Understanding Illinois Legal Aid Online's Users and their Needs

Team

Analysts: Jess Ray, João Caldeira, Nathan Bartley, Nupoor Gandhi
Project Manager: Emma Remy

Background on ILAO

Illinois Legal Aid Online (ILAO) is a small, non-profit organization that offers free, 24/7, online tools in English, Spanish and Polish, so that people can use the law to get the justice they seek.

ILAO's services and tools are designed for both motivated learners, who are considering taking action, and people in crisis, who may be thrust unwillingly into a situation that is beyond their control. Legal help is generally around sensitive topics like domestic abuse, divorce, debt, eviction, deportation, and access to medical care. Most recently, it is about unemployment and food stamp benefits, child support changes, and money/debt problems.

Project goals

The goal of the project was to better understand who ILAO's users are and what is being offered to them that is most helpful and least helpful to them. To date, ILAO relied exclusively on anecdotal feedback, user interviews, and Google Analytics to patch together a profile of who is using the website and how it helps. This allows ILAO to have an accurate view of how many people are viewing which pages at what time, but does not connect different sessions from the same user. With this project and the more detailed data from Acquia used in it, ILAO will be able to understand usage of the site in a more specific way by understanding user journeys throughout different sessions spaced apart in time.

In particular, ILAO wanted to identify patterns in usage that may not have been obvious to them otherwise. As a result of this analysis, ILAO will be able to add design elements to the website that will make it easier for users to get access to the resources they need before they know that they need them. For example, if the analysis found that users who look at information about unemployment frequently later search for information about eviction, ILAO could add a link to information about eviction on the page about unemployment. As a result, users might be able to better prepare for that possible future outcome. ILAO will conduct A/B testing to confirm the efficacy of these links after adding them according to our recommendations.

Timeline

May 2020 - Initial team is recruited; project begins with legal paperwork

August 2020 - Legal paperwork completed, data exploration begins

September 2020 - Cleaning data for loading, beginning exploratory data analysis (EDA)

October-November 2020 - Continuing to troubleshoot data that doesn't load with limited RAM or is missing documentation; continued EDA

December 2020-January 2021 - Generating maps and other descriptive metrics and graphics demonstrating the types of users ("professional" vs. "non-professional" users). Presented at ILAO's January board committee meeting

February 2021 - Generating tables of common pairings of pages visited, presented at ILAO's regular board committee meeting in March. ILAO prepares for A/B testing based on preliminary results.

March-April 2021 - Cleaning code and finalizing results; preparing for QA.

How we accessed the data

ILAO's data is stored on their own servers, and we accessed it using remote desktop software to connect to an ILAO machine, where we conducted all of our analysis. None of ILAO's data was ever saved in any other location, save a few small samples and summary tables and graphs. Only one user could access the remote desktop at a time, so we used Slack to coordinate access and ensure that no one conflicted.

The data

The data for this project was collected using Acquia Lift from March 1, 2019 through May 1, 2020 and is stored in a series of .tsv files as described here:

Touch. Table for each session of the user, including platform, date, duration, location.

Event. Table for each event within a specific touch, including the specific type of event, page url, language, touch.

Person. Table for each person object including their first touch timestamp, last touch timestamp, whether they are anonymous, engagement score, among others

Person Identifier. Table identifying individual persons with information like email addresses, if available.

Person Ranking. Table containing the raw metadata for each person and their highest ranked content sections, engagement score, etc.

Person Ranking Item. Table containing each item which is ranked for each person_id, describing the frequency and rank of each item

Person Ranking Summary. Joined table of *Person Ranking* and *Person Ranking Item* summarizing each ranked item for each person.

Segment. Supplied labels for different types of users observed by Acquia.

Matched Segment. Matched population segment labels for different person ids.

Data cleaning and transformation

In some files, particularly the touch and event tables, some of the rows contained new line characters and tab characters which were part of the data fields. This made it impossible to directly open the files with a package like dask, which uses new line characters to partition a csv file into chunks that fit on memory individually. Some effort was necessary to remove those characters, either manually or with a script, before working with those larger tables.

We originally had 25.4 million unique touches covering 20.3 million persons, which was unwieldy to analyze given the total number of columns. We observed that 19.6 million of the 20.3 million persons had only one touch in the dataset, so we decided to filter the data to only refer to person_id's that had more than one touch as we were mostly interested in the experience of users who used the website more than once. This allowed us to cut down significantly on the total size of the data, allowing some tables to be read completely into memory.

We then identified several columns that were not clearly described in the data dictionary, and given that they had missing values for many of the persons, we were able to filter and ignore those columns in analysis.

At this point, the data contained 5.8 million touches. However, we noticed that many of these touches referred to the same person and were less than one second apart, and so could not truly correspond to different user sessions. We deduplicated these touches, keeping the record

with the maximum information and claimed number of events (usually, all but one of the duplicated touches claimed one event and had very little metadata). This reduced the sample to 2.1 million touches from 0.6 million persons.

The event table originally contained 71.2 million events. We selected the events corresponding to touches in the selected sample above, resulting in 19.4 million total events.

Accessing cleaned data

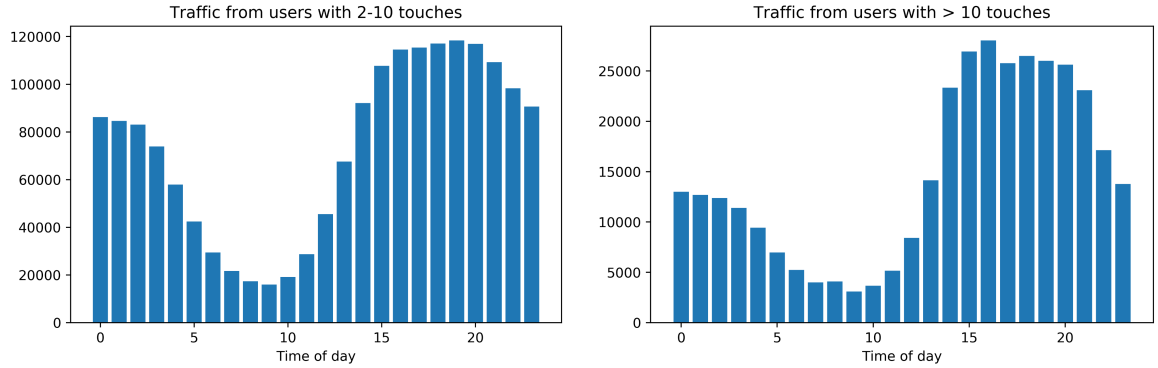
The following cleaned data files were generated over the course of the analysis, and are accessible from the same location as the original data files -- please refer to the original data dictionary for details about the contents of the original files:

- *touch_fixed*: Same as the original touch file, without tab and new line characters in data fields.
- *event_fixed*: Same as the original event file, without tab and new line characters in data fields. Transformation in clean_events.py.
- *touch_multiple_deduped:* Created from touch_fixed by selecting persons with more than one touch and deduplicating touches by the same person within one second of each other as found here.
- *person_ids_appear_more_than_once*: IDs of the persons that have more than one touch, as created here.
- *event_professionals*: events corresponding to touches in touch_multiple_deduped and persons whose majority (>50%) of touches are between 13:00 and 22:00 UTC and on a desktop, as created here. Holidays not taken into account.
- *event_non_professionals*: events corresponding to touches in touch_multiple_deduped and persons not part of event_professionals, as created here.
- *event_desktop_ppl*: events corresponding to touches in touch_multiple_deduped and persons whose majority (>50%) of touches are on a desktop, as created here.
- *event_mobile_ppl*: events corresponding to touches in touch_multiple_deduped and persons not part of event_desktop_ppl, as created here.
- *event_desktop_ppl2, event_mobile_ppl2, event_professionals2*: events whose persons are present in touch_multiple_deduced - i.e., similar to the versions without the number 2, but selected on person IDs rather than touch IDs. Made to check whether touches with 0 events were caused by our filtering, which turned out not to be the case, so these can be ignored. Created here.
- *people_business_desktop*: persons whose majority (>50%) of touches are between 13:00 and 22:00 UTC and not on a weekend or holiday in Illinois and on a desktop, as created here.
- *{event,touch}_{mobile,desktop}_{business,night}*: Created here from touch_multiple_deduped, event_desktop_ppl, and event_mobile_ppl. Desktop is from persons whose majority (>50%) of touches are on a desktop, and mobile is the complement of that set of persons. Business is from persons whose majority (>50%) of

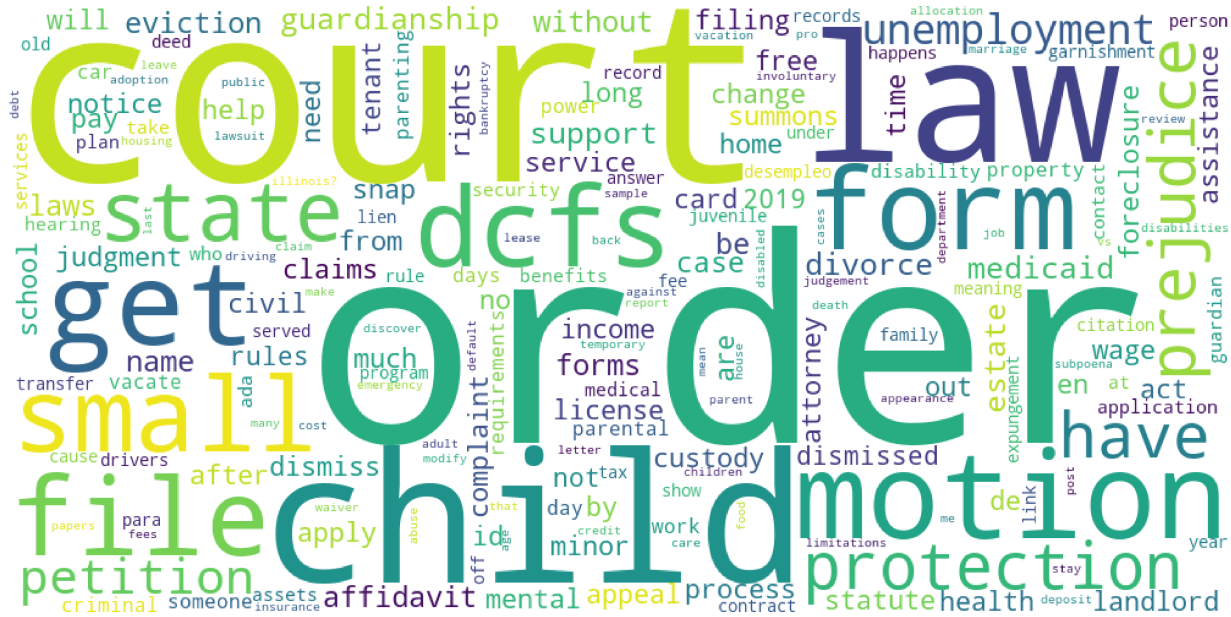
touches are between 13:00 and 22:00 UTC and not on a weekend or holiday in Illinois, and night is the complement of that.

Exploratory/summary metrics on the data

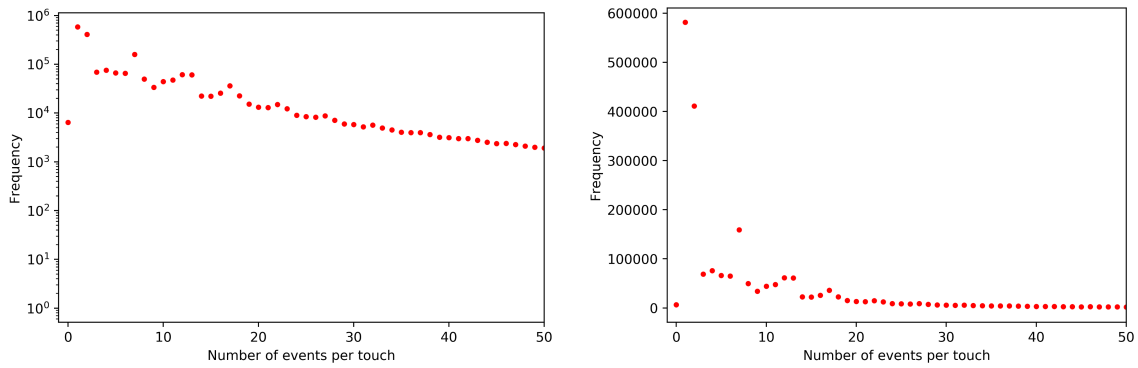
One initial hypothesis we wanted to corroborate was that professional users are more likely to make more visits to the website, and also more likely to visit during business hours. Then we should see that the traffic from users with more touches comes mostly during business hours, and this is indeed the case (note time of day is in UTC):



Each touch is associated with the search terms that led the user to the site. This allowed us to generate a word cloud (after excluding generic common words in English and direct references to ILAO):



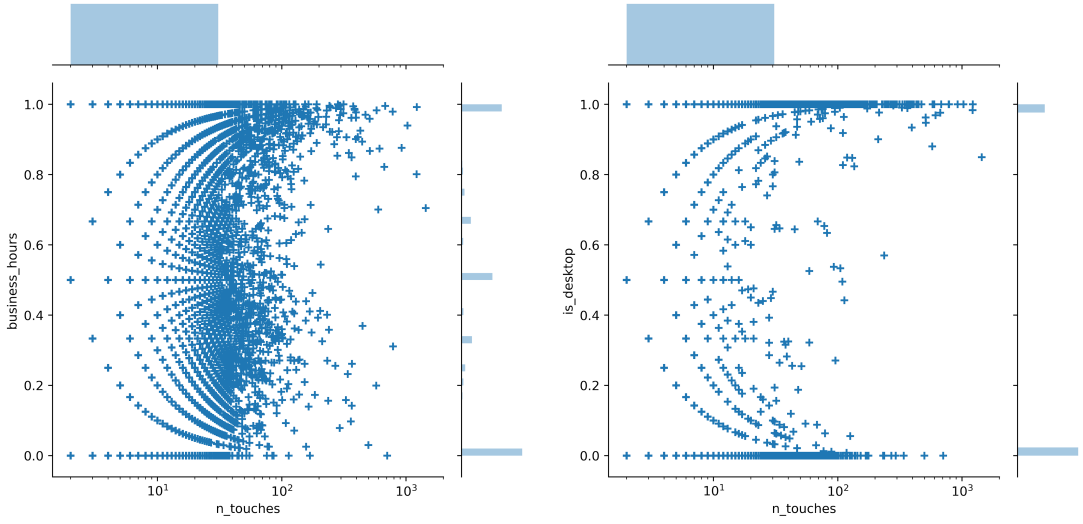
One interesting fact is related to the number of events per touch. We found a regular exponential distribution, with some outliers:



A particularly interesting outlier is events with 7 touches, much more likely than any number between 3 and 6. This most likely corresponds to a path of Get legal help -> triage -> program triage -> about -> income -> address -> confirmation.

Below, we will partition the data using two dimensions which we believed were good selectors for the type of user each person is: whether more than 50% of their touches were made from a desktop computer (as opposed to mobile, tablet or other), and whether more than half of their touches were made during business hours (defined as between 13:00 and 22:00 UTC on weekdays). A plot of how all persons in the dataset line up relative to these two measures is shown below.

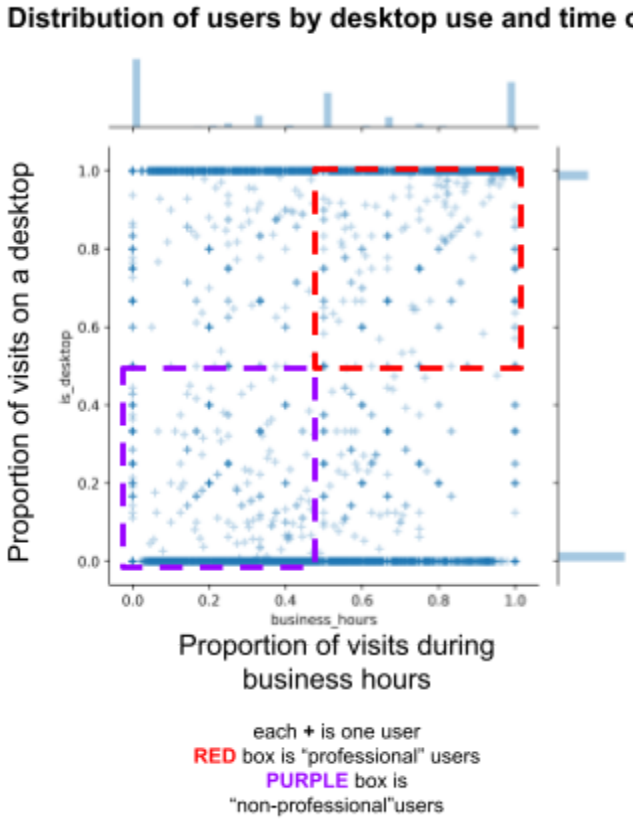
It is interesting once again to see how the two dimensions correlate with the number of touches from each given user:



Recall that we expect professional users to have a larger number of touches, and to access the website mostly during business hours from a desktop platform. We see that indeed most of the users with a high number of touches are primarily desktop business-hours users, though this is not a perfect association.

Identification and comparison of different user groups

Given the platform type (mobile vs desktop), and the times which the website in general were accessed, we were able to identify groups of users that seem to use the website on a desktop during regular business hours, and groups of users that seem to use the website primarily on mobile devices outside of regular business hours.



In this graphic, each "+" indicates one user. For each user, we looked at all of their visits to the website and calculated the proportion of those visits that were during business hours (x axis) and the proportion of those visits that were from a desktop, rather than a mobile device (y axis). We used the 50% threshold for each of these axes to roughly separate the users into four groups, two of which we focused on: desktop + business hour users (purple box) and mobile + non-business hour users (red box). We use the set of users in the purple box to represent "professional" users (people who use ILAO resources for work or volunteer positions to help clients) and the red box to represent "non-professional" users (who use ILAO resources for themselves or their friends/families.

Below are some metrics on the two groups of users, focusing on the timing of their visits to the website (created in Tableau).

Days Between Users First and Last Event Dates

	Desktop Business Hours	Mobile Evening
Average	42	21
Median	11	2
Maximum	430	426
Minimum	0	0
Number of Users	98335	307435

Days Between Users Individual Event Dates

	Desktop Business Hours	Mobile Evening
Average	22	19
Median	7	3
Maximum	396	409
Minimum	0	0
Number of Users	98335	307435

We repeated the analysis leading to the word cloud above by user type, and found different patterns for these different groups:

Search terms used to find ILAO's resources

Professional users
(primarily on desktops + use the site during business hours)

Non-professional users
(primarily on mobile + use the site outside business hours -- nighttime)

Within these segments of users we can then identify the content sections of the pages that the users travel from and to. For instance, the typical content sections visited (mobile users on the left, desktop and business hours users on the right:

('Procedure', 128309),

('Unemployment Compensation', 57795),

('Immigration/Naturalization', 57134),

('Food stamps', 56666),

('Divorce/Separation/Annulment', 42215),

('Abuse', 37723),

('Custody and visitation', 35420),

('Landlord/Tenant,Eviction/Lockouts', 30521),

('Child support', 27196),

('Procedure', 105887),

('Divorce/Separation/Annulment', 35812),

('Custody and visitation', 28002),

('Abuse,Domestic violence', 23865),

('Unemployment Compensation', 23725),

('Landlord/Tenant,Eviction/Lockouts', 20445),

('Child support', 18395),

('Wills and estates', 14053),

('Criminal records', 13591),

We can also look at content section pairs that are most commonly visited together by users (here we exclude "Procedure", as that is a generic content section not related to any particular legal issue):

(('Child support', 'Custody and visitation'), 1236),

(('Food stamps', 'Unemployment Compensation'), 1216),

(('Abuse,Domestic violence', 'Domestic violence'), 1070),

(('Custody and visitation', 'Divorce/Separation/Annulment'), 1038),

(('Custody and visitation', 'Paternity'), 983),

(('Child support', 'Divorce/Separation/Annulment'), 894),

(('Abuse', 'Custody and visitation'), 798),

(('LIHEAP,Public utilities', 'Public utilities'), 731),

(('Abuse,Domestic violence', 'Custody and visitation'), 718),

(('Custody and visitation', 'Divorce/Separation/Annulment'), 1794),

(('Child support', 'Custody and visitation'), 1565),

(('Child support', 'Divorce/Separation/Annulment'), 1464),

(('Custody and visitation', 'Paternity'), 1163),

(('Landlord/Tenant', 'Landlord/Tenant,Eviction/Lockouts'), 1005),

(('Abuse,Domestic violence', 'Custody and visitation'), 897),

(('Abuse,Domestic violence', 'Divorce/Separation/Annulment'), 864

(('Criminal records', 'Other employment'), 853),

(('Custody and visitation', 'Minor guardianship'), 804),

If we want to look at content sections that are visited by people at different points in time, we can specify the content section visits to be, for instance, more than 7 days apart:

(('Unemployment Compensation', 'Food stamps'), 406),
(('Food stamps', 'Unemployment Compensation'), 404),
(('Custody and visitation', 'Child support'), 348),
(('Child support', 'Custody and visitation'), 340),
(('Abuse', 'Custody and visitation'), 287),
(('Custody and visitation', 'Divorce/Separation/Annulment'), 281),
(('Divorce/Separation/Annulment', 'Custody and visitation'), 281),
(('Custody and visitation', 'Abuse'), 276),
(('Divorce/Separation/Annulment', 'Custody and visitation'), 860),
(('Custody and visitation', 'Divorce/Separation/Annulment'), 847),
(('Divorce/Separation/Annulment', 'Child support'), 694),
(('Child support', 'Divorce/Separation/Annulment'), 684),
(('Custody and visitation', 'Child support'), 671),
(('Child support', 'Custody and visitation'), 671),
(('Custody and visitation', 'Abuse,Domestic violence'), 527),
(('Divorce/Separation/Annulment', 'Abuse,Domestic violence'), 512)

We then calculate for each content section pair the proportion of users for which that pair appears, and find the pairs where that proportion is most different between the two user groups:

	counts_mobile	counts_business	mobile_ratio	business_ratio	diff
(Wills and estates, Divorce/Separation/Annulment)	13	299	0.000294	0.001381	-0.001087
(Divorce/Separation/Annulment, Wills and estates)	23	336	0.000520	0.001551	-0.001032
(Landlord/Tenant,Eviction/Lockouts, Wills and estates)	22	323	0.000497	0.001491	-0.000994
(Wills and estates, Landlord/Tenant,Eviction/Lockouts)	21	316	0.000474	0.001459	-0.000985
(Custody and visitation, Wills and estates)	18	295	0.000407	0.001362	-0.000955
...	...	...	...	...	...
(Custody and visitation, Child support)	348	671	0.007861	0.003098	0.004763
(Custody and visitation, Abuse)	276	257	0.006234	0.001187	0.005048
(Abuse, Custody and visitation)	287	229	0.006483	0.001057	0.005425
(Food stamps, Unemployment Compensation)	404	175	0.009126	0.000808	0.008318
(Unemployment Compensation, Food stamps)	406	151	0.009171	0.000697	0.008474

Final analysis methodology

Content Section Pairings

Using Tableau and the deduplicated tables filtered to mobile nighttime and desktop business hours, we created sets of users that viewed the same content sections. We also created a parameter of grouped content sections to use as a dynamic filter on sets. Using these resources, we excluded the content section of interest (displayed in the parameter) and displayed the other content sections and pages the user set also viewed.

Sample table (see results document for tables corresponding with other content areas):

Mobile Nighttime Users: Top 15 Pages from Other Content Sections Viewed by People Who Looked at **Public Benefits** Pages

Content Title	Other Content Section	Number of Users
Applying for unemployment benefits	Unemployment Compensation	341
Am I eligible for Medicaid?	Health, public benefits	262
Getting unemployment benefits	Unemployment Compensation	231
Can I receive unemployment benefits if I work a temp job?	Unemployment Compensation	134
Illinois State ID basics	Traffic	131
Application for a person with a disability ID card - blank	Traffic	128
DCFS cases and child protection services	Safety/Abuse	128
Application for unemployment benefits - online	Unemployment Compensation	121
Understanding DCFS investigations	Safety/Abuse	118
The low income home energy assistance program (LIHEAP)	Utilities	115
If I quit my job, can I get unemployment benefits?	Unemployment Compensation	108
Presentar una solicitud para beneficios de desempleo	Unemployment Compensation	105
Getting a divorce	Family (divorce)	96
3 types of orders of protection	Safety/Abuse	92
Applying for Medicaid	Health, public benefits	89

Note: excludes “procedural” content

Mapping methodology

Using Tableau, combined the segmented deduplicated data files with a spatial join to a shapefile of Illinois counties acquired from the Illinois Geospatial Data Clearinghouse. Due to the wide variance of users per county, decided to review the distribution of users across the segments and determined buckets to more meaningfully color code the counties by number of users. Mapped each dataset using the Tableau generated latitude and longitude and grouped color legend.

Conclusions and next steps

Throughout the course of this project, we struggled to identify results that were both unique to our analysis (that is, results not available through Google Analytics) and useful new insights for the experts at ILAO. In order to try and identify sets of pages that wouldn’t already have existing links but also were frequently visited by the same users, we created many of the tables by content area discussed previously based on a list of content areas recommended by ILAO.

From here, ILAO used their expertise to help identify pages that might be good candidates for our planned intervention: adding new links. Here are some specific connections ILAO would like to investigate:

- Public benefits -> state ID
- Debt -> Car issues (repossession / impounded)
- Rental housing -> Foreclosure

- Safety/abuse -> Family (divorce + custody issues)
- Crime pages -> DCFS content
- Eviction -> Utility shut off, DCFS cases

Using these results, ILAO will perform pilot A/B tests on page pairings to either add a “Users who looked at this page also looked at this other page” or update the existing recommended/related block to include these newly identified page pairings. ILAO has significant experience conducting these kinds of A/B tests using Google Optimize, and will evaluate success using a combination of metrics including: increase in session duration, number of pages accessed, decrease in bounce rates for users who viewed the updated pages vs. the original pages.

This analysis focused on the content area of each page, the number of users who visited each page, and the time and device type of each access. Further analysis of this data could include focusing on specific geographical areas to get more granular results about different types of users based on their locations, including cross-referencing those locations with the positions with the locations of publicly available resources such as libraries. This would get at more specifics of the different ways different types of users utilize the site. Another method of identifying more granular groups of users would be using the demographics that are provided for users with ILAO accounts. Perhaps some patterns of usage could be identified focusing specifically only on these users. Continuing in the current vein of results, the tables of page/content area pairings could be further extended: were there other overlapping pages that users also viewed?

Another particularly interesting follow-up would be to try to quantify more precisely certain trends in how users return to the website after visiting certain pages. The type of question we would like to answer can be formulated as: if a user visits page x, how likely is that user to return and visit page y, and how much time is likely to pass between the two visits? A well-structured answer would allow ILAO to make better recommendations of the other pages that their users should visit.

It would also be interesting to investigate whether shifts in traffic are visible around the start of the COVID-19 pandemic, as the data period goes until May 1, 2020. It is likely that some legal issues became more pressing in those months as compared to the period before, and that analysis could be valuable to inform interventions while the pandemic is still ongoing as well as after.