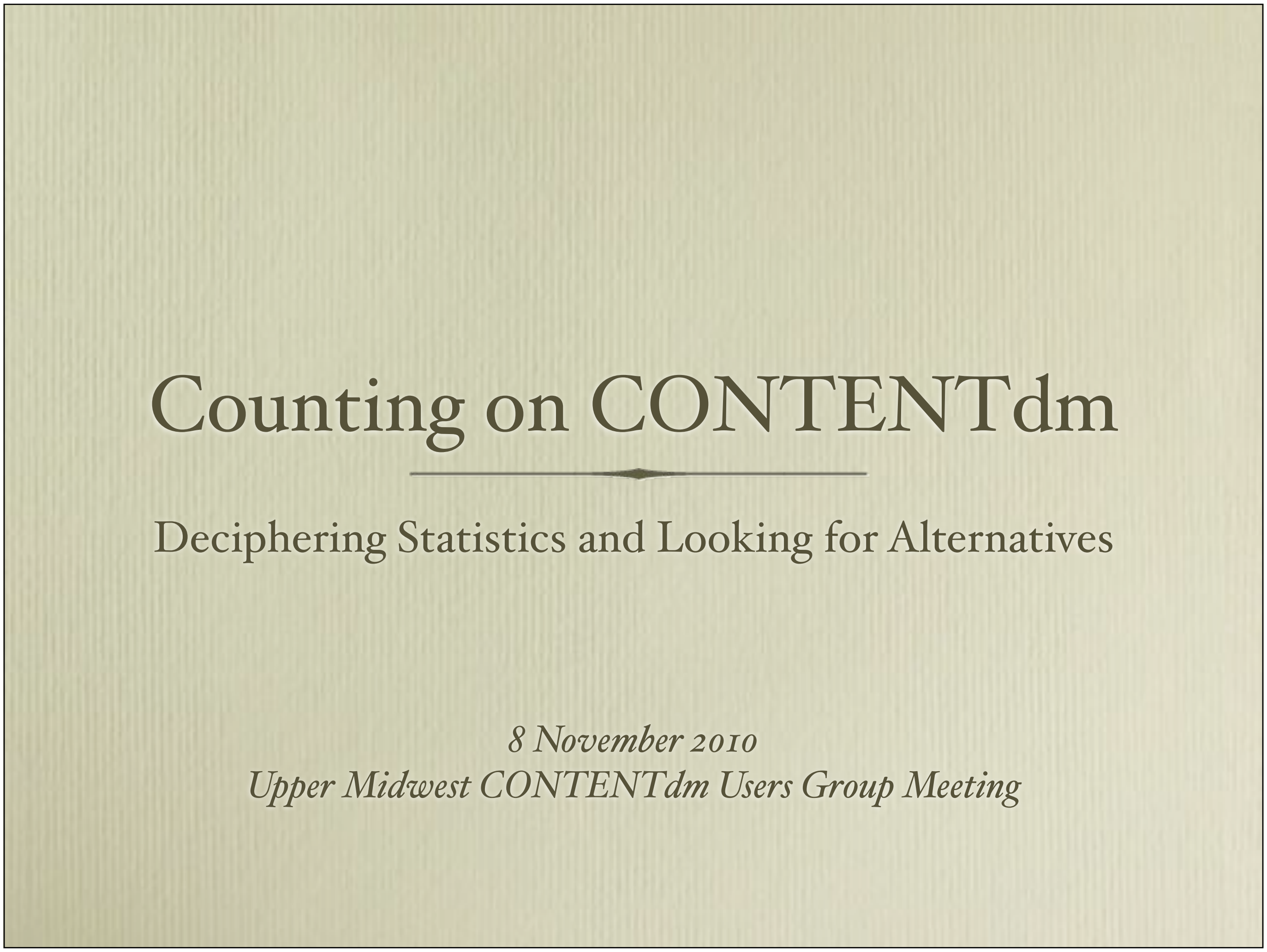
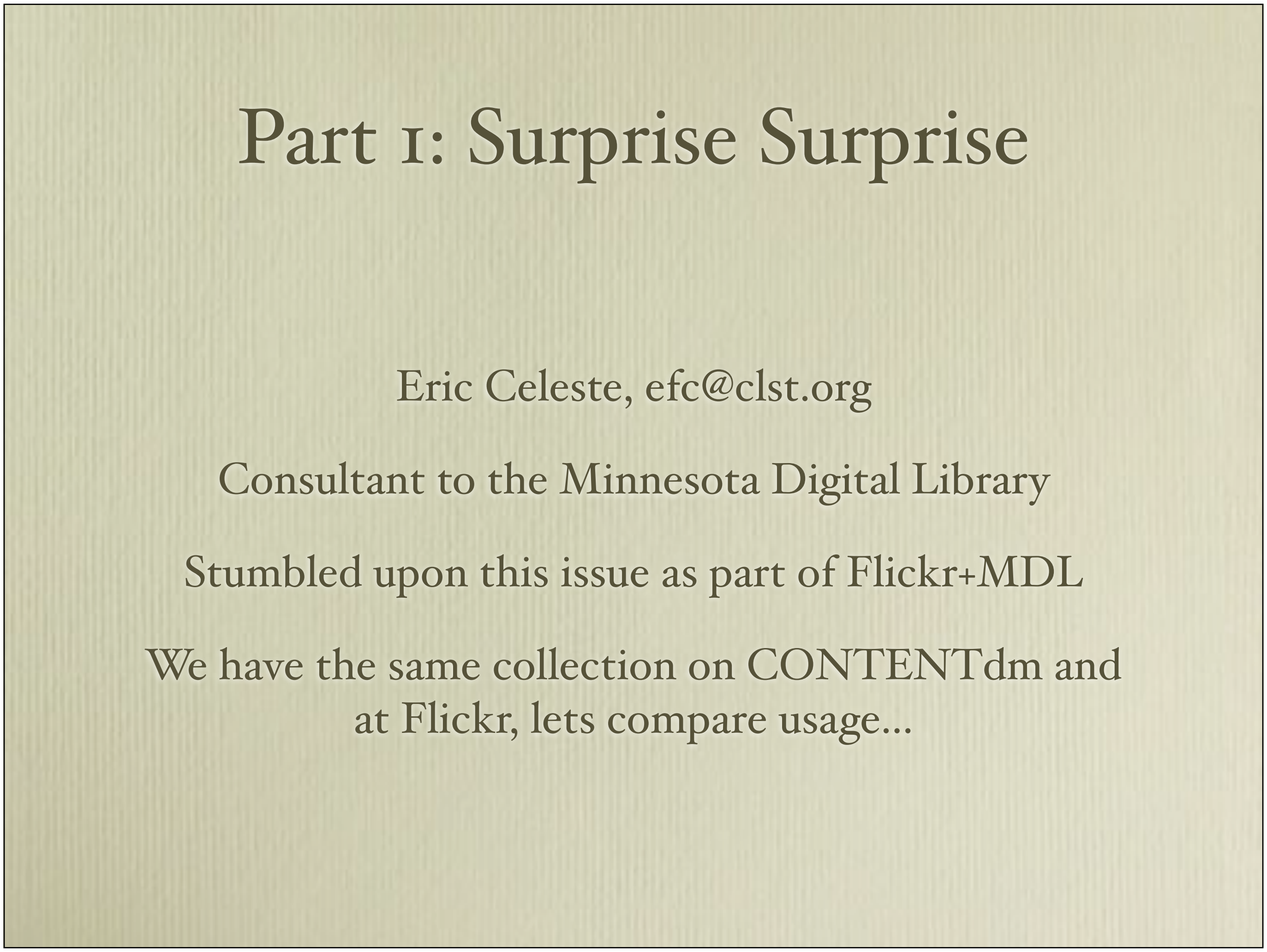


Counting on CONTENTdm

Deciphering Statistics and Looking for Alternatives

8 November 2010
Upper Midwest CONTENTdm Users Group Meeting

This was a presentation prepared for the Upper Midwest CONTENTdm Users Group Meeting, November 2010. You can reach me with questions at efc@clst.org and find more information about this issue at http://eric.clst.org/MDL/CONTENTdmStats.



Part 1: Surprise Surprise

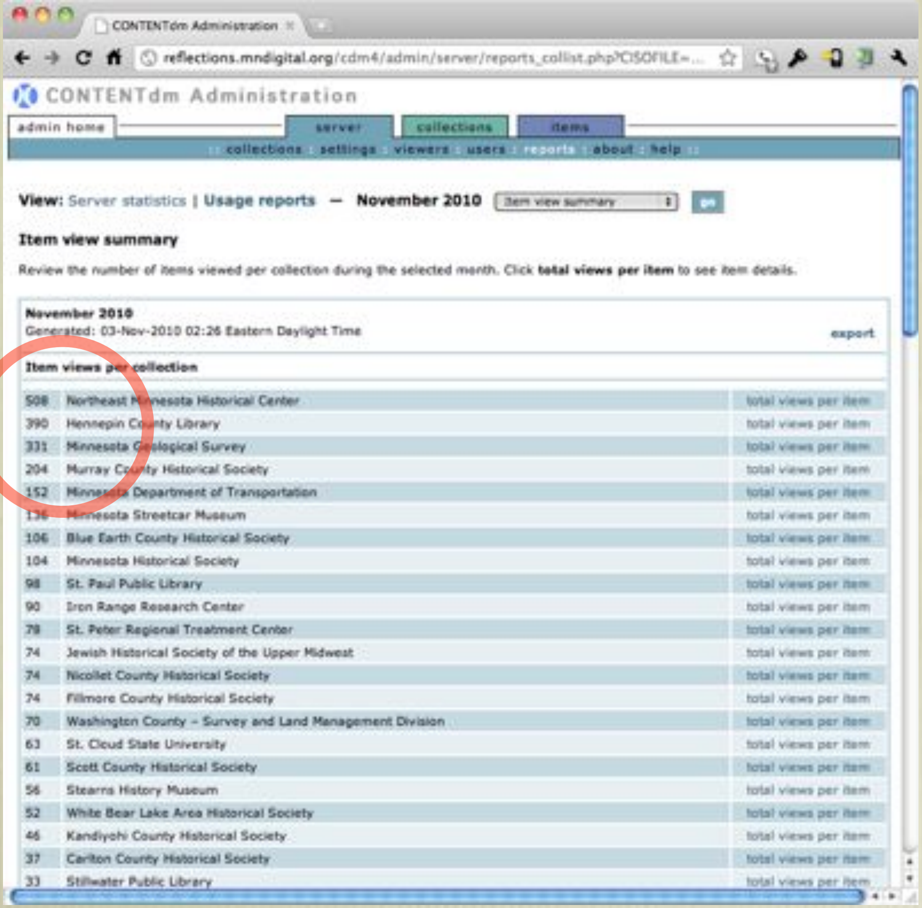
Eric Celeste, efc@clst.org

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, lets compare usage...

This work was done as part of my consulting work for the Minnesota Digital Library during 2009 and 2010.


CONTENTdm Administration
reflections.mndigital.org/cdm4/admin/server/reports_collist.php?CISOFILE=...
CONTENTdm Administration
admin home
server
collections
items
:: collections : settings : viewers : users : reports : about : help ::
Item view summary
go
export

View: Server statistics | **Usage reports** — **November 2010**

Item view summary

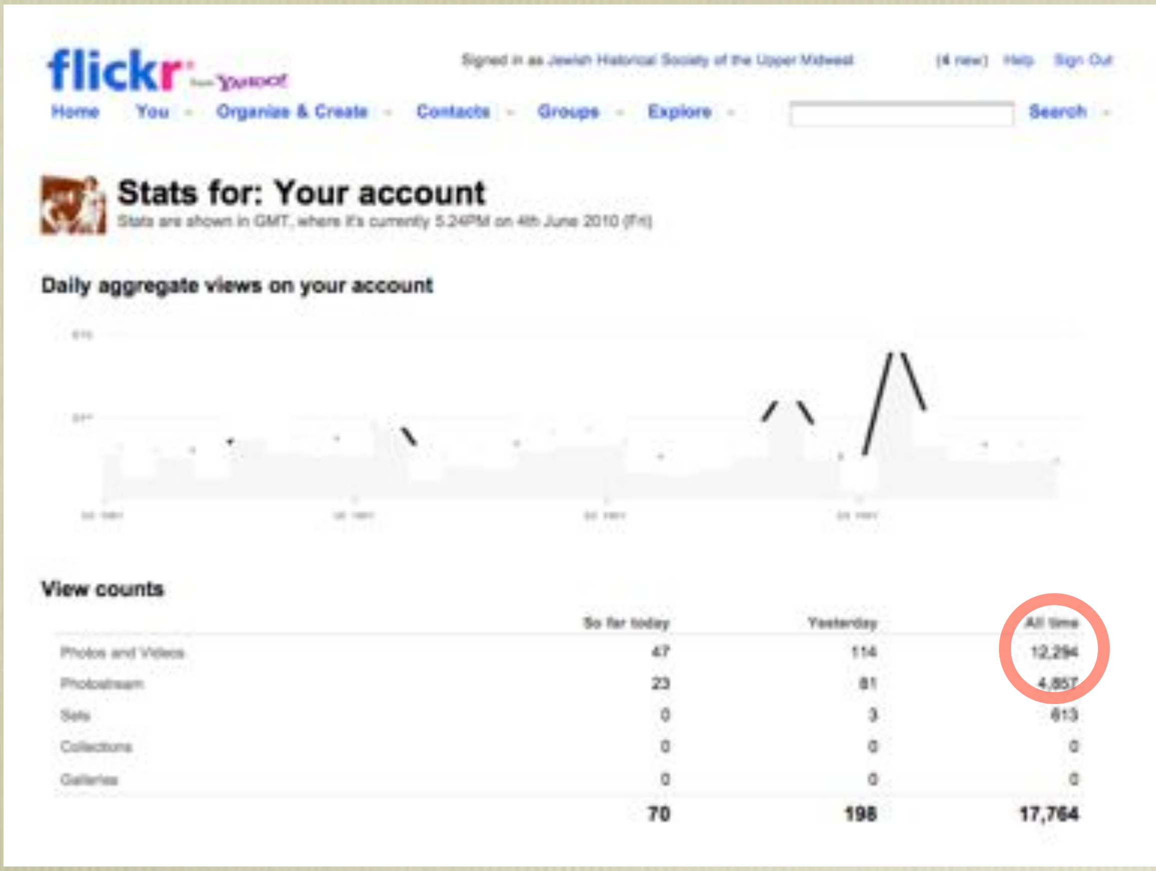
Review the number of items viewed per collection during the selected month. Click **total views per item** to see item details.

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

Item views per collection

508	Northeast Minnesota Historical Center	total views per item
390	Hennepin County Library	total views per item
331	Minnesota Geological Survey	total views per item
204	Murray County Historical Society	total views per item
152	Minnesota Department of Transportation	total views per item
136	Minnesota Streetcar Museum	total views per item
106	Blue Earth County Historical Society	total views per item
104	Minnesota Historical Society	total views per item
98	St. Paul Public Library	total views per item
90	Iron Range Research Center	total views per item
78	St. Peter Regional Treatment Center	total views per item
74	Jewish Historical Society of the Upper Midwest	total views per item
74	Nicollet County Historical Society	total views per item
74	Fillmore County Historical Society	total views per item
70	Washington County – Survey and Land Management Division	total views per item
63	St. Cloud State University	total views per item
61	Scott County Historical Society	total views per item
56	Stearns History Museum	total views per item
52	White Bear Lake Area Historical Society	total views per item
46	Kandiyohi County Historical Society	total views per item
37	Carlton County Historical Society	total views per item
33	Stillwater Public Library	total views per item

I first noticed that the statistics being reported for a given collection in a given month appeared quite high. The usage reports from CONTENTdm showed me what CDM thought the count was. Note that our Minnesota Digital Library CONTENTdm instance is hosted at OCLC and called Minnesota Reflections (http://reflections.mndigital.org).



Signed in as Jewish Historical Society of the Upper Midwest

Stats for: Your account

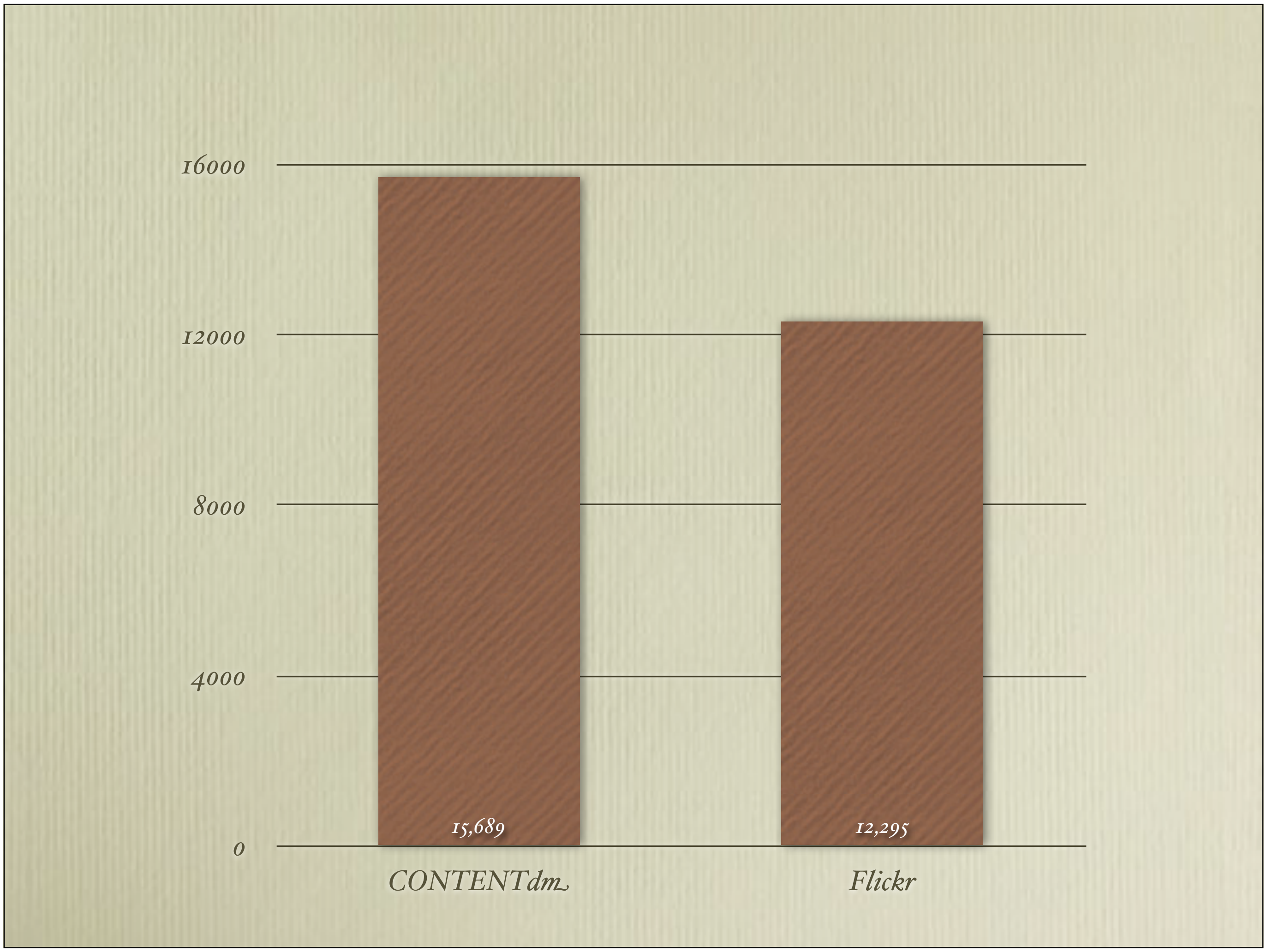
Stats are shown in GMT, where it's currently 5.24PM on 4th June 2010 (Fri)

Daily aggregate views on your account

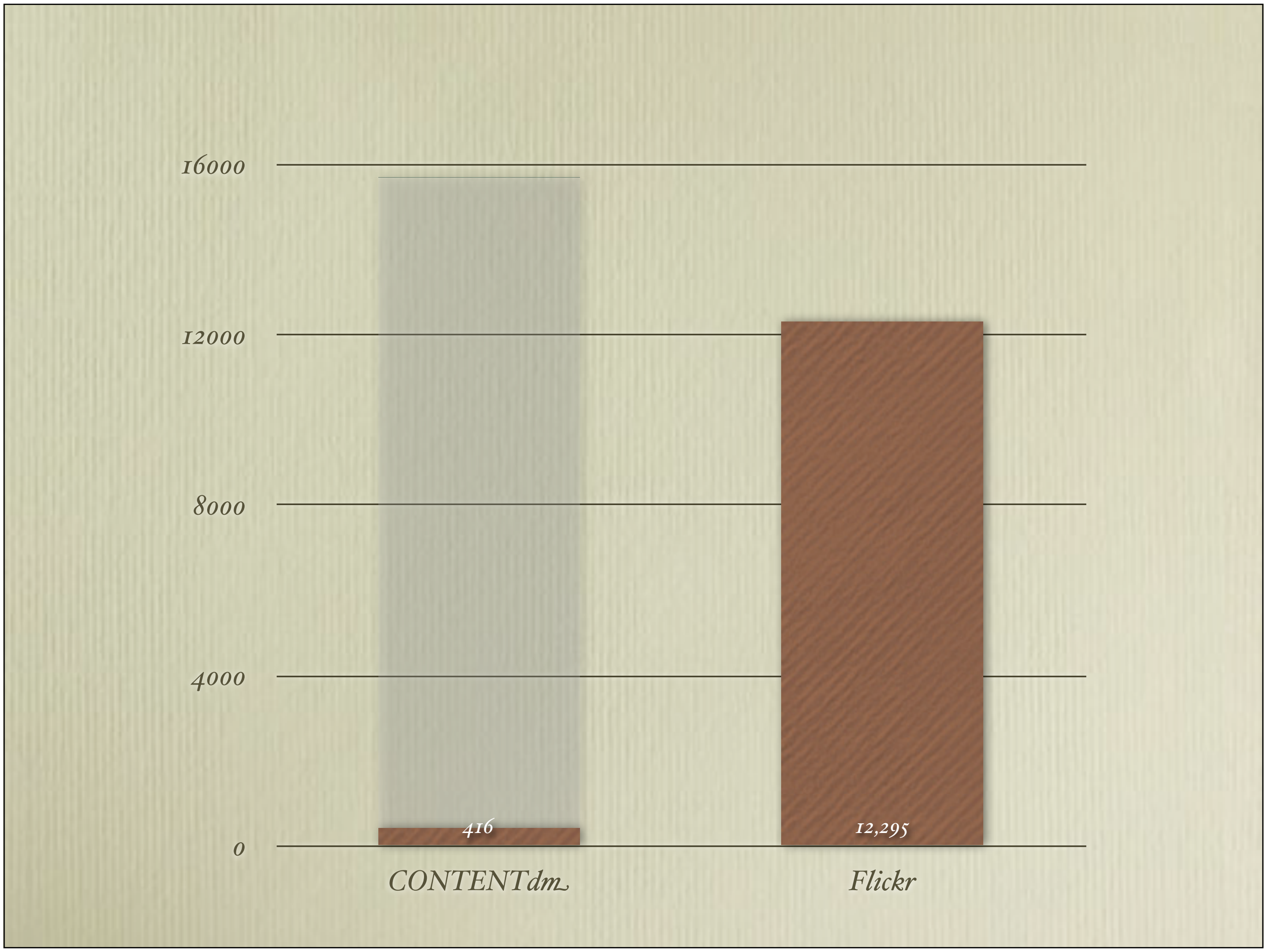
View counts

	So far today	Yesterday	All time
Photos and Videos	47	114	12,294
Photostream	23	81	4,857
Sets	0	3	613
Collections	0	0	0
Galleries	0	0	0
	70	**198**	**17,764**

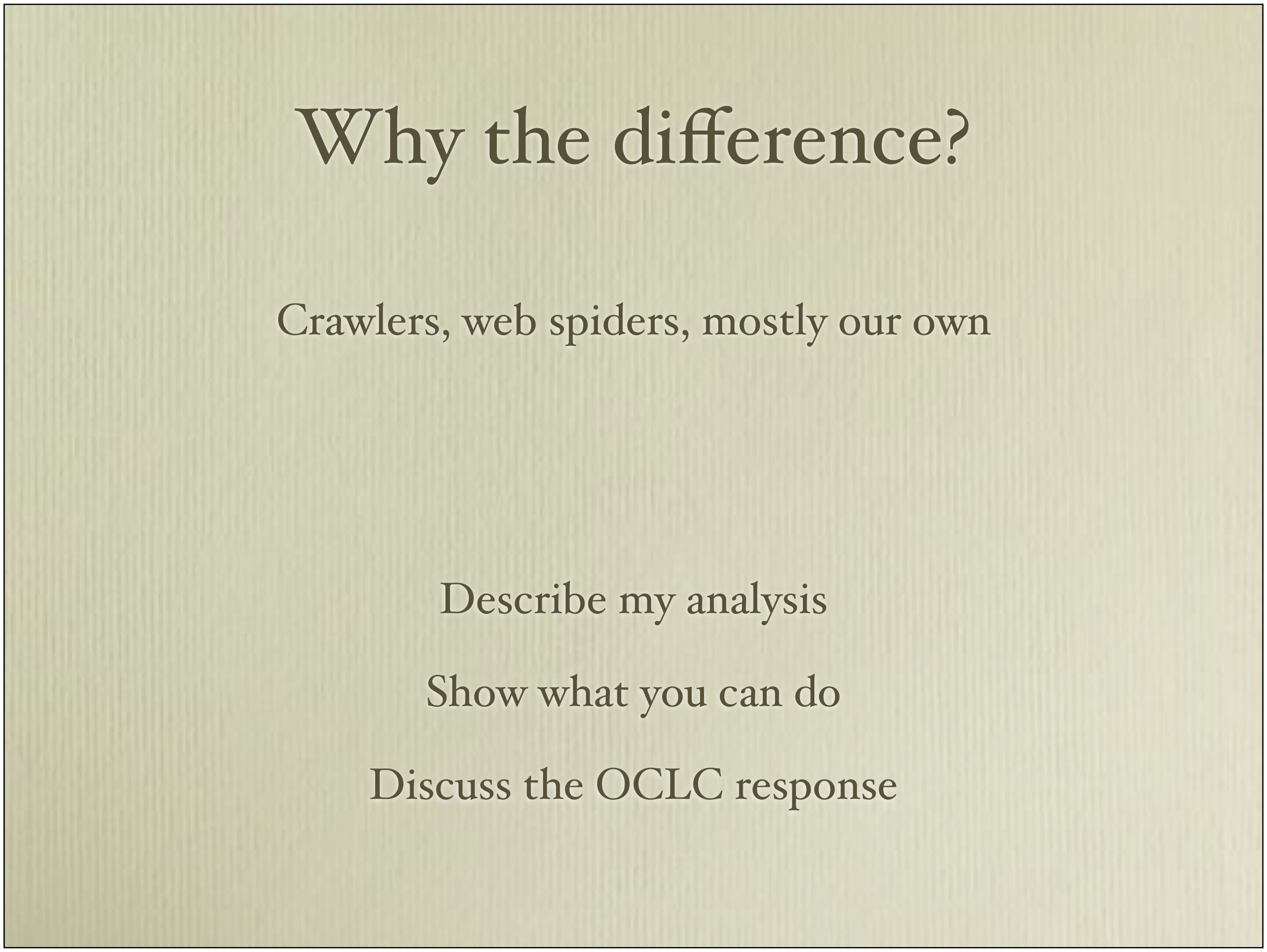
Sorry for the quality of this chart, but it shows the number of views for a collection in Flickr. We duplicated one of our Minnesota Reflections collections in the Flickr Commons. I expected Flickr to show much higher usage than CONTENTdm.



It seemed very suspicious to me that our CONTENTdm stats showed more hits than our Flickr Commons stats. In fact, I didn't believe this could be the case, I began to look for an explanation.



What I found was that CONTENTdm overcounts by quite a large margin. Once the overcount was removed, it was clear that Flickr had actually brought us many many more users than our local CONTENTdm collection.



Why the difference?

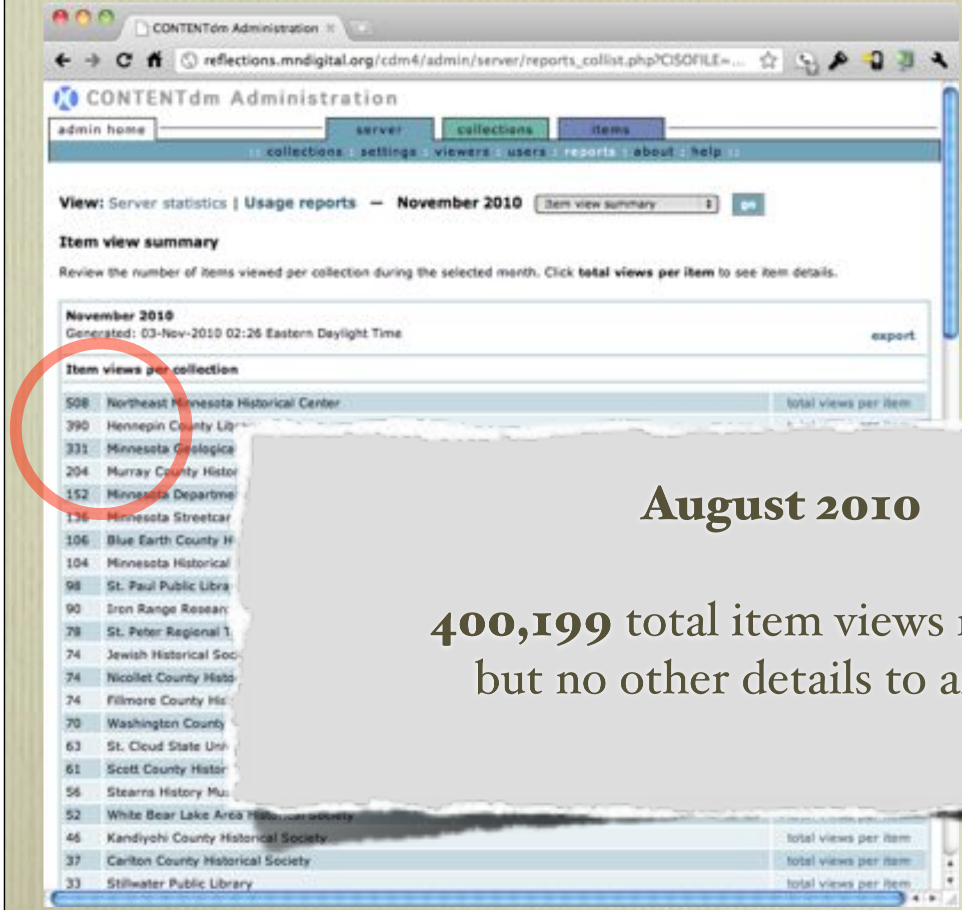
Crawlers, web spiders, mostly our own

Describe my analysis

Show what you can do

Discuss the OCLC response

The rest of this presentation explains what I found and how OCLC responded to the problem.


CONTENTdm Administration
reflections.mndigital.org/cdm4/admin/server/reports_collist.php?CISOFILE=...
CONTENTdm Administration
admin home
server
collections
items
:: collections : settings : viewers : users : reports : about : help ::
Item view summary
go

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

export

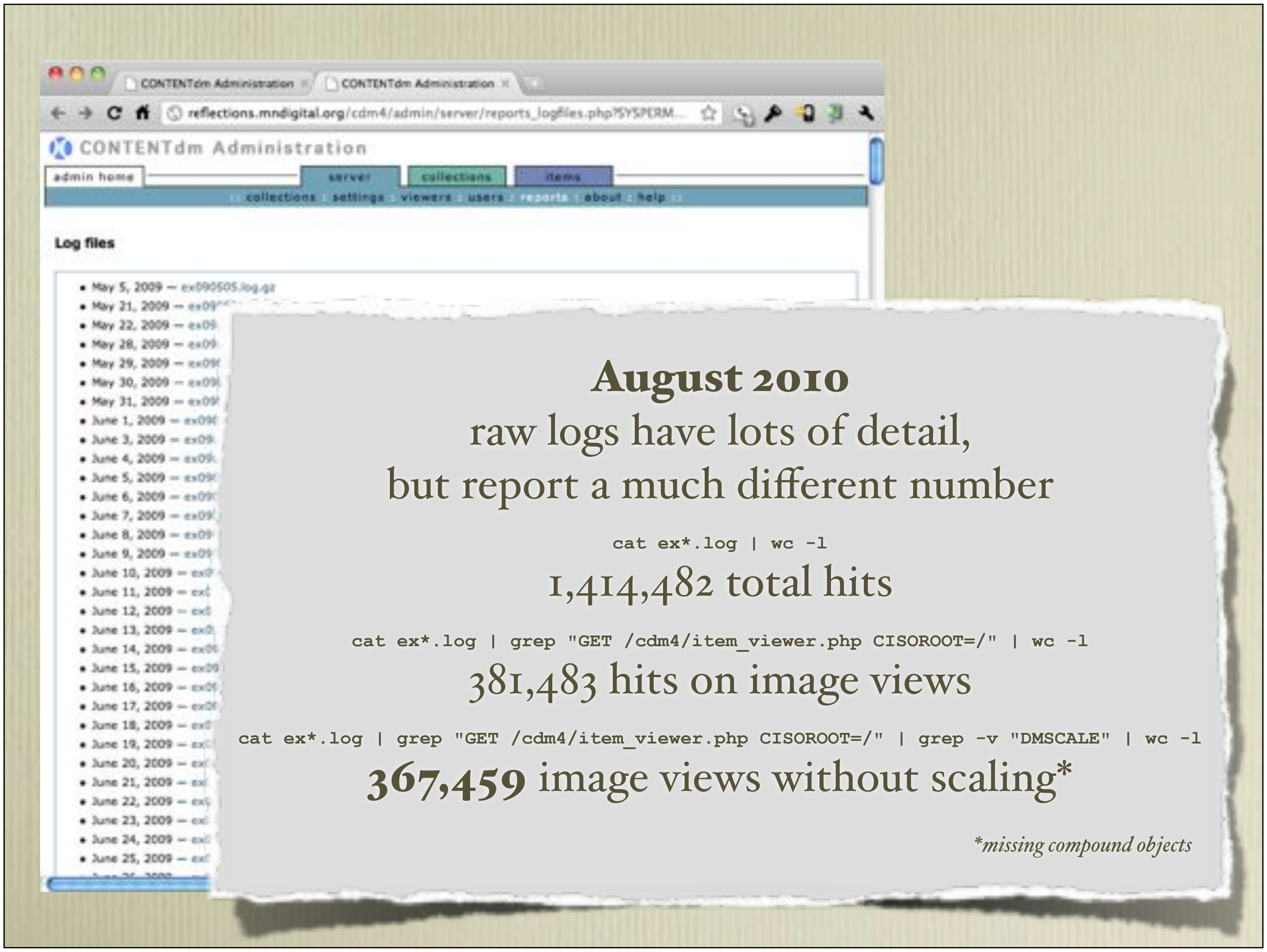
Item views per collection

508	Northeast Minnesota Historical Center	total views per item
390	Hennepin County Libr	
331	Minnesota Geologica	
204	Murray County Histor	
152	Minnesota Departme	
136	Minnesota Streetcar	
106	Blue Earth County H	
104	Minnesota Historical	
98	St. Paul Public Libra	
90	Iron Range Research	
78	St. Peter Regional T	
74	Jewish Historical Soc	
74	Nicollet County Histo	
74	Fillmore County His	
70	Washington County	
63	St. Cloud State Uni	
61	Scott County Histor	
56	Stearns History Mu	
52	White Bear Lake Area Historical Society	
46	Kandiyohi County Historical Society	total views per item
37	Carlton County Historical Society	total views per item
33	Stillwater Public Library	total views per item

August 2010

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

Using the CONTENTdm Usage Reports (the export to spreadsheet function) I found that CONTENTdm thought we'd had 400,199 item views in August 2010.


CONTENTdm Administration
CONTENTdm Administration
reflections.mndigital.org/cdm4/admin/server/reports_logfiles.php?SYSPERM...
CONTENTdm Administration
admin home
server
collections
items
:: collections : settings : viewers : users : reports : about : help ::

Log files

- May 5, 2009 — ex090505.log.gz
- May 21, 2009 — ex09
- May 22, 2009 — ex09
- May 28, 2009 — ex09
- May 29, 2009 — ex09
- May 30, 2009 — ex09
- May 31, 2009 — ex09
- June 1, 2009 — ex09
- June 3, 2009 — ex09
- June 4, 2009 — ex09
- June 5, 2009 — ex09
- June 6, 2009 — ex09
- June 7, 2009 — ex09
- June 8, 2009 — ex09
- June 9, 2009 — ex09
- June 10, 2009 — ex0
- June 11, 2009 — ex
- June 12, 2009 — ex
- June 13, 2009 — ex0
- June 14, 2009 — ex0
- June 15, 2009 — ex09
- June 16, 2009 — ex0
- June 17, 2009 — ex0
- June 18, 2009 — ex0
- June 19, 2009 — ex
- June 20, 2009 — ex
- June 21, 2009 — ex
- June 22, 2009 — ex
- June 23, 2009 — ex
- June 24, 2009 — ex
- June 25, 2009 — ex

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
```

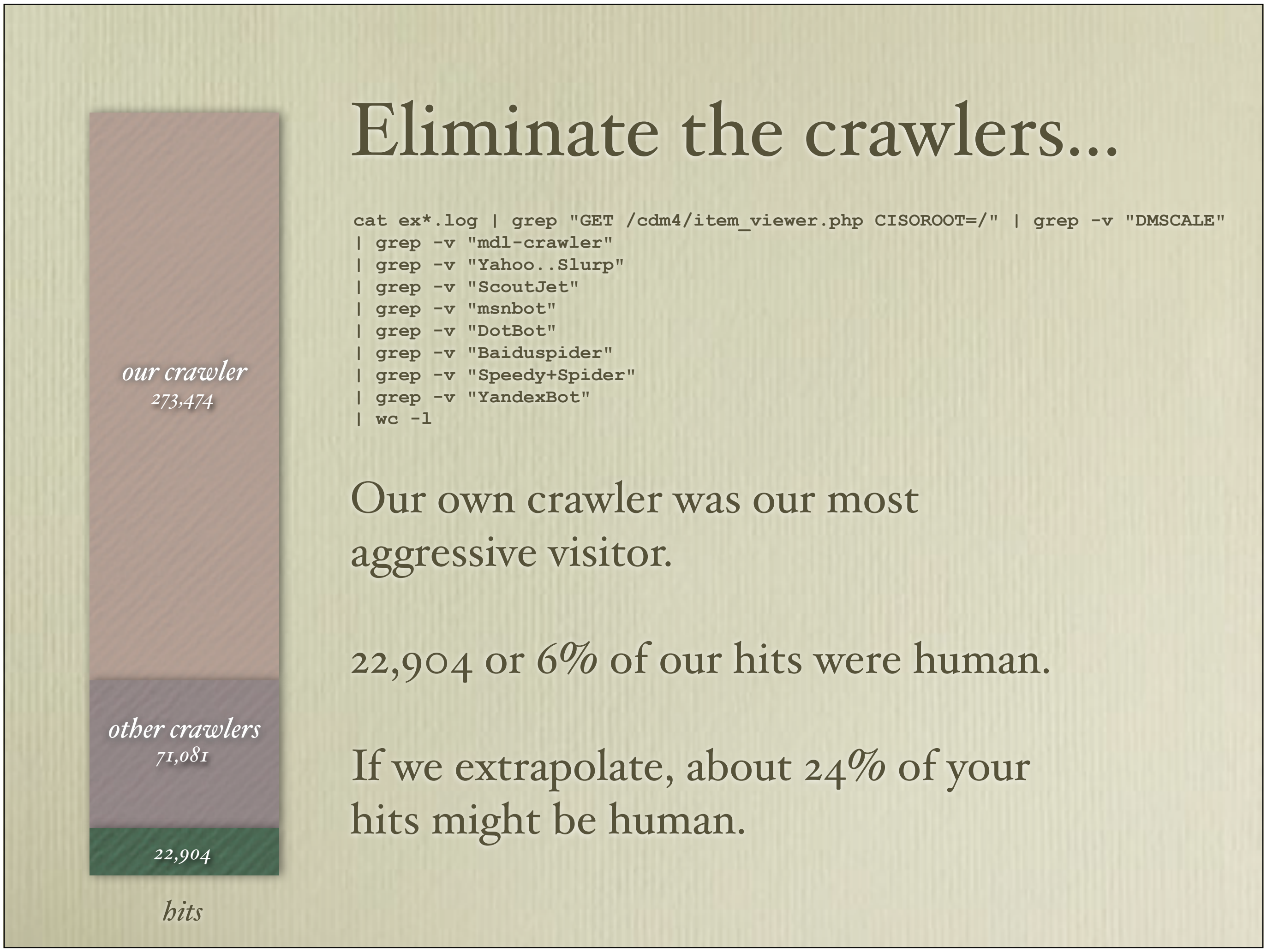
381,483 hits on image views

```
cat ex*.log | grep "GET /cdm4/item_viewer.php CISOROOT=/" | grep -v "DMSCALE" | wc -l
```

367,459 image views without scaling*

**missing compound objects*

I then retrieved the raw log files for each day in August from the CONTENTdm administration module. Using some Unix command line tools on my Mac, I analyzed these raw files to try to figure out what CONTENTdm might be calling an “item view.” As you can see, I got close to the Usage Report figure, showing 367,459 “hits” without counting compound objects. I think this is probably a similar “filter” to what the Usage Reports use, but I have not been able to confirm this with OCLC.



Eliminate the crawlers...

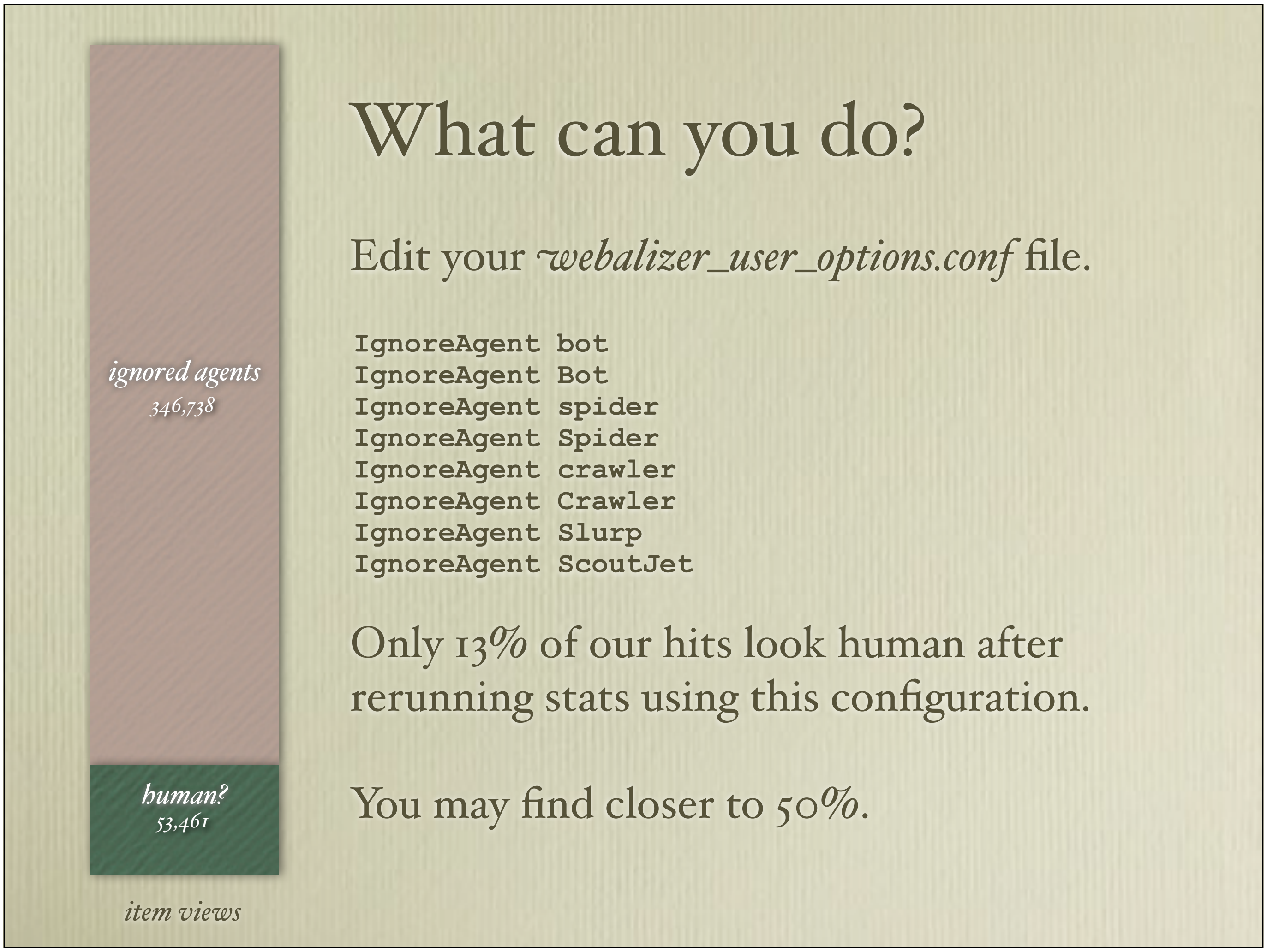
```
cat ex*.log | grep "GET /cdm4/item_viewer.php CISOROOT=/" | grep -v "DMSCALE"
| grep -v "mdl-crawler"
| grep -v "Yahoo..Slurp"
| grep -v "ScoutJet"
| grep -v "msnbot"
| grep -v "DotBot"
| grep -v "Baiduspider"
| grep -v "Speedy+Spider"
| grep -v "YandexBot"
| wc -l
```

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.

Taking that 367,459 as my starting point for August 2010, I then applied a further filter using more command line tools. For one thing, it appears that the Usage Reports were counting hits from a crawler that we use to regularly index our own system. That accounted for a huge number of hits. But even after removing our own crawler from the figure, it appeared CONTENTdm was counting a whole host of other crawlers, inflating the Usage Reports by about 4 times.


ignored agents
346,738
human?
53,461
item views

What can you do?

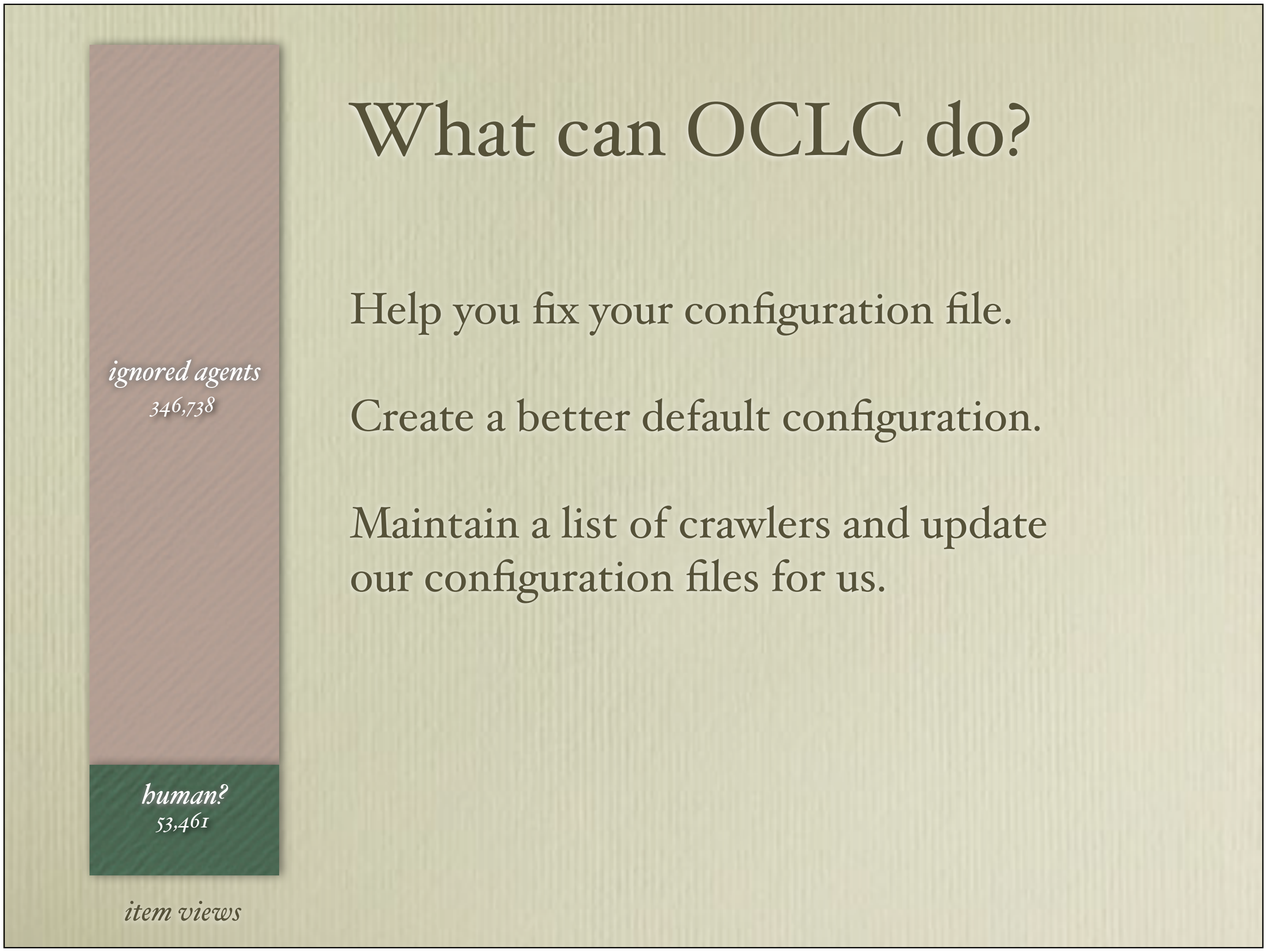
Edit your *webalizer_user_options.conf* file.

```
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%.

I contacted OCLC support about this huge overcount and learned that they could edit a file called “webalizer_user_options.conf” that was not normally accessible to me as a customer. Their configuration file only tried to ignore “agents” (web clients) with the text “bot” in their name. It turns out there were a ton of crawlers hitting our server with names other than “bot” and since the configuration file is case sensitive, OCLC was not even filtering out “Bot”. OCLC added a whole set of “IgnoreAgent” directives to the configuration file for us. Your reduction may not be as dramatic, remember we had our own crawler hitting our system an awful lot. Still, I estimate that taking this step with OCLC support will probably reduce your Usage Report stats by about 50%.



ignored agents
346,738
human?
53,461
item views

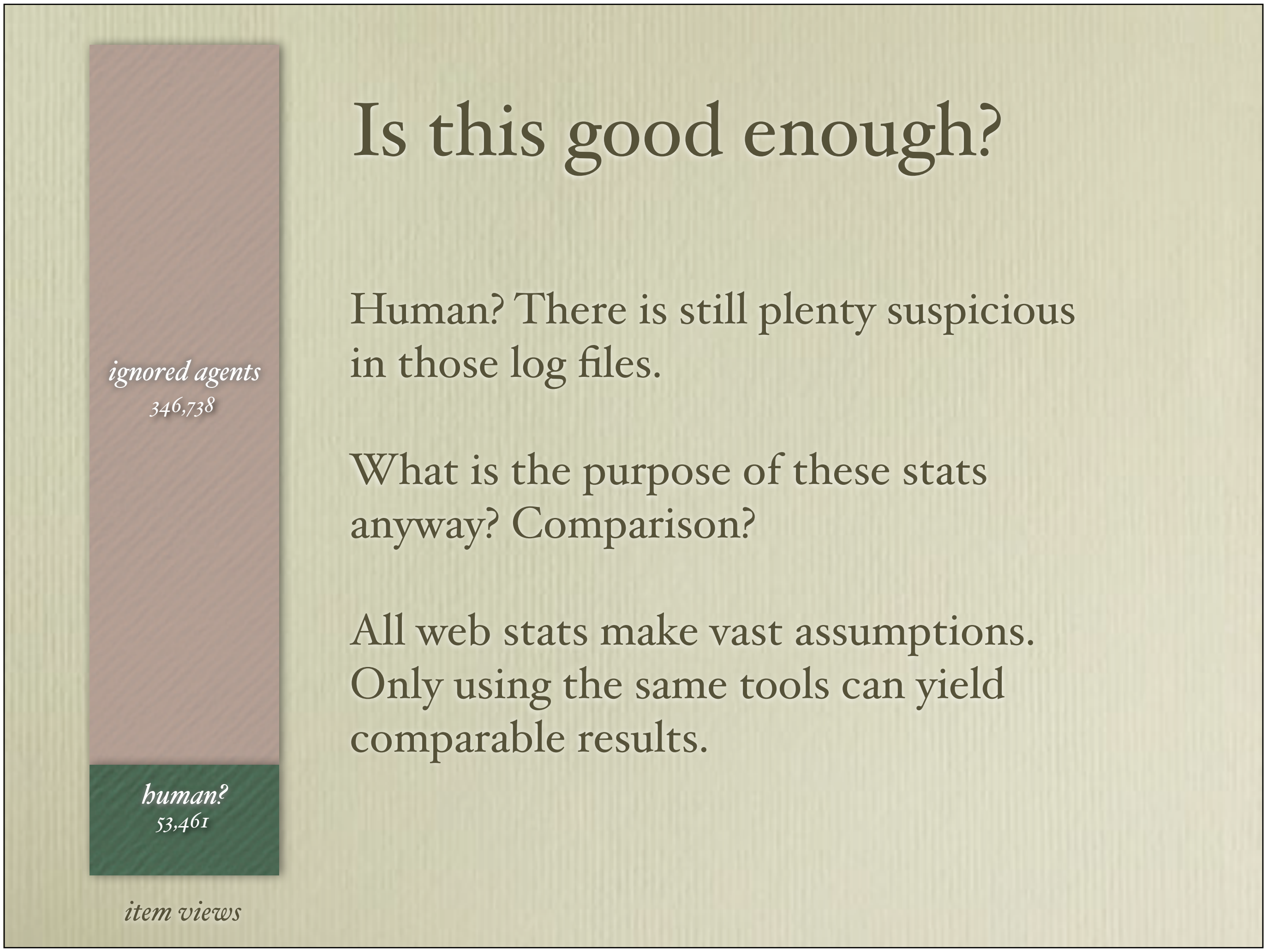
What can OCLC do?

Help you fix your configuration file.

Create a better default configuration.

Maintain a list of crawlers and update our configuration files for us.

Of course, we don’t all want to keep track of the web crawlers hunting the internet. It would be nice if OCLC added as a service for hosted sites, at least, an automatic maintenance of the “webalizer_user_options.conf” file using what it knows of crawlers on the net. OCLC runs a lot of services, I’m sure someone there is already maintaining this kind of list to assure reasonable stats for other parts of OCLC’s good works. I hope the CONTENTdm team can leverage that work for its customers.



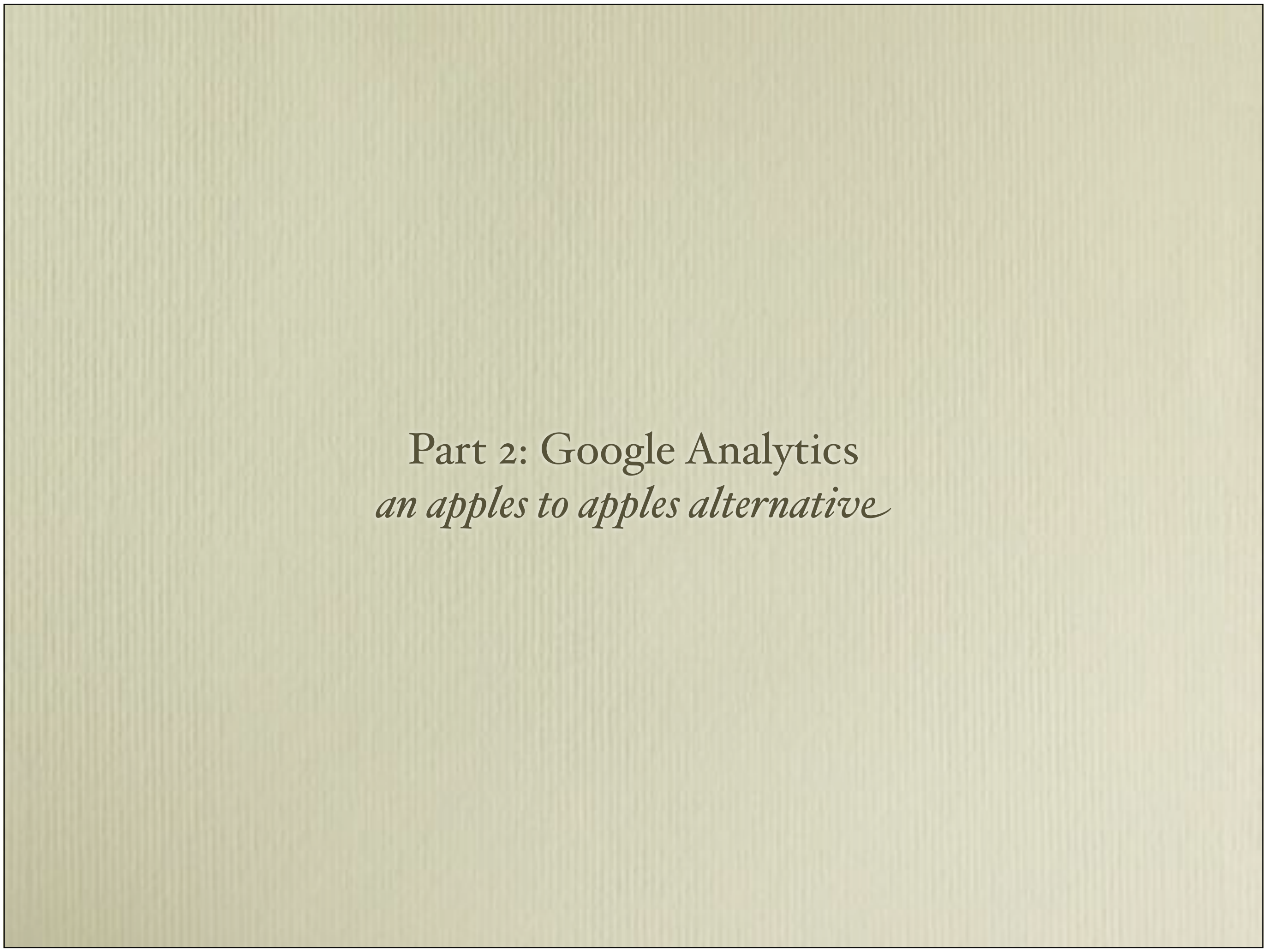
Is this good enough?

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.

Still, this is only a partial solution. In fact, there are many other oddities in those raw logs once you start digging in to them. I only did the simplest thing, screening out honest web crawlers that I could identify. There appear to be many other systems that hit our CONTENTdm server much more rapidly and regularly than any human could, but without identifying themselves as crawlers. I could develop heuristics to screen those out as well, but then I’d just be reducing our stats further, and who does that help? The incentive for cleaning up CONTENTdm stats is limited. The only way to really get useful numbers would be to have some sort of apples-to-apples comparison with other sites.



Part 2: Google Analytics

an apples to apples alternative

That’s where part two of this presentation came in. The folks from the Macalester College libraries showed us what they have done to integrate Google Analytics with CONTENTdm. Google Analytics has its own quirks, but at least it is widely used and available to anyone, it can provide an apples-to-apples comparison.

Again, you can reach me at efc@clst.org with questions, or check http://eric.clst.org/MDL/CONTENTdmStats for any updates.