Counting on CONTENTdm

Deciphering Statistics and Looking for Alternatives

8 November 2010
Upper Midwest CONTENTdm Users Group Meeting
Part 1: Surprise Surprise

Eric Celeste, efc@clst.org
Consultant to the Minnesota Digital Library
Stumbled upon this issue as part of Flickr+MDL

We have the same collection on CONTENTdm and at Flickr, let's compare usage...
View: Server statistics | Usage reports — November 2010

Item view summary

Review the number of items viewed per collection during the selected month. Click total views per item to see item details.

November 2010
Generated: 03-Nov-2010 02:26 Eastern Daylight Time

<table>
<thead>
<tr>
<th>Item views per collection</th>
</tr>
</thead>
<tbody>
<tr>
<td>508</td>
</tr>
<tr>
<td>390</td>
</tr>
<tr>
<td>331</td>
</tr>
<tr>
<td>204</td>
</tr>
<tr>
<td>152</td>
</tr>
<tr>
<td>136</td>
</tr>
<tr>
<td>106</td>
</tr>
<tr>
<td>104</td>
</tr>
<tr>
<td>98</td>
</tr>
<tr>
<td>90</td>
</tr>
<tr>
<td>78</td>
</tr>
<tr>
<td>74</td>
</tr>
<tr>
<td>74</td>
</tr>
<tr>
<td>74</td>
</tr>
<tr>
<td>70</td>
</tr>
<tr>
<td>63</td>
</tr>
<tr>
<td>61</td>
</tr>
<tr>
<td>56</td>
</tr>
<tr>
<td>52</td>
</tr>
<tr>
<td>46</td>
</tr>
<tr>
<td>37</td>
</tr>
<tr>
<td>33</td>
</tr>
</tbody>
</table>
Stats for: Your account
Stats are shown in GMT, where it's currently 5.24PM on 4th June 2010 (Fri)

Daily aggregate views on your account

View counts

<table>
<thead>
<tr>
<th></th>
<th>So far today</th>
<th>Yesterday</th>
<th>All time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Photos and Videos</td>
<td>47</td>
<td>114</td>
<td>12,294</td>
</tr>
<tr>
<td>Photostream</td>
<td>23</td>
<td>81</td>
<td>4,857</td>
</tr>
<tr>
<td>Sets</td>
<td>0</td>
<td>3</td>
<td>613</td>
</tr>
<tr>
<td>Collections</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Galleries</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

70                      | 198          | 17,764    |
Why the difference?

Crawlers, web spiders, mostly our own

Describe my analysis
Show what you can do
Discuss the OCLC response
August 2010

400,199 total item views reported but no other details to analyze
August 2010

raw logs have lots of detail, but report a much different number

cat ex*.log | wc -l
1,414,482 total hits

cat ex*.log | grep "GET /cdm4/item_viewer.php CISOROOT="/" | wc -l
381,483 hits on image views

cat ex*.log | grep "GET /cdm4/item_viewer.php CISOROOT="/" | grep -v "DMSCALE" | wc -l
367,459 image views without scaling*

*missing compound objects
Eliminate the crawlers...

cat ex*.log | grep "GET /cdm4/item_viewer.php CISOROOT=/" | grep -v "DMSCALE"
| grep -v "mdl-crawler"
| grep -v "Yahoo..Slurp"
| grep -v "ScoutJet"
| grep -v "msnbot"
| grep -v "DotBot"
| grep -v "Baiduspider"
| grep -v "Speedy+Spider"
| grep -v "YandexBot"
| wc -l

Our own crawler was our most aggressive visitor.

22,904 or 6% of our hits were human.

If we extrapolate, about 24% of your hits might be human.
What can you do?

Edit your `webalizer_user_options.conf` file.

```
IgnoreAgent bot
IgnoreAgent Bot
IgnoreAgent spider
IgnoreAgent Spider
IgnoreAgent crawler
IgnoreAgent Crawler
IgnoreAgent Slurp
IgnoreAgent ScoutJet
```

Only 13% of our hits look human after rerunning stats using this configuration.

You may find closer to 50%.
What can OCLC do?

Help you fix your configuration file.

Create a better default configuration.

Maintain a list of crawlers and update our configuration files for us.
Is this good enough?

Human? There is still plenty suspicious in those log files.

What is the purpose of these stats anyway? Comparison?

All web stats make vast assumptions. Only using the same tools can yield comparable results.
Part 2: Google Analytics

an apples to apples alternative