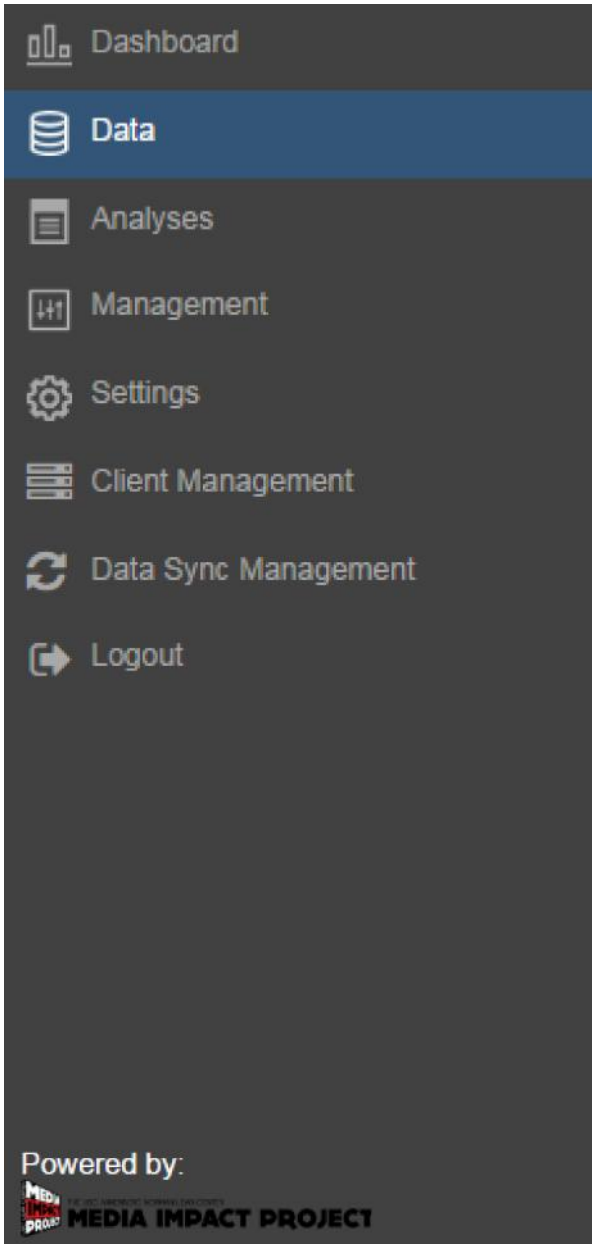


USC Media Impact Project
Data Repository Overview

December 2016

Dana Chinn
USC Annenberg School for Communication & Journalism

Mike Lee
USC Viterbi School of Engineering

Jonathan Weber
LunaMetrics

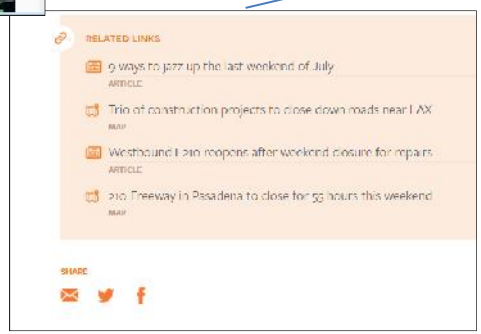
Development team
Daniel Nicolas, Developer
Justice Haque, Rebecca Wu
Steve Yin, Arpi Mardirossian

MIP Data Repository Publicly Available Documentation

No.	Document	Description
1	**Data Repository Overview** (this document)	Overview of the MIP methodologies for user and story metrics for media organization managers and analysts. Description of the MIP dashboard reporting interface used by the three MIP pilot organizations to access some of the data collected in the MIP data repository.
2	**Google Analytics Custom Reports**	Overview of all of the custom metrics gathered with the MIP Google Tag Manager that are now available to the three pilot organizations through custom reports in the standard Google Analytics reporting interface.
3	**Architecture Overview**	Technical architecture overview
4	**Data Repository Technical Overview**	Technical architecture whitepaper
5	**GitHub account**	Open source code and documentation for the repository, Google Tag Manager event tracking, the dashboard reporting interface and MIP + NewsLynx.

Purpose

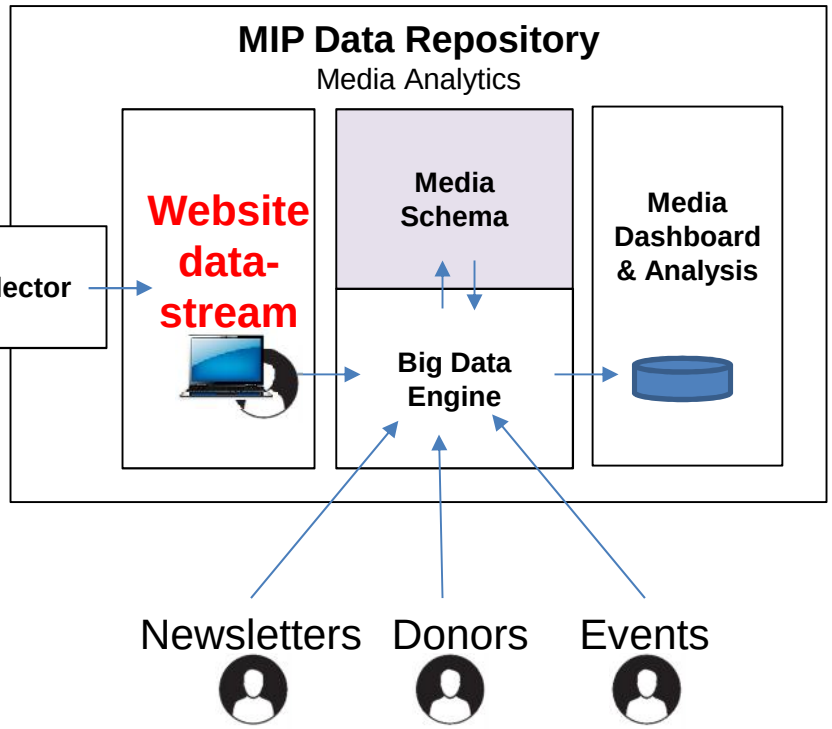
To understand actual people's digital media behavior across products vs. analyzing user data in product silos



Segmentation by user behavior varies by org and site architecture.

Example: KPCC has 24-hour livestreaming, but Texas Tribune and Wisconsin Watch don't.

The repository also includes custom metrics that indicate interactions after exposure to content such as scroll rate, attention minutes, and shares from the site.

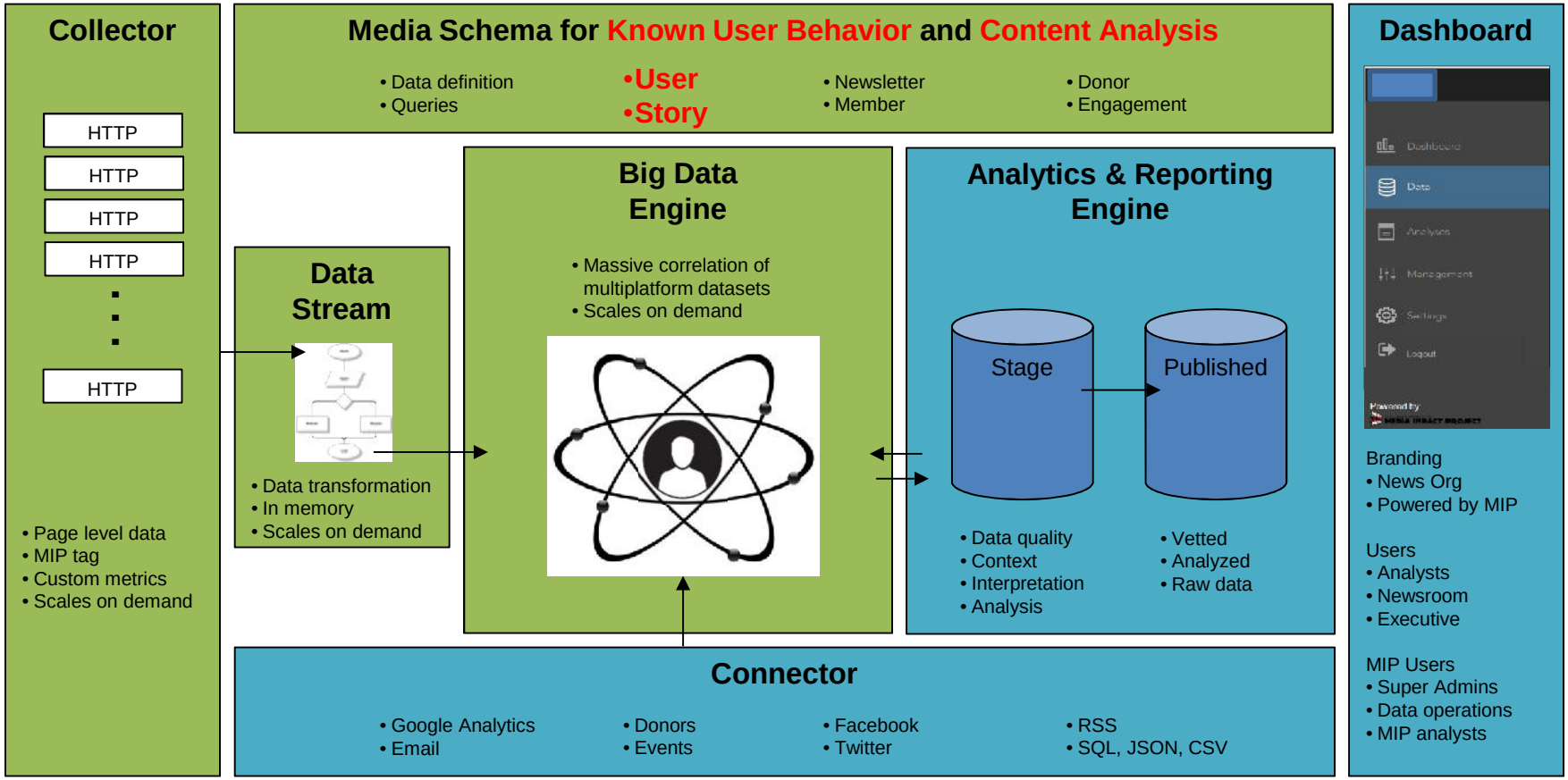


Media products that indicate behavior by an organization's most loyal users

MIP Data Repository Application Architecture

News orgs have to spend more resources on grouping and classifying page URLs than other industries due to the sheer volume of content that's updated, put into context, part of a series, etc.
We have found that we can't use the raw data that come from analytics software, content management and RSS feeds, so we have to develop both automated and manual methodologies to define what a "story" is.

See page 18 for a detailed explanation of how the MIP data repository gives media organizations data about their stories that's not available through the standard Google Analytics reporting interface.

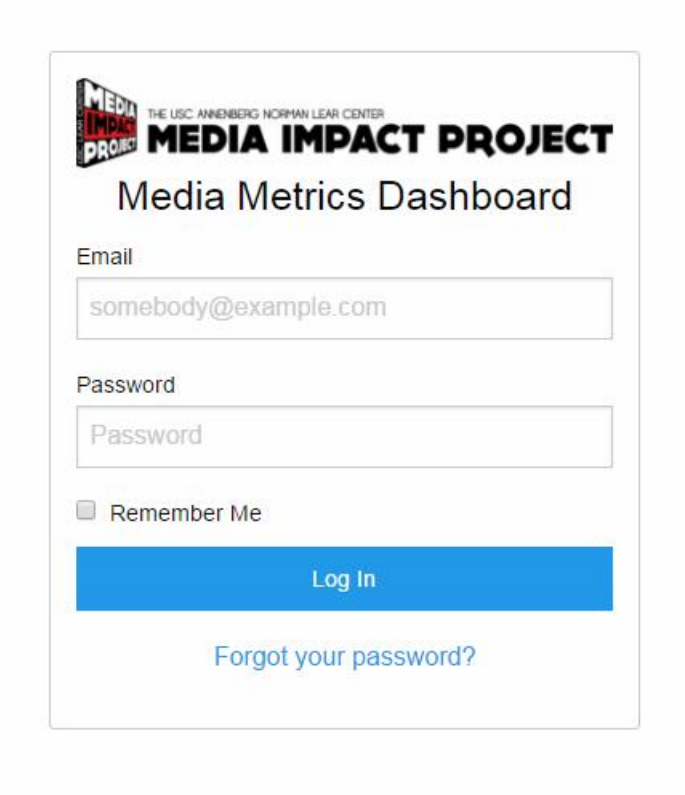


MIP Dashboard Reporting Interface Summary and Release Dates

No.	Section name	Section description	Page
1	**Dashboard**	Known user metric highlights by week	7
2	**Data**	• Known users: email subscribers, donors, members • Stories • Email newsletters • MIP user vs. standard Google Analytics user counts	8 13 21 22
3	**Analyses**	MIP and/or client reports	23
4	**Management**	System information that could affect analyses	24
5	**Settings**	Client-controlled user set-up	25
6	**Client Management**	MIP admin: Setting up clients	26
7	**Data Sync Management**	MIP admin: Checklist for releasing weekly data	27

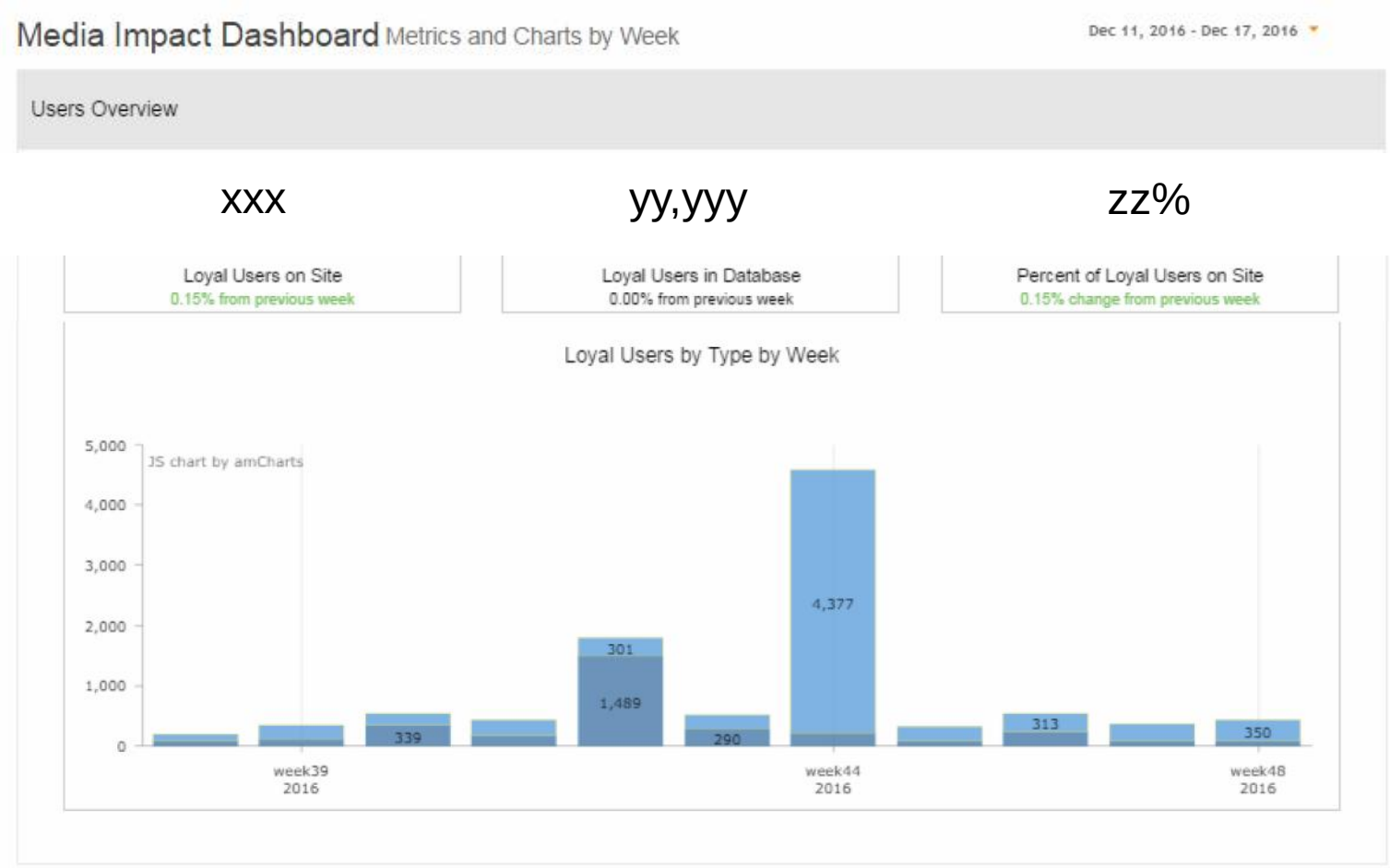
Version	Release date	Clients
0.2	July 1, 2016	KPCC (data since July 2015)
0.3	August 1, 2016	KPCC, Texas Tribune (data since July 2015)
0.4	September 1, 2016	KPCC, Texas Tribune, Wisconsin Watch (data since July 2016)
1.0	October 3, 2016	KPCC, Texas Tribune, Wisconsin Watch
	December 17, 2016	Last day of data collected

URL: http://app.mediaimpactproject.org



Section: **1-Dashboard**
Tab: **Users Overview**

Use: **How many of our most loyal users came to the site this week?**



Section: **2-Data**
Tab: **Users**

Table: **Total Known Users**

Use: **How many of our most loyal users came to the site this week?**

Loyal users are those who have come to the site since MIP started collecting data and whose e-mail addresses are known to MIP as:

1. Email newsletter subscribers only
   OR
2. Donors only
   OR
3. Members who have logged in (Texas Tribune only)
   OR
4. Both email newsletter subscribers and donors, or other combinations of subscribers, donors and logged-in members

The **number of loyal users who come to the site each week** is a Key Performance Indicator for email newsletter performance and donor drives.

The **total number of loyal users in the MIP database** and **percent of loyal users on site** will be KPIs for total audience growth after MIP has collected data for 6-12 months.

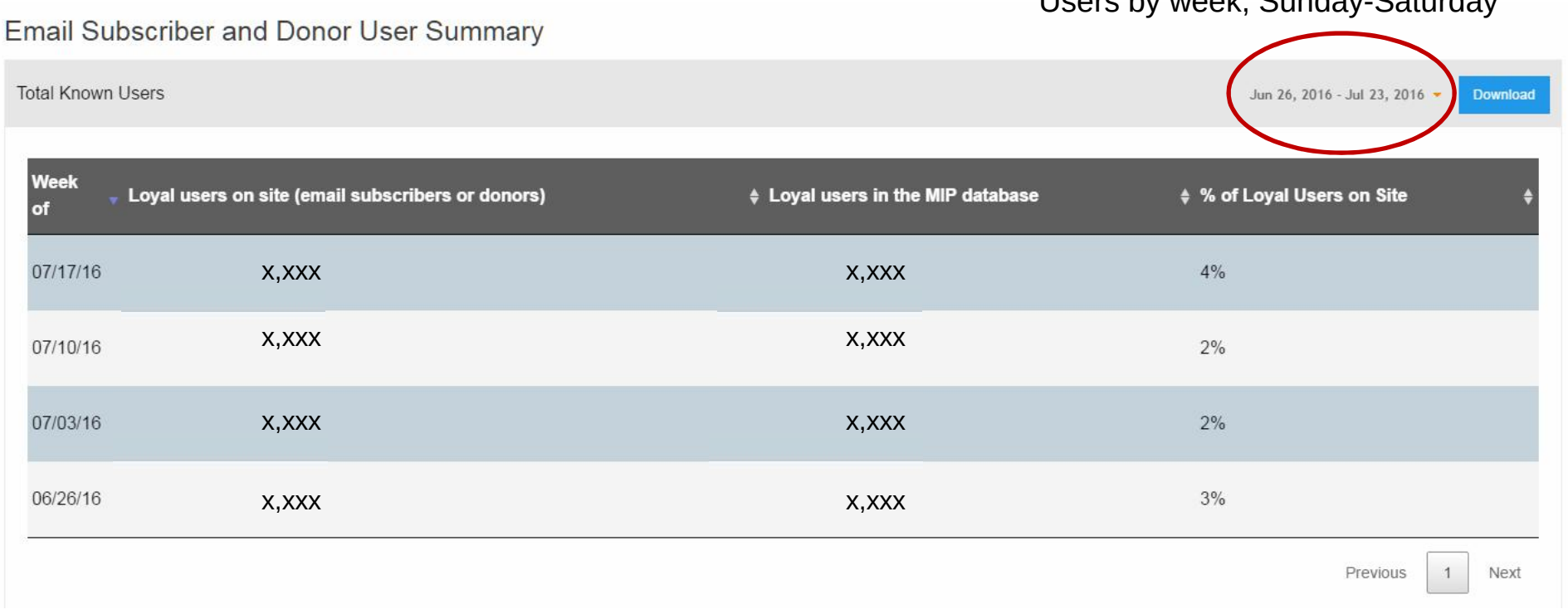
Note: See Appendix A for more detailed explanations of how known users are identified and counted.

Section: **2-Data**
Tab: **Users**

Table: **Total Known Users**

Use: **How many of our most loyal users came to the site this week?**

Users by week, Sunday-Saturday



Email Subscriber and Donor User Summary

Total Known Users

Jun 26, 2016 - Jul 23, 2016 | Download

Week of	Loyal users on site (email subscribers or donors)	Loyal users in the MIP database	% of Loyal Users on Site
07/17/16	X,XXX	X,XXX	4%
07/10/16	X,XXX	X,XXX	2%
07/03/16	X,XXX	X,XXX	2%
06/26/16	X,XXX	X,XXX	3%

Previous 1 Next

Note: See Appendix A for more detailed explanations of how known users are identified and counted.

Section: **2-Data**
Tab: **Users**

Additional notes on known users

1. The MIP custom tracking code allows us to identify individuals identifiable by e-mail address. Once an individual has come to the site via an e-mail newsletter or a donation, the organization can track his/her *past and future* digital product behavior.

2. The MIP media metrics focus on identifying donors as a type of user behavior, i.e., people who donate online, with the goal of analyzing what content donors view vs. non-donors. The data repository doesn't provide metrics for analyzing donor behavior, e.g., how many times they donate, the median amounts, types of donations, etc.

3. Email subscriber and newsletter fields are standardized based on MailChimp (Texas Tribune and Wisconsin Watch).

   Texas Tribune and Wisconsin Watch known user counts were continuously updated via direct access to MailChimp feeds. SCPR/KPCC email and donor data was from batch uploads of Oracle Eloqua data.

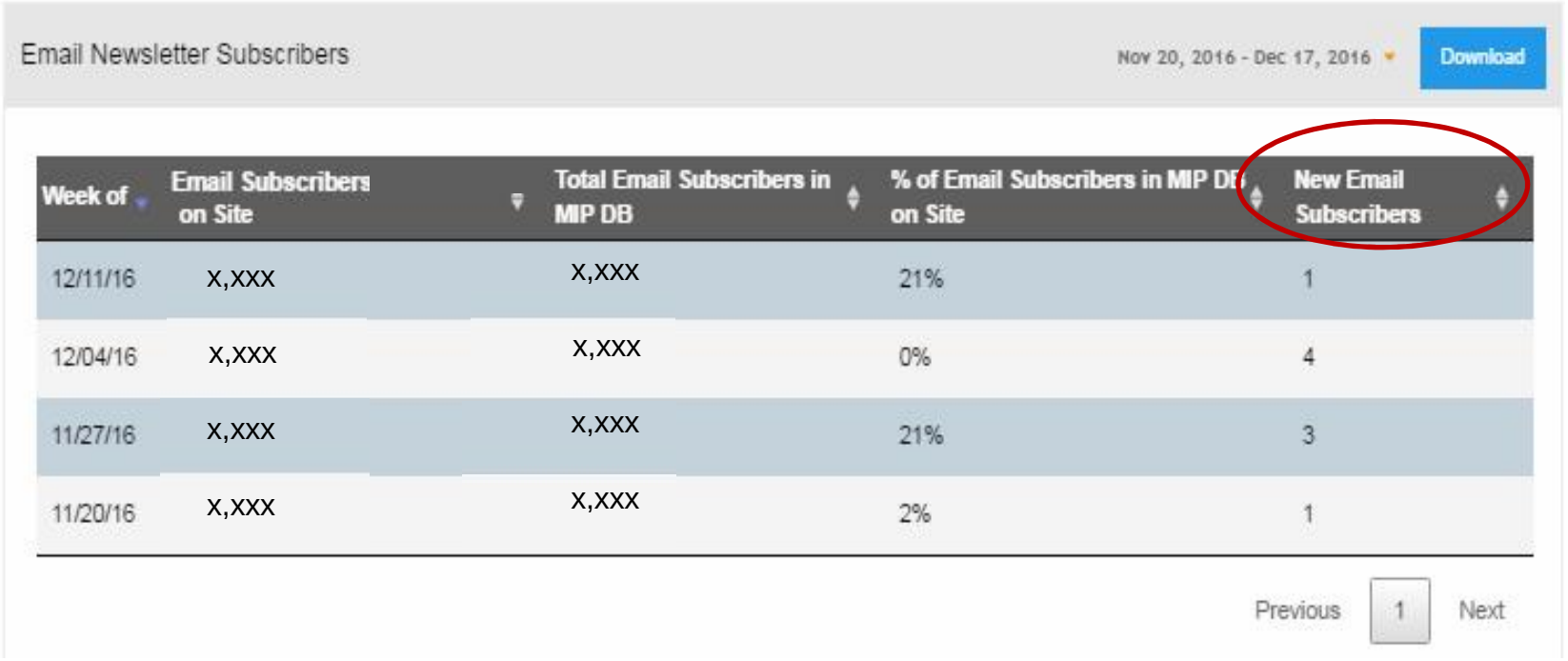
Section: **2-Data**
Tab: **Users**

Table: **Email Newsletter Subscribers**

Use: **How many email subscribers came to the site this week?**

The trend in the number of email subscribers per week is a KPI for newsletter performance. The number by specific newsletter may be available in a future Data/Newsletter tab, as will the open and click-through rates, number of clicks, and other actionable KPIs.

The number of new email subscribers per week, a possible KPI, will vary widely depending on marketing campaigns and new newsletter launches.



Section: **2-Data**
Tab: **Users**

Table: **Donors**
Members (Texas Tribune only)

Use: **How many donors came to the site this week?**

The number of donors who come to the site is a KPI for analyzing hypotheses about the correlation between site/content engagement and donation revenue.

Donors

Jul 10, 2016 - Aug 6, 2016 Download

Week of	Donors on Site	Donors in MIP DB	% of Donors in MIP DB on Site
07/31/16	X,XXX	X,XXX	1%
07/24/16	X,XXX	X,XXX	1%

Members

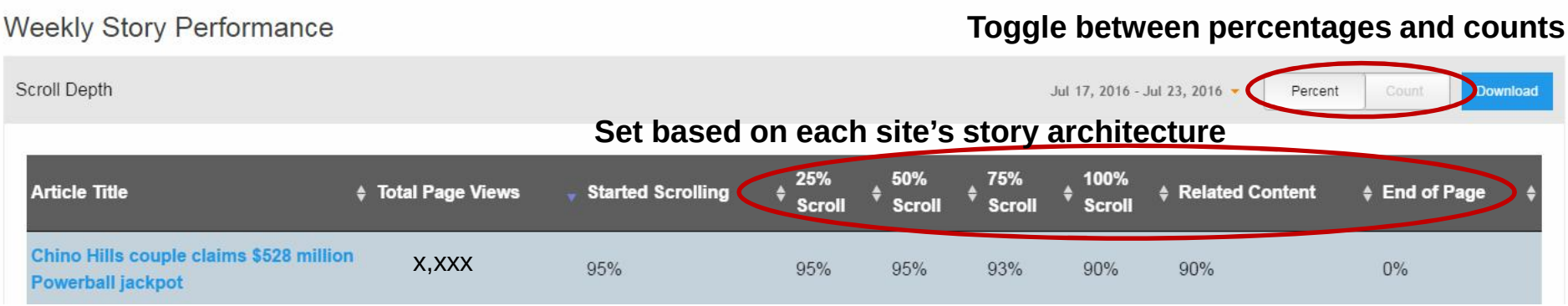
Jul 10, 2016 - Aug 6, 2016 Download

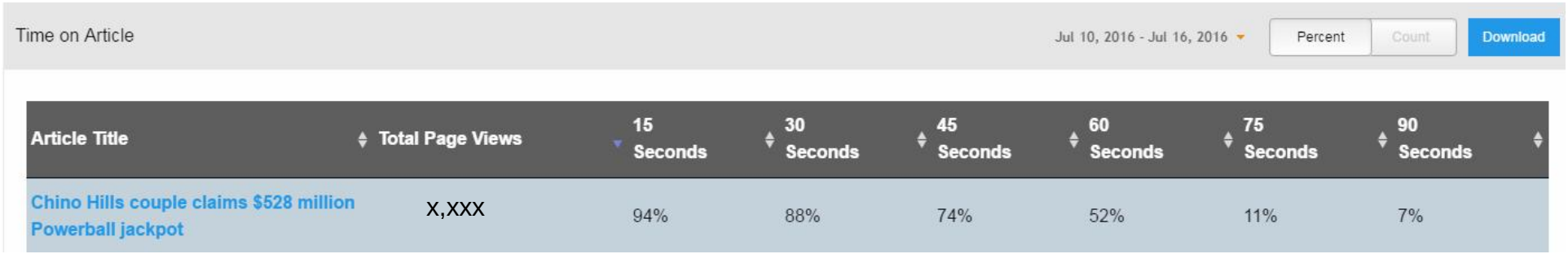
Week of	Known members on the site this week	Known: Total known members in the MIP database	Known: Percent of members in the MIP database who logged in this week
07/31/16	X,XXX	X,XXX	2%
07/24/16	X,XXX	X,XXX	2%

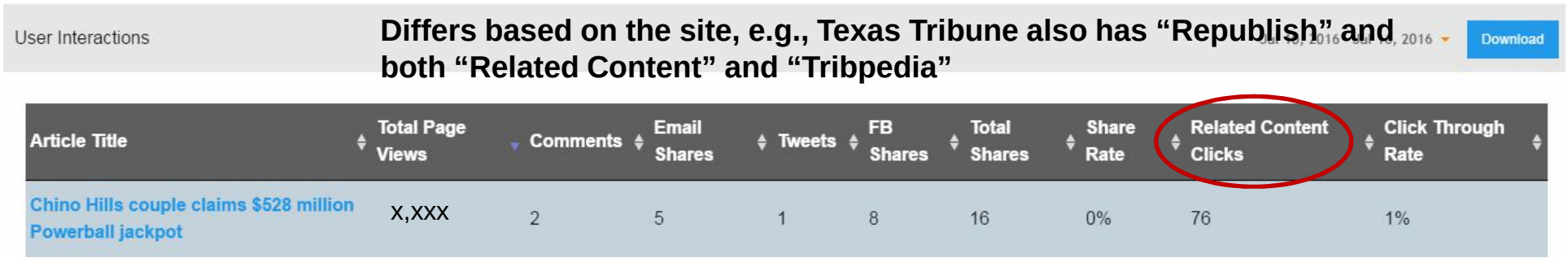
Section: **2-Data**
Tab: **2.2 Stories**

The Stories tab has weekly numbers that include:

- Stories that are posted during the specified week, and the data for that week
- Stories that were posted in previous weeks but still got traffic during the specified week







Section: **2-Data**
Tab: **2.2 Stories**

Table: **Scroll depth**

Use: **How many pageviews involved scrolling?**

Scroll depth is a proxy indicator that a user read all or part of a story. It can be used as a proxy for awareness of a particular part of a story based on the part's location.

The scroll depth for Related Content at the end of a story is combined with the Related Content clicks in the User Interactions table to calculated Related Content click rate.

MIP will be exploring hypotheses on how scroll depths, segmented by topic, can be analyzed with attention minutes, share rate and related content click rates.

Weekly Story Performance

"Related Content" and "Tribpedia" on the Texas Tribune site are at the same place.

Scroll Depth

Jul 10, 2016 - Jul 16, 2016 | Percent | Count | Download

Article Title	Total Page Views	Started Scrolling	25% Scroll	50% Scroll	75% Scroll	100% Scroll	Related Content	End of Page
Chino Hills couple claims $528 million Powerball jackpot	X,XXX	95%	95%	95%	93%	90%	90%	0%
Slideshow: Winds swell Sand Fire in Santa Clarita, forcing evacuations	X,XXX	89%	75%	56%	45%	30%	29%	0%
5 weeks later, California's primary results officially official	X,XXX	95%	94%	93%	86%	68%	67%	0%

Section: **2-Data**
Tab: **2.2 Stories**

Table: **Time on article, or attention minutes**

Use: **How many pageviews involved users seemingly paying attention?**

Counting is suspended:

- If the tab isn't active in the browser
- If more than 60 seconds have elapsed since the last mouse activity on the page
- After 30 minutes on the same page (same as session timeout)

Time on Article

Jul 10, 2016 - Jul 16, 2016 | Percent | Count | Download

Article Title	Total Page Views	15 Seconds	30 Seconds	45 Seconds	60 Seconds	75 Seconds	90 Seconds
Chino Hills couple claims $528 million Powerball jackpot	X,XXX	94%	88%	74%	52%	11%	7%
Slideshow: Winds swell Sand Fire in Santa Clarita, forcing evacuations	X,XXX	85%	71%	60%	50%	19%	13%
5 weeks later, California's primary results officially official	X,XXX	93%	77%	58%	42%	12%	7%

Google Analytics standard time on site (or session duration) is based on calculating when a page is sent from the server. It doesn't capture time spent on the last page a user got, or the time spent during sessions with only one page view (i.e., a bounce).

Section: **2-Data**
Tab: **2.2 Stories**

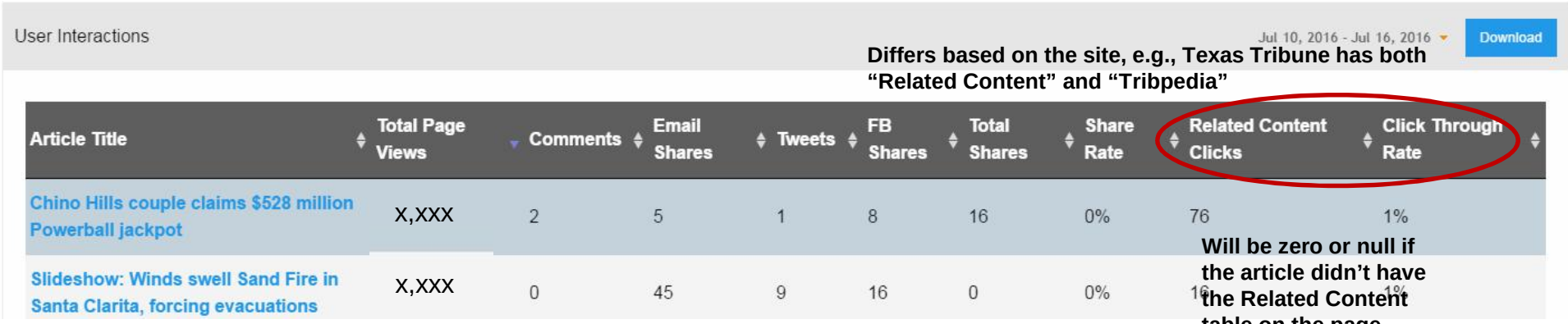
Table: **User interactions – comments, shares, related content clicks**

Use: **How much interaction did a story have?**

We assume that users who click on “email,” “Twitter,” “Facebook” and other action buttons at the top and bottom on a story are more likely to have read the story than those who shared it from outside the site.

User interactions will vary greatly by topic and whether there are related stories, e.g., a user may comment on the second story he/she reads. Thus, we recommend that user interactions be analyzed by topic, story series, story format and other variables. Site-wide medians and rankings shouldn’t be KPIs.

Also, each interaction type should probably be weighted differently. As they need to input email addresses and messages, users who email a story could be more highly engaged than users who share a story via Twitter or Facebook.



The scroll depth for Related Content at the end of a story is combined with the Related Content clicks in the User Interactions table to calculated Related Content click rate.

Section: **2-Data**
Tab: **2.2 Stories**

The Stories tab has weekly numbers that include:

- Stories that are posted during the specified week, and the data for that week
- Stories that were posted in previous weeks but still got traffic during the specified week

Not in the dashboard reporting interface, but available via custom queries or reports:

- **Total traffic/metrics per story over a specified time period.** This report would include:
  - Stories that were posted during the specified time period
  - The cumulative traffic/metrics per story from the time the story was posted to the date the report is run.

  Note: This type of cumulative report is similar to what's available by post in Facebook Insights.

  Example: The Texas Tribune needs to report on the impact of its second quarter 2016 higher education coverage. The report would include:
    - Stories that were posted April 1, 2016 through June 30, 2016
    - Pageviews and other metrics per story from April 1, 2016 through July 23, 2016, when the report was run.

- **Trending by week for a story or group of stories**
  This report would be used to track the impact of a story in response to external and internal events, e.g., explainers, stories in a series.

Section: **2-Data**
Tab: **2.2 Stories**

Definition of a "story"

- **The number of stories is essential to analyze content strategy and is often required for funder reporting.**

- **However, each organization's site architecture and analytics tagging strategy requires its own story counting methodology for URLs vs. page titles.**
  - Comments pages in the Texas Tribune are separate from the story and have different URLs. On KPCC, comments are on the same page/URL as the story.
  - The MIP Google Tag Manager code is not included on all pages of a site. For example, it is not included on Texas Tribune app pages, so those stories and data aren't in the dashboard.

- **In both the MIP data repository and in Google Analytics, pageview reports are based on unique URLs.** Selecting page title or adding page title as a secondary dimension results in getting different data. The number of rows in a pageview report is usually used as a count of the total number of articles, as each unique URL usually only has one page title.

- **A media site's URL and page title structure often includes too many irregularities for organizations for the number of rows to be an accurate count of the number of articles, or for any one row to include an accurate aggregation of the metrics for any one story.** The irregularities are usually due to media organizations updating a page multiple times a day, with each update sometimes generating different unique URLs and/or page titles for the same article.
  - In the standard Google Analytics reporting interface, Google Analytics will aggregate pageviews and other metrics for a site that exceeds 50,000 URLs a day. Stories will be listed as "(other)" rather than by their URLs and page titles. This is a known issue with the Texas Tribune.
  - MIP data repository reports show all URLs and page titles, regardless of whether a site exceeds 50,000 URLs a day.

Section: **2-Data**
Tab: **2.2 Stories**

Definition of a "story" (continued)

- **A story that's updated or is in a different language may have a different page title but the same URL.** For example, all of the following page titles have the same URL:
    - Lawsuit Won, Texas Moves to Cut Therapy Programs
    - Lawsuit Won, Texas Moves to Cut Children's Therapy Programs
    - Demanda Won, Texas mueve a recortar los programas de terapia
- **Similarly, the Interactions report can include data in multiple rows that need to be aggregated.** For example, the number of Facebook Recommends for an article on the SCPR/KPCC site may be in three different rows:
    - Facebook : recommend
    - N/A – www.scpr.org : N/A Facebook
    - N/A – www.scpr.org : N/A Facebook ã§ã‚·ã‚§ã¢
- **The Stories tab lists articles by page title**, and each page title clicks to a URL. Thus, the tab has rows with duplicate stories, and can't be used to get the count of stories or to rank the stories. However, the downloaded report includes both page titles and URLs, with each row being a unique combination of page title and URL.
    - To get an accurate story counts and metrics by stories, organizations can download the story report, group together the stories, and aggregate the data.
    - In a test of three months of Texas Tribune story data, we found that:
        - A report of about 3,000 rows actually represented only about 1,000 stories.
        - After trying out one algorithm, there were no obvious consistent patterns to the irregularities, so the groupings of rows may need to be done manually. See Appendix B for a detailed breakdown of the irregularities.

Section: **2-Data**
Tab: **2.2 Stories**

Implications for media metrics

- **Defining a "story" is essential for all media metrics operations and research, as the raw story-level data is incomplete** and thus leads to underestimating a story's impact and mispresenting its relative importance to other stories.
  - Coding by topic, part of series, story format and other metadata should not be done until the row data is aggregated.
  - Medians by week and other descriptive statistics based on the raw data are probably irrelevant.
  - Combining methodologies will get more complicated if we add datasets from other analytics systems (e.g., MailChimp), content management systems and RSS feeds.

- **The root problem is not in Google Analytics and other digital analytics tool but rather in each organization's site architecture** and in its investment in the resources needed to maintain a systematic, consistent URL and page title assigning process across its content management system and all of the digital platforms it controls – site, email, mobile apps, etc.
  - Media organizations can get far more from the standard Google Analytics reports with standardized, canonical URLs and page titles. For example, summarizing content performance by category could be a simple query rather than a process that involves applying a taxonomy manually after data is collected.
  - Developing best practices site architecture maintenance processes has operational benefits beyond media metrics, such as optimizing a site for search engines (SEO).

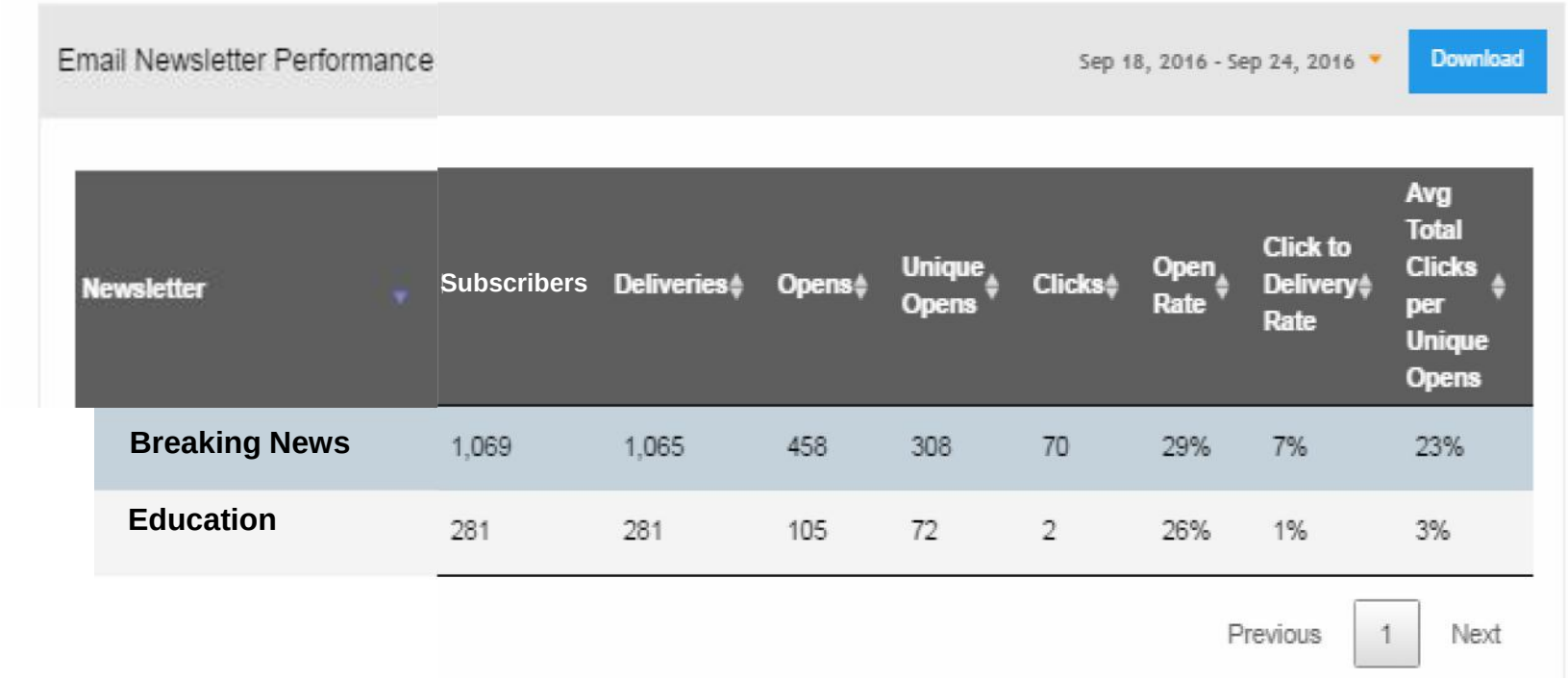
Section: **2-Data**
Tab: **2.3 Newsletters**

Table: **Email Newsletter Performance**

Use: **How much traffic did email newsletters send to our site? From whom?**

The click per unique open rate by type of email newsletter is a KPI that can lead to more in-depth analyses of:

- What stories appealed to an organization's most loyal and known users, email newsletter subscribers
- The number and type of known users who click on an email newsletter could also be a KPI.



Newsletter	Subscribers	Deliveries	Opens	Unique Opens	Clicks	Open Rate	Click to Delivery Rate	Avg Total Clicks per Unique Opens
Breaking News	1,069	1,065	458	308	70	29%	7%	23%
Education	281	281	105	72	2	26%	1%	3%

Section: **2-Data**
Tab: **2.4 Data Quality**

Table: **Variance: Identified Users**

Use: **How does the number of MIP users differ from the number in an organization's standard Google Analytics account?**

The MIP Google Tag Manager code is on most but not all of an organization's pages. This table shows the variance in identified users from week to week. A variance higher than 10 percent may indicate a problem with the MIP GTM code.

Or, it may be due to high usage of a section of a site where the MIP GTM code doesn't exist. For example, the MIP GTM code is not on the Texas Tribune app pages, or sometimes there was a lag in putting the MIP GTM code on a new section of the site.

Variance: Identified Users

Standard Google Analytics Account vs MIP Google Analytics / Google Tag Manager

Nov 20, 2016 - Dec 17, 2016 Download

Week of	Events	GA Users	MIP GTM Users	Variance
12/11/16		3,449	3,576	-4%
12/04/16		4,168	4,524	-8%
11/27/16		5,084	5,279	-4%
11/20/16		2,989	3,242	-8%

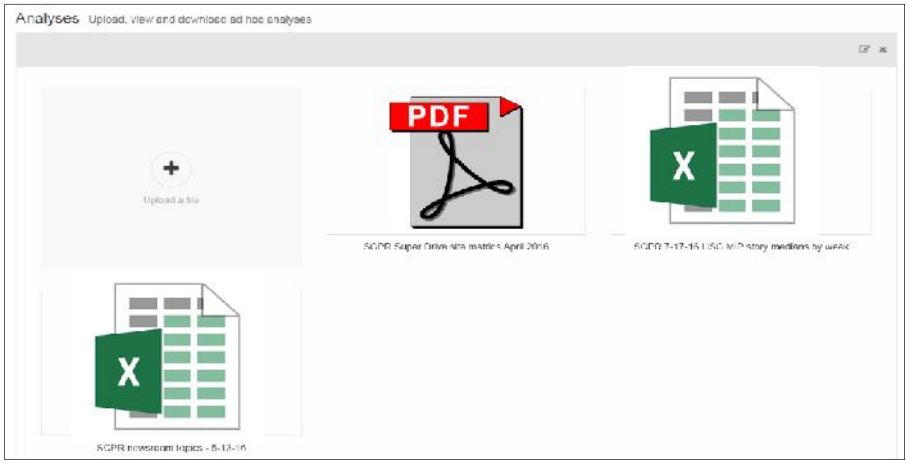
Section: **3-Analyses**

Use: **What analyses are available? What taxonomies should we use?**

The primary purpose of the Data section is to select and download data for analysis outside of the dashboard.

The purpose of the Analyses section is for MIP and the organization to post:

- analyses that answer specific questions using data from multiple sources
- files such as taxonomies and weekly medians



Custom Texas Tribune higher education report compiled with MIP custom metrics and topic coding

		Reach		"Read" and responded				Shares after exposure to stories on the site					
Topic	No. of stories	Pageviews	Percent of Total Pageviews	Pageviews scrolled ("read")	Percent of PVs scrolled ("read")	Clicks to related stories	Percent of clicks to related stories	No. of comments	Emailed	Tweeted	Shared on Facebook	Total shares from site	Percent of shares	Clicked on "Republish"

Section: **4-Management**

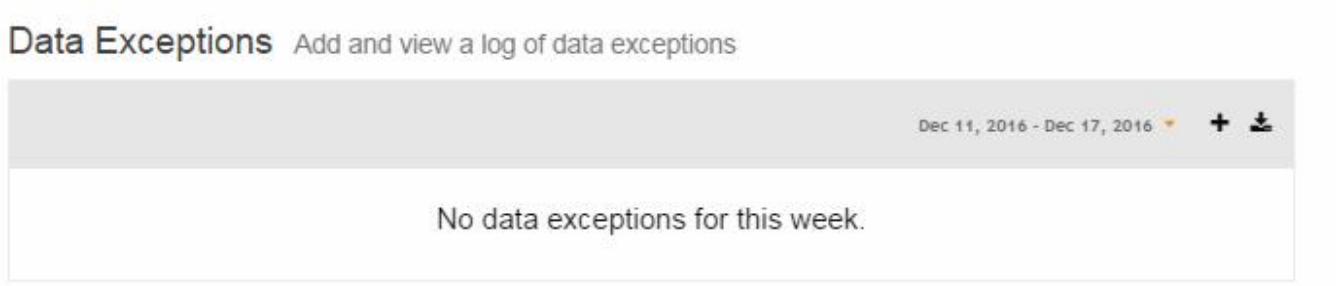
Table: **Data exceptions**

Use: **Data quality caveats for analysts**

This section is a manually updated log of any MIP or client issues that would affect the analysis and interpretation of the MIP data repository data.

Examples:

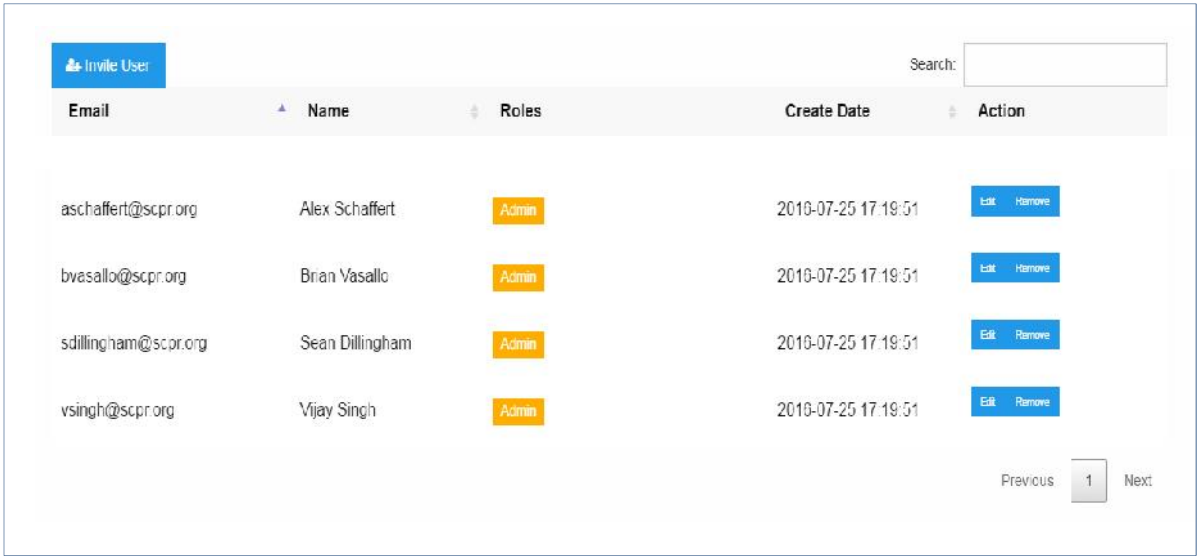
- January 19-28, 2016: GTM tracking code was removed from all but one page. This resulted in nearly no data being collected from most of the website.
- March 23, 2016: No comments were tracked due to a JavaScript race condition where the comment tracking code would not fire properly. Fix was added to the GTM update on March 24, 2016.

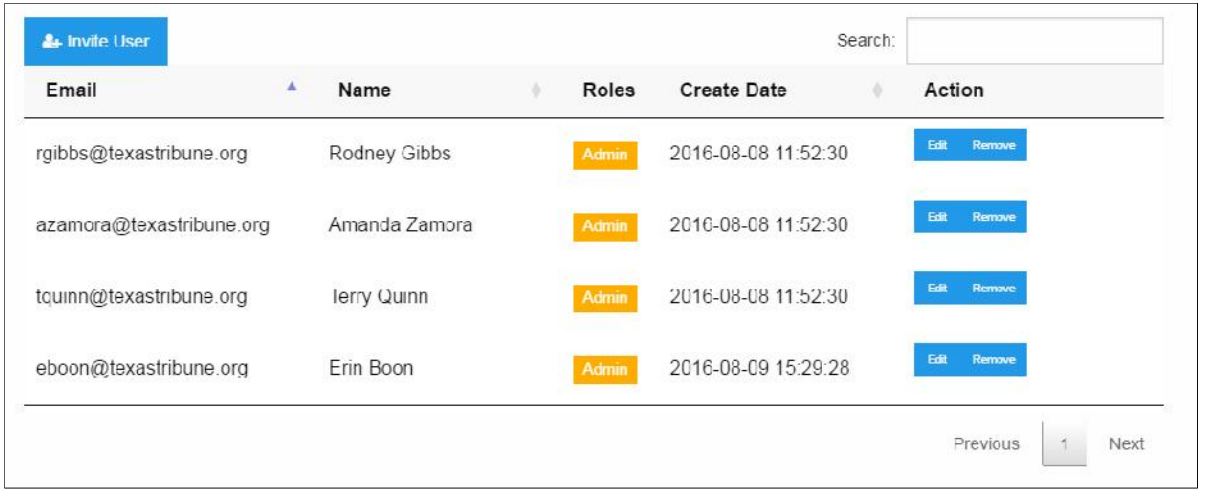


Section: **5-Settings**

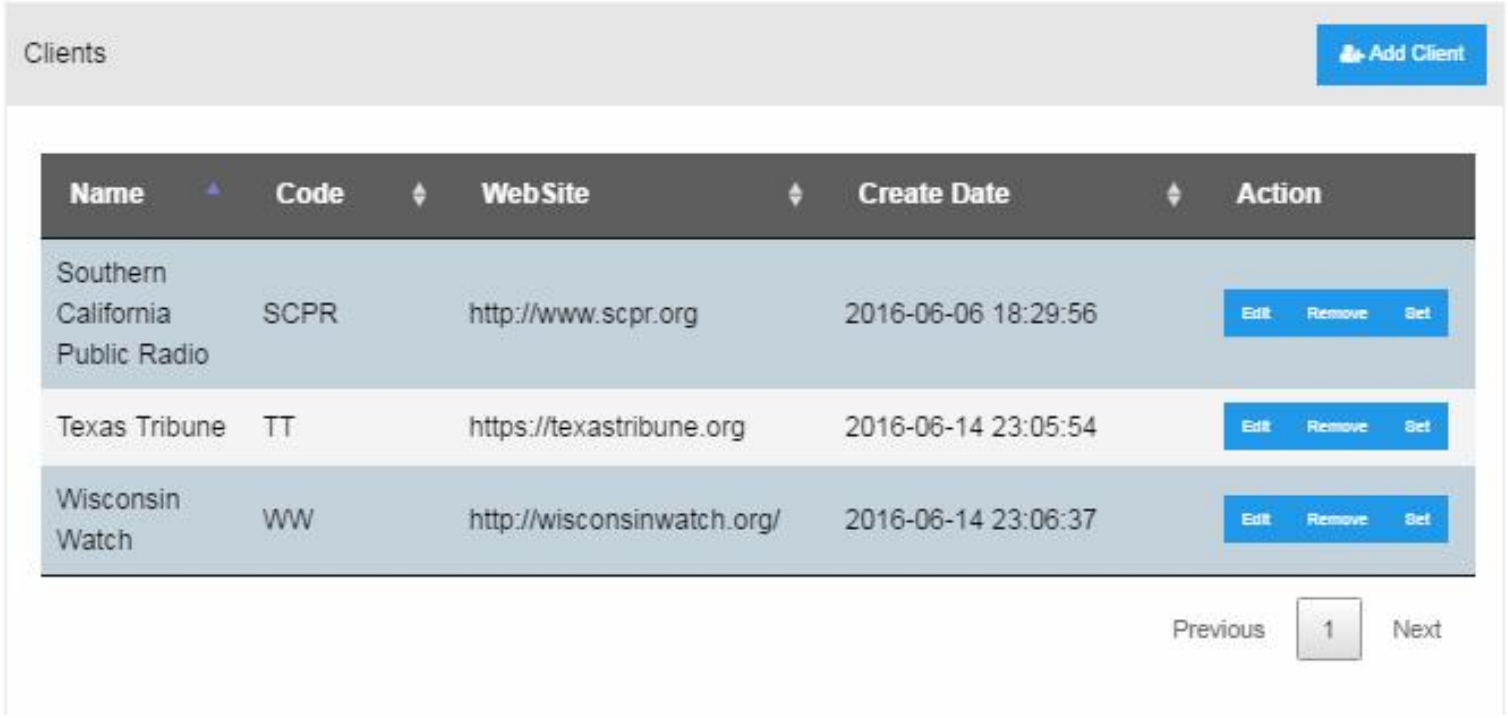
Use: **Client-controlled user set-up and editing**

- An organization can sign up an unlimited number of users.
- Hundreds of users across the three MIP pilot organizations can be on the system simultaneously.
- Each organization sees only its own users in this Settings section.





Section: **6-Client Management (MIP admin)**



Section: **7-Data Sync Management (MIP admin)**

Each data section was checked by an analyst each week before being released to the client.

Data Sync Monitor

Week	Southern California Public Radio			Texas Tribune			Wisconsin Watch			Action
	Users	Stories	Data Quality	Users	Stories	Data Quality	Users	Stories	Data Quality	
2016-12-11	mark as ready	mark as ready		mark as ready	mark as ready	mark as ready	mark as ready	mark as ready	mark as ready	
2016-12-04	Ready	Ready	Ready	Ready	Ready	Ready	Ready	Ready	Ready	
2016-11-27	Ready	Ready	Ready	Ready	Ready	Ready	Ready	Ready	Ready	
2016-11-20	Ready	Ready	Ready	Ready	Ready	Ready	Ready	Ready	Ready	
2016-11-13	Ready	Ready	Ready	Ready	Ready	Ready	Ready	Ready	Ready	
2016-11-06	Ready	Ready	Ready	Ready	Ready	Ready	Ready	Ready	Ready	
2016-10-30	Ready	Ready	Ready	Ready	Ready	Ready	Ready	Ready	Ready	

Appendix

A. Known User Methodology

B. Texas Tribune URL and Page Title Analysis

Appendix A-1: Identified vs. Known Users – SCPR/KPCC Example

Identified: Number of unique Google Customer ID (CID) with an ELQ code						
Known: Number of unique Eloqua customer codes (ELQ)						
Click number	**CID**	**ELQ**	**Date of click**	**Time of click**	**E-mail address from Eloqua database**	**Name from Eloqua database**
1	C6748	E2738	Monday	9:00 a.m.	Jean@gmail.com	Jean Smith
2	C9854	E2738	Tuesday	9:00 a.m.	Jean@gmail.com	Jean Smith
3	C0987	E2738	Wednesday	3:00 p.m.	Jean@gmail.com	Jean Smith
4	C0987	E5934	Thursday	4:00 p.m.	Jean@scpr.org	Jean Smith
5	C2637	E6473	Tuesday	9:05 a.m.	Joan@yahoo.com	Joan Jones
6	C2637	E6473	Tuesday	10:05 a.m.	Joan@yahoo.com	Joan Jones
	Number of Identified Users		4			
	Number of Known Users		3			
An e-mail newsletter subscriber who clicks from a newsletter from multiple devices will have multiple unique Google cookie IDs but only one Eloqua ELQ= code.						
1	C6748 (Home PC, Home e-mail)	E2738	Monday	9:00 a.m.	Jean@gmail.com	Jean Smith
2	C9854 (Phone, Home e-mail)	E2738	Tuesday	9:00 a.m.	Jean@gmail.com	Jean Smith
3	C0987 (Work PC, Home e-mail)	E2738	Wednesday	3:00 p.m.	Jean@gmail.com	Jean Smith
One person with two different e-mail addresses in the Eloqua database will be two known e-mail newsletter subscribers.						
3	C0987 (Work PC, Home e-mail)	E2738	Wednesday	3:00 p.m.	Jean@gmail.com	Jean Smith
4	C0987 (Work PC, Work e-mail)	E5934	Thursday	4:00 p.m.	Jean@scpr.org	Jean Smith
Current subscribers who want to add newsletters would probably do so from the "update subscriptions" link in a newsletter.						
Clicks to the site from a forwarded e-mail newsletter are captured with the CID/ELQ of the person who forwarded it.						

Appendix A: Loyal User Calculations – SCPR/KPCC Example

Reference	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
	How many e-mail subscribers came to the site THIS WEEK? How many were (1 new subscribers and/or (2 new to MIP?										How many e-mail subscribers do we have in the MIP database?		Calculated field	E-mail newsletter clicks per week
	Identified: Subscribers already in MIP database who came to the site this week	Known: Subscribers already in MIP database who came to the site this week	Identified: Subscribers who came to the site through an e-mail this week for the first time since MIP started collecting data	Known: Subscribers who came to the site through an e-mail this week for the first time since MIP started collecting data	Identified: New subscribers this week who also clicked on an e-mail this week	Known: New subscribers this week who also clicked on an e-mail this week	**Identified e-mail newsletter subscribers THIS WEEK**	**Known e-mail newsletter subscribers THIS WEEK (unique ELQs)**	Identified: New e-mail subscribers this week	Known: New e-mail subscribers this week	**Identified: Total identified e-mail newsletter subscribers in the MIP database**	**Known: Total number of known e-mail newsletter subscribers in the MIP database**	Known: Percent of subscribers in the MIP database who clicked on an e-mail this week	The number of pageviews with an ELQtrackID-code from unique Google cookie IDs (weekly unique users) that also have an Eloqua ELQ=code
Validation	Should be less than P	Should be less than Q	May vary widely	May vary widely	Should be less than N	Should be less than O	Should equal I+H+J=L. Should be compared to the number of sessions coming from e-mail in Google Analytics	Should be less than L and equal G+H+K=M. Should be less than or equal to the total number of subscribers in the Eloqua database	Identified new subscribers should be equal to Known new subscribers?		P should equal previous week P plus H + N	Q should equal previous week Q plus I + O	M/O = R should be less than 100%. May vary widely.	Should be compared to the number of sessions coming from e-mail in Google Analytics Acquisition/Channels/Email/Landing Page-Campaign

Reference	T	U	V	W	X	Y	Z	AA	BB	CC	DD	EE
	How many donors came to the site THIS WEEK?						How many donors do we have in the MIP database?		Calculated field	How many known individuals came to the site THIS WEEK?	How many known individuals are in the MIP database?	Calculated field
	Identified: Donors already in MIP database who came to the site this week	Known: Donors already in MIP database who came to the site this week	Identified: Users who donated on the site for the first time since MIP started collecting data	Known: Users who donated on the site for the first time since MIP started collecting data	**Identified donors on the site THIS WEEK**	**Known donors on the site THIS WEEK**	**Identified: Total identified donors in the MIP database**	**Known: Total known donors in the MIP database**	Known: Percent of subscribers in the MIP database who clicked on an e-mail this week	**Known: Total known donors and/or e-mail newsletter subscribers who came to the site THIS WEEK**	**Known: Total known donors and/or e-mail newsletter subscribers in the MIP database**	Known: Percent of known individuals in the MIP database who came to the site this week
Validation	Should be less than Z	Should be less than AA	Will vary widely based on pledge drives. Track vs. pledge drive calendar (internal events).		Should equal T+V=X	Should be less than AA and equal U+W=Y. Should be less than or equal to the total number of donors in the Eloqua database.	Z should equal previous week Z plus V	AA should equal previous week AA plus W	M/Q = R should be less than 100%. May vary widely.	The total number of unique e-mail addresses that can be segmented into three groups: 1. People who have ONLY come to the site via one or more e-mail newsletters only. 2. People who have ONLY donated online. 3. People who have both come to the site via an e-mail newsletter AND who have donated online.		BB/CC = DD should be less than 100%. May vary widely.

Each unique e-mail address in the MIP database should match to an e-mail in the Eloqua e-mail and donor databases. Unique identifier is e-mail ad dress, so manual clean-up/merge-purge would be needed to reduce the number of duplicates/triplicates.

Appendix B: Texas Tribune URL and Page Title Analysis

Texas Tribune stories in the MIP data repository
Stories posted April 1-June 30, 2016
Does not include stories posted prior to April 1, 2016 that got pageviews April 1-June 30, 2016

Goal **How many stories were posted?**

Ideal: Each story has one unique URL and one corresponding page title.
You should be able to see a clean page title and then click on a link to go to the story.

A. **Thus, we ran a report with both page title and URL. Each row is a unique combination of page title and URL.**
Report query: What URLs first appeared during second quarter?

B. **However, each unique URL has one or many page titles, for various valid operational reasons.**

Row_ID	Page_Title_ID	Page_Title	URL_ID	URL
52	22	Abbott to Release Book on Life, Convention of States	13	/2016/04/04/abbott-writes-book-life-constitutional-convention
53	23	Abbott Writes Book on Life, Constitutional Convention	13	/2016/04/04/abbott-writes-book-life-constitutional-convention
54	24	Abbott Writes Book on Life, Convention of States	13	/2016/04/04/abbott-writes-book-life-constitutional-convention

Thus, to get an accurate story count, you should be able to run a report based on unique URLs.
To make the report useful, the TT will need to:

1. Select which of the duplicate page titles should be marked as the primary page title.
   This can be semi-automated but needs to be reviewed and finalized manually.

2. Combine the pageview and other data.
   This can be automated.

Some of the duplicate page titles may only have one or two pageviews, but we don't know if these are the rows that have comments, shares, and other custom metrics that we really want to capture.

1. Need to analyze past data to determine if rows with less than x pageviews have zero comments, shares and other custom metrics. If so, reports could exclude these rows, and there would be fewer duplicate page titles to manually mark and combine.

Appendix B: Texas Tribune URL and Page Title Analysis

	C.	There are "duplicate" URLs			
		1	This may be a problem with the way Google Analytics reports URLs.		
		2	This perhaps can be fixed with Google Tag Manager.		
		3	This may be a problem with the site architecture itself.		
	Row_ID	Page_Title_ID	Page_Title	URL_ID	URL
	2758	8	Page Not Found	789	/2016/06/10/citing-ong..
	2759	8	Page Not Found	789	/2016/06/10/citing-ongoing-issues-
	2760	8	Page Not Found	789	/2016/06/10/citing-ongoing-issues-state
	2761	8	Page Not Found	789	/2016/06/10/citing-ongoing-issues-state-
	2762	8	Page Not Found	789	/2016/06/10/citing-ongoing-issues-state-scraps-=
	2763	8	Page Not Found	789	/2016/06/10/citing-ongoing-issues-state-scraps-staar
	2764	1460	Falta sÃ-mbolo del Estado Ups para el desguace de la escuela Puntuaciones de prueba	789	/2016/06/10/citing-ongoing-issues-state-scraps-staar-scores
	2765	1460	Falta Ups sÃ-mbolo del Estado para el desguace de la escuela Puntuaciones de prueba	789	/2016/06/10/citing-ongoing-issues-state-scraps-staar-scores
	2766	1462	Foul Ups Prompt State to Scrap School Test Scores	789	/2016/06/10/citing-ongoing-issues-state-scraps-staar-scores
	2767	3	Login to Your Account	789	/2016/06/10/citing-ongoing-issues-state-scraps-staar-scores
	2768	8	Page Not Found	789	/2016/06/10/citing-ongoing-issues-state-scraps-staar-scores
	2769	4	Null	789	/2016/06/10/citing-ongoing-issues-state-scraps-staar-scores

Appendix B: Texas Tribune URL and Page Title Analysis

	D.	Some "duplicate" page titles have nonwords.		
		1	We can automate removing nonwords that are repeatedly and consistently used.	
	Row_ID	**Page_Title_ID**	**Page_Title**	**URL_ID**	**URL**
	59	25	Analysis: Texas Government's Shrinking Financial Buffer	14	/2016/04/04/analysis-texas-govts-shrinking-financial-buffer
	60	26	L1 - Analysis: Texas Government's Shrinking Financial Buffer	14	/2016/04/04/analysis-texas-govts-shrinking-financial-buffer
	228	101	54.7 MB: UT Steps Up Security After Body of Student Found on Campus	63	/2016/04/07/ut-steps-security-after-body-student-found-campus
	229	102	58.3 MB: UT Steps Up Security After Body of Student Found on Campus	63	/2016/04/07/ut-steps-security-after-body-student-found-campus
	230	103	61.7 MB: UT Steps Up Security After Body of Student Found on Campus	63	/2016/04/07/ut-steps-security-after-body-student-found-campus
	231	104	62.0 MB: UT Steps Up Security After Body of Student Found on Campus	63	/2016/04/07/ut-steps-security-after-body-student-found-campus
	232	105	72.7 MB: UT Steps Up Security After Body of Student Found on Campus	63	/2016/04/07/ut-steps-security-after-body-student-found-campus

Appendix B: Texas Tribune URL and Page Title Analysis

	E.	**Some "duplicate" page titles are garbled, preventing completely automated matching. Only some of the symbols could be filtered automatically, so the page titles would need to fixed manually.**			
	Row_ID	**Page_Title_ID**	**Page_Title**	**URL_ID**	**URL**
	92	37	Emails Show Abbott\u2019s Involvement at Child Welfare Agency	24	/2016/04/05/emails-show-abbotts-involvement-child-welfare-agen
	93	38	Emails Show Abbottâ€™s Involvement at Child Welfare Agency	24	/2016/04/05/emails-show-abbotts-involvement-child-welfare-agen
	F.	**Some URLs have no page titles, or are null.**			
		See the example in C. Duplicate URLs.			
		1	Need to analyze past data to determine if rows with null page titles have less than x pageviews and less than x number of zero comments, shares and other custom metrics. If so, reports could exclude these rows, and there would be fewer duplicate page titles to manually mark and combine.		
	G.	**Some URLs and recurring page titles don't give any information on the story topic.**			
		1	This hampers tagging stories and adding other metadata.		
		2	This could be corrected with site design and would be better SEO.		
	Row_ID	**Page_Title_ID**	**Page_Title**	**URL_ID**	**URL**
	84	35	The Bookshelf: April 6, 2016	21	/2016/04/05/bookshelf-april-6-2016