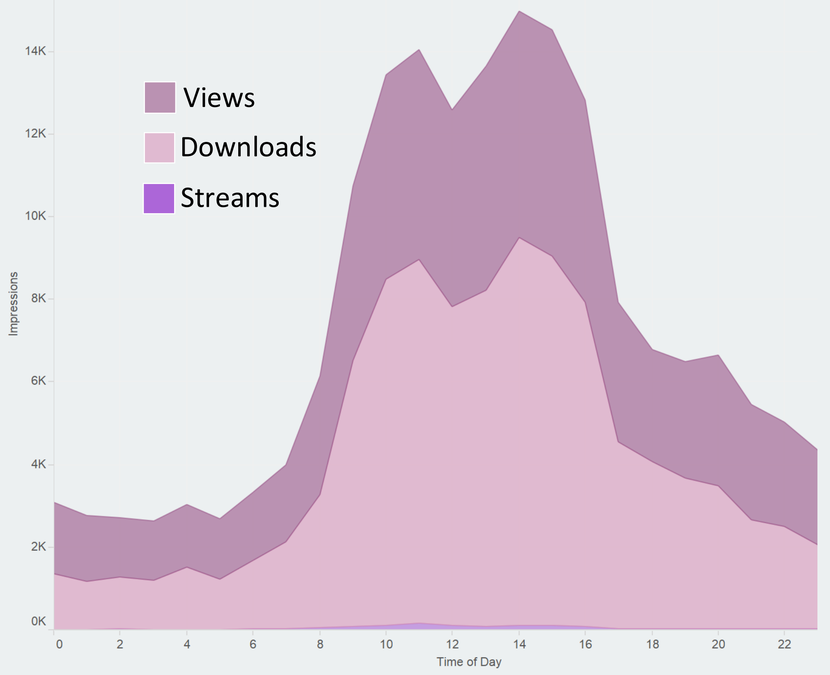
Categorizing our traffic and sharing our statistics gathering process has enabled us to more accurately track how the DRS is being used and more confidently defend our usage statistics as a reflection of genuine use of repository content. In 2016 we plan to improve our data gathering practices by:

- Inserting additional data points into the impressions table
- Operationalizing the process for adding new agents to the bot list
- Improving our statistical displays, including aggregated statistics and geographic visualizations

Improved Workflows

We also would like to use the data to improve our workflows. For example, the chart on the right can tell us the best time of day to schedule system deploys (between midnight and 6 am).

Measuring impact accurately is a difficult task, as is being able to confidently defend how measurements are recorded. Although there is important statistical value in recording automated bot traffic, measuring and reporting genuine repository use can improve our ability to communicate the true impact of repository content.



DRS *Digital Repository Service*
repository.library.northeastern.edu

Learn More

For more information about the DRS visit dsg.neu.edu/resources/drs
or github.com/NEU-Libraries/cerberus

Title

Northeastern University Library

Sarah Sweeney sj.sweeney@neu.edu

repository.library.northeastern.edu

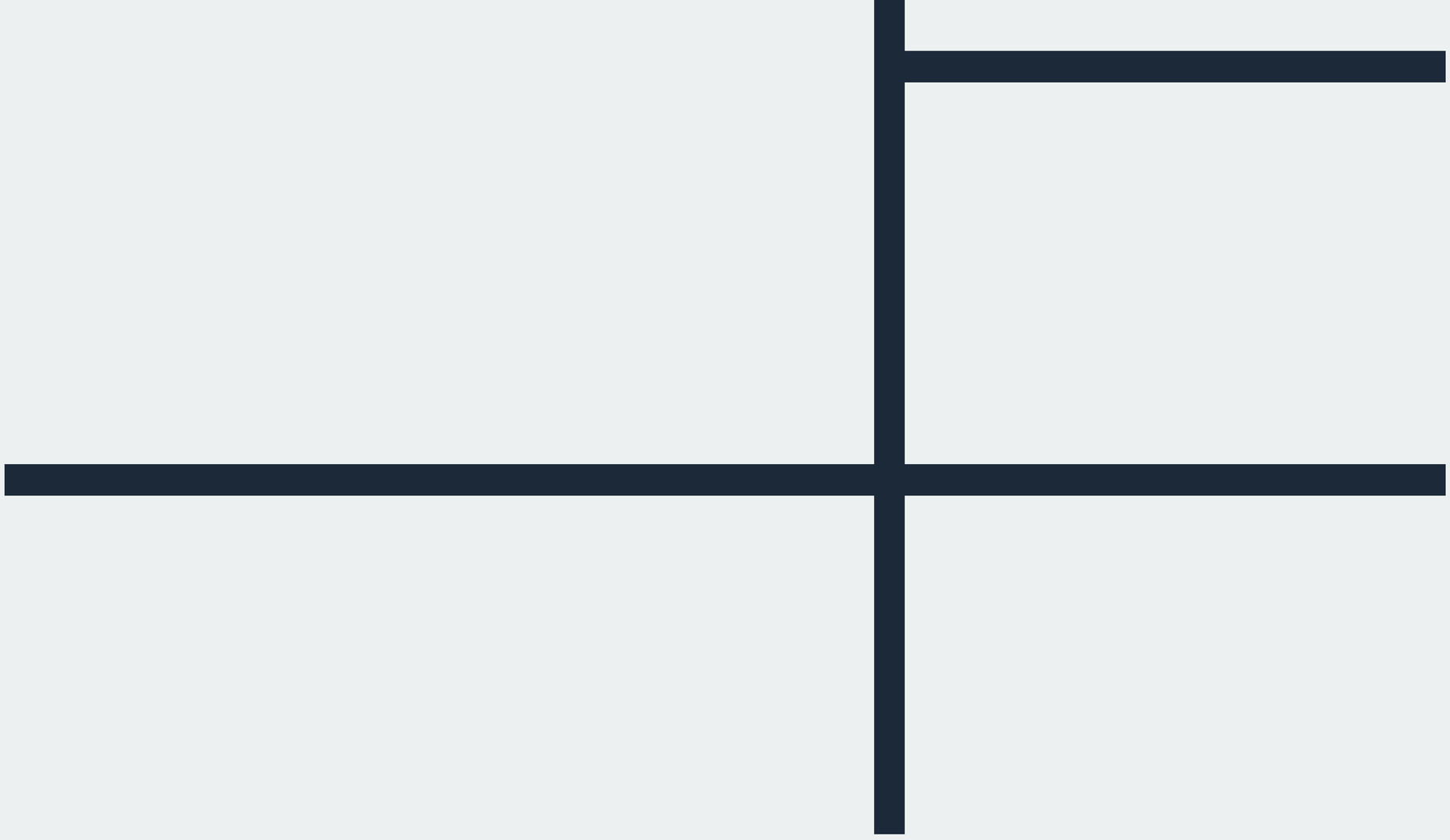
Heading 1

Paragraph text

The community framework has not just neatly organized repository content according to the existing Northeastern college and department structure, it has made it easier for the system to leverage the relationships between objects to enhance the discoverability of scholarly content in the repository.

Heading 2

- Valuable repository content can be discovered through multiple search and browse options.
- Communities and collections are easily organized according to an existing authoritative framework.
- The repository structure follows a model that is quickly understood by Northeastern users.



Learn More

For more information about the DRS visit
dsg.neu.edu/resources/drs or
github.com/NEU-Libraries/cerberus

DRS