Polluted Leftovers: Repository Metrics from the Perspective of a Most Downloaded Item

Jon Wheeler jwheel01@unm.edu | Kenning Arlitsch kenning.arlitsch@montana.edu
Patrick OBrien, Montana State University | Jeff Mixter, OCLC Research | Leila Sterman, Montana State University |

MONTANA STATE UNIVERSITY
Mountains & Minds

OCLC Research

Association of Research Libraries

UNM College of University Libraries & Learning Sciences

Measuring Up grant: About

- IMLS-funded grant, 2014-2017
  - "Measuring Up: Assessing Accuracy of Reported Use and Impact of Digital Repositories"
  - http://scholarworks.montana.edu/xmlui/handle/1/8924
  - Partners: MSU, UNM, OCLC Research, ARL
- Main driver
  - Improving accuracy and consistency of reporting
    - Research demonstrates under and over-counting of IR use
    - Solution = RAMP (Repository Analytics & Metrics Portal)

Overview

- Problems of establishing consistent, reliable metrics of Institutional Repository (IR) usage
- Current research into systemic over-counting and under-counting
- Mapping IR log data to Google Analytics and Search Console data and characterizing the differences
  - Assumptions underlying analytics services can rule out legitimate user interactions

Problems of IR Reporting

- Over- and Undercounting
- Variety of analytics services and methods
  - Log files
  - Software services (page tagging)
- General concerns
  - Bots
  - Variety of configurations == maintenance and continuity questions for a UL, consistency and reliability across IR ecosystem more broadly

DSpace Configuration Options

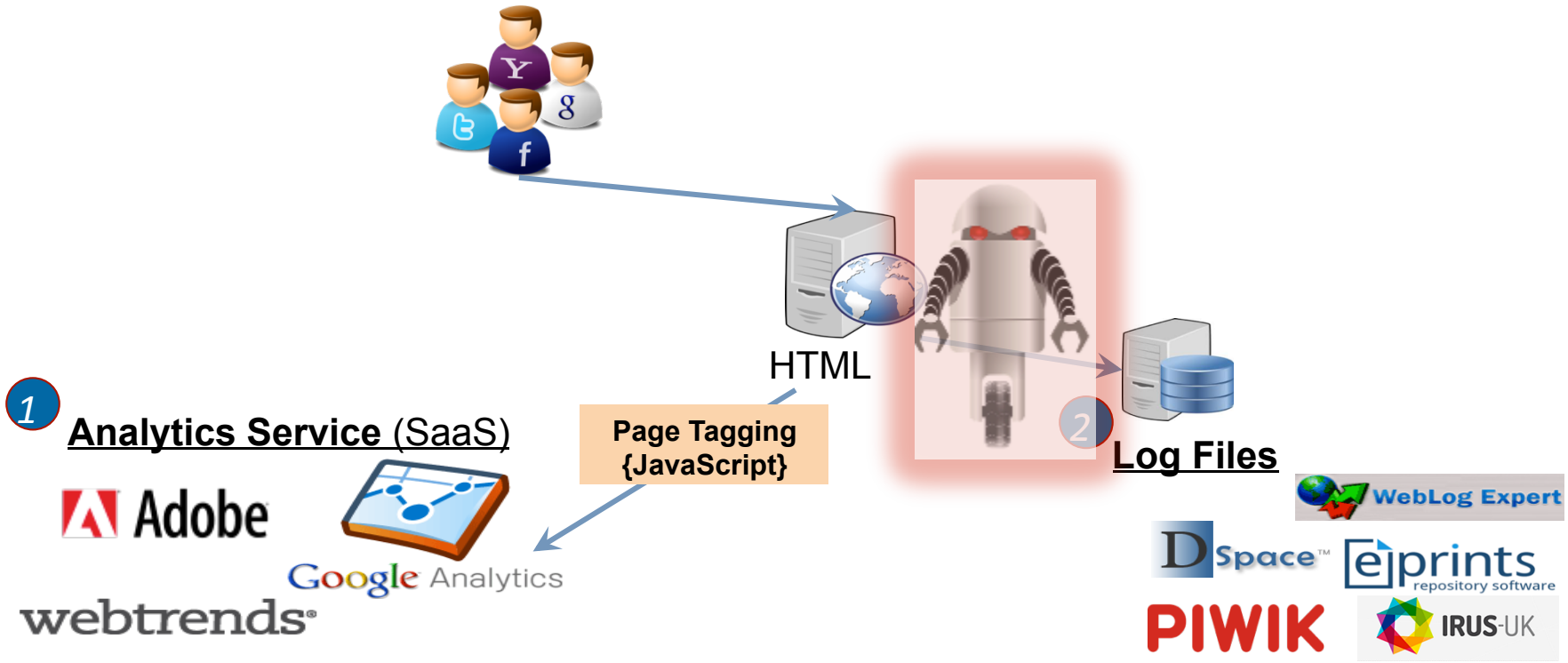
Option	DSpace Version	DSpace UI Compatibility	Integration with Earlier Stats	Other Notes
Google Analytics (GA)	GA UI since pre v5, DSpace plugin since v5. "Events" since v5.	XML UI, Mirage 2 theme only.	NA	DS plugin must be enabled
Elasticsearch (ES)	Since v3, deprecated in v6.	XML UI only.	Legacy Solr data can be converted.	Must be enabled. Default set of fields is not configurable.
Solr	Since v1.6.	XML and JSP UI, some difference in how facet events are logged.	Legacy system stats can be converted.	Same as ES, plus downloads, workflow, and events.

https://wiki.duraspace.org/display/DSDOC5x/Statistics+and+Metrics

A New Reporting Model

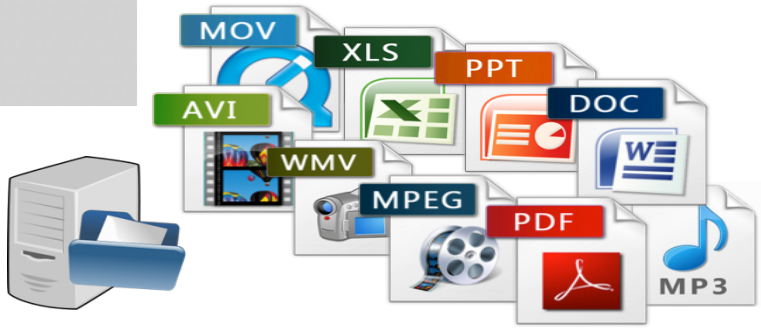
Page Type	Definition	Examples
Citable Content Downloads	Non-HTML scholarly content that may be formally cited in the research process	• Publication (.pdf) • Presentation (.ppt) • Data Sets (.csv)
Item Summary	HTML pages to help user decide to download the full publication	• Title & Abstract • Item Metadata
Ancillary	HTML pages that provide general information or navigation	• Search Results • Browse by Author • Statistics

Two Classes of Web Analytics



Page tagging methods do not tack non-HTML Citable Content Downloads

Does!



Non-HTML

GA Ancillary PV and Item Summary PV vs CCD

IR	Item Sumary PV	Ancillary PV	Total Google Analytics HTML PV	Download Events	Citable Content Downloads
scholarworks.montana.edu	26,735	23,350	50,085	7,129	77,380
macsphere.mcmaster.ca	51,150	71,585	122,735	n/a	133,342
repository.unm.edu	83,491	59,289	142,780	n/a	166,320
content.lib.utah.edu	122,927	47,569	170,496	19,226	159,536

134-day period in Spring 2016

GA HTML versus GSC CCD tracking

CCD Tracking Improvement

Page Type	GA		Search Console
	Pages	Events	Search Analytics
Citable Content Downloads	-	26,355	562,933
Item Summary	284,[illegible]03	[illegible]	-
Ancillary	201,793	-	-

+2,000%

Montana Method Challenges

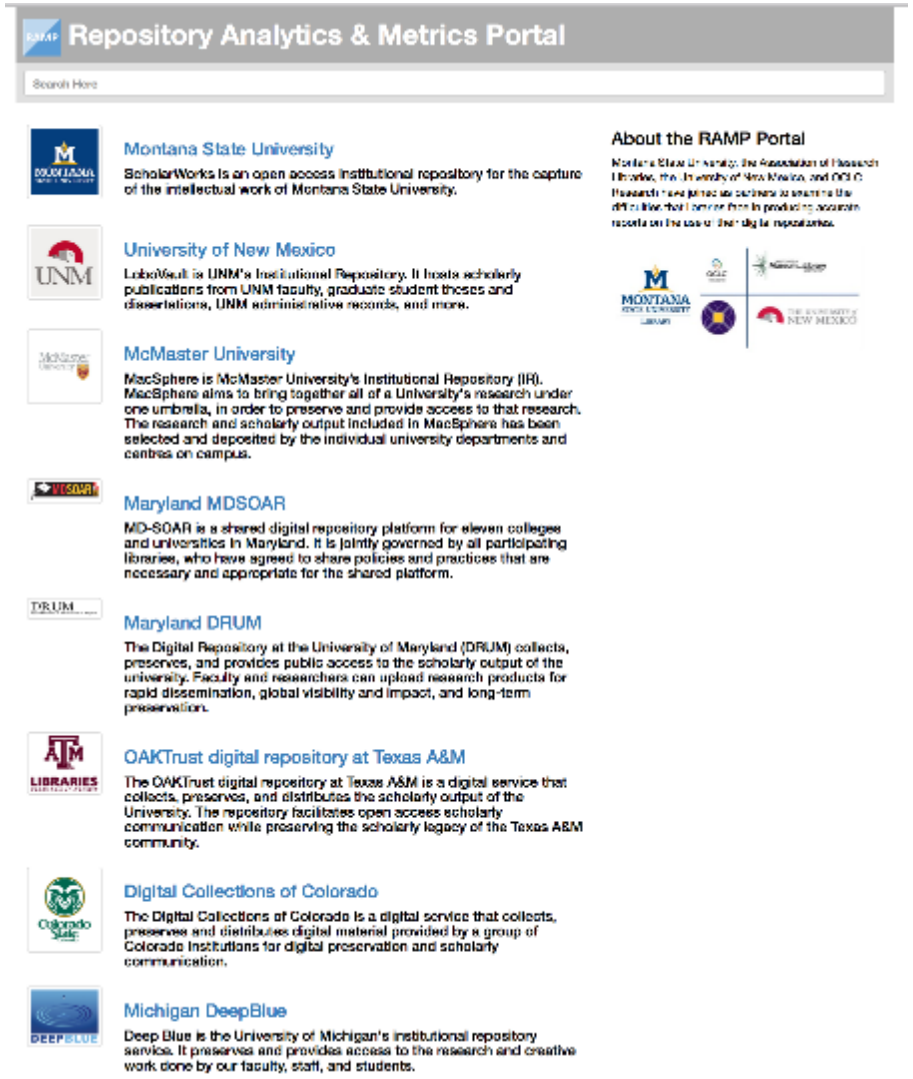
- Missing non-Google direct link CCD events
  - Yahoo
  - Bing
  - Email
  - Facebook & social media
- GSC limits time and access
  - Moving 90-day window
  - Granular data = programming skills to access API

RAMP: Repository Analytics & Metrics Portal

- ramp.montana.edu
- A benchmarking tool
- Prototype application
  - No local installation or configuration required
  - Integrated reporting from Google APIs

RAMP IR
as of March 27, 2017

- 8 IR registered
- Tracking over 250,000 digital items
- Capturing over 20,000 CCD per day that were previously invisible through GA.



RAMP

Repository Analytics & Metrics Portal

Montana State University

Access Data

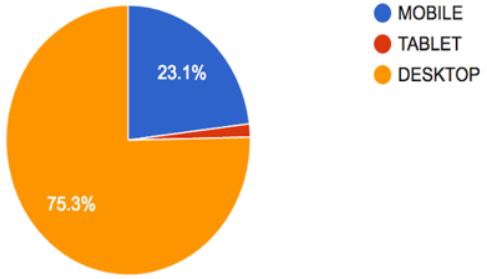
Selected Date: Wednesday, January 11, 2017

January 2017

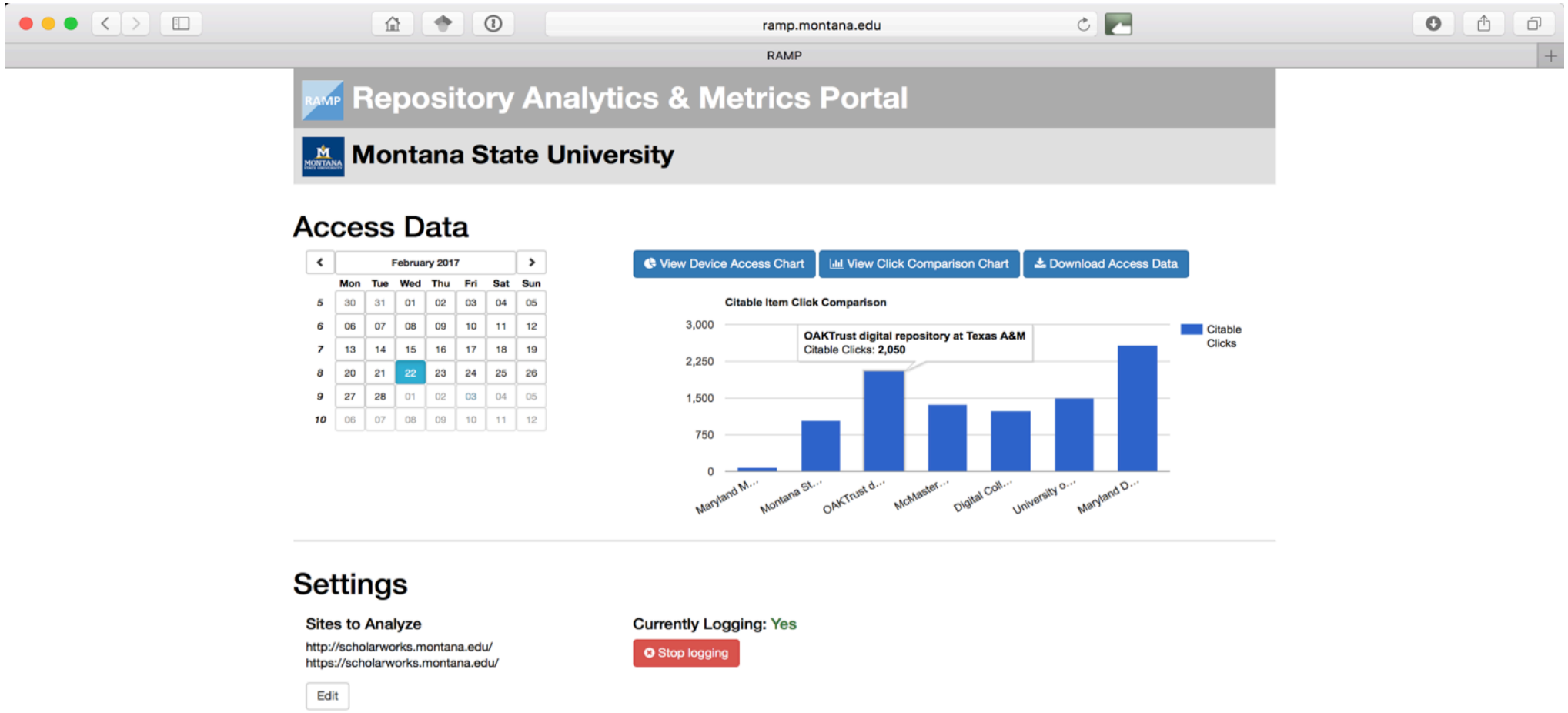
	Mon	Tue	Wed	Thu	Fri	Sat	Sun
52	26	27	28	29	30	31	01
1	02	03	04	05	06	07	08
2	09	10	11	12	13	14	15
3	16	17	18	19	20	21	22
4	23	24	25	26	27	28	29
5	30	31	01	02	03	04	05

View Device Access Chart | Download Access Data

Access by Device



RAMP


ramp.montana.edu
RAMP
Repository Analytics & Metrics Portal
Montana State University
Access Data
February 2017
Mon Tue Wed Thu Fri Sat Sun
View Device Access Chart
View Click Comparison Chart
Download Access Data
Citable Item Click Comparison
OAKTrust digital repository at Texas A&M
Citable Clicks: 2,050
Citable Clicks
3,000
2,250
1,500
750
0
Maryland M...
Montana St...
OAKTrust d...
McMaster...
Digital Coll...
University o...
Maryland D...
Settings
Sites to Analyze
http://scholarworks.montana.edu/
https://scholarworks.montana.edu/
Edit
Currently Logging: Yes
Stop logging

Sample Rows from Data Set

Citable Content	Click Through	URL	Country	Device	Position	Date	Impressions	Clicks
No	0	http://scholarworks.montana.edu/xmlui/handle/1/9348	hrv	DESKTOP	31	3/8/17	1	0
Yes	0	http://scholarworks.montana.edu/xmlui/bitstream/handle/1/8705/WhitenS0814.pdf;sequence=1	pan	MOBILE	6	3/8/17	1	0
Yes	0	http://scholarworks.montana.edu/xmlui/bitstream/handle/1/3670/31762001131281.pdf;sequence=1	fra	DESKTOP	24	3/8/17	1	0
Yes	0	http://scholarworks.montana.edu/xmlui/bitstream/handle/1/7215/31762101989810.pdf?sequence=1	chn	DESKTOP	13	3/8/17	2	0
Yes	0	http://scholarworks.montana.edu/xmlui/bitstream/handle/1/11518/15-002_Surface-attached_cells_biofilms_A1b.pdf?sequence=1	gbr	DESKTOP	10	3/8/17	1	0
Yes	0	http://scholarworks.montana.edu/xmlui/bitstream/1/1091/1/ColemanT1212.pdf	kwt	MOBILE	3	3/8/17	1	0
No	0	http://scholarworks.montana.edu/xmlui/handle/1/9049	gbr	DESKTOP	9	3/8/17	1	0
No	0	http://scholarworks.montana.edu/xmlui/handle/1/2567	egy	DESKTOP	44	3/8/17	1	0
Yes	0	http://scholarworks.montana.edu/xmlui/bitstream/handle/1/7546/31762102468723.pdf;sequence=1	twn	DESKTOP	14	3/8/17	1	0
No	0	http://scholarworks.montana.edu/xmlui/handle/1/1854	tur	DESKTOP	128	3/8/17	1	0

Polluted Leftovers

- How much DSpace Solr data is not also captured by Google Analytics (GA) or Google Search Console (GSC) data?
- How much of this activity is human and how much is bot? What kinds of human activity are not being captured by third party services?
- By focusing on GA/GSC data are we missing a significant percentage of citable content downloads? A fraction? A fraction of a fraction?

Looking for Stories in the Data

Processing Solr Logs

- Query stats core for citable content downloads (CCD)
  - BundleName = ORIGINAL
  - isBot = False
  - StatisticsType = view
  - Type = 0
- Query search core for metadata
  - search.resourcetype = 2 (bitstreams)
- Map CCD log events to metadata using metadata.search.resourceid = log.owningItem
- Consolidate logged events per Handle per day

Processing GA/GSC Data

- Harvest data dimensions via API
  - Event category
  - Page
  - Unique Events
- Filter for CCD
  - Page has “bitstream” in URL
  - Clicks > 0
- Add a column for Handles extracted from URLs
- Join to Solr data on combined key of Handle_date

Overview

Table	Solr CCD	Google CCD
Joined Solr statistics and search core data	854,781	
GA/GSC data with Handles		166,199
Solr events with corresponding GA/GSC events	356,557	166,199
Solr events without corresponding GA/GSC events	498,224	

A Closer Look at a Most Downloaded Item

	handle	hCount	owningItem	
	Filter	> 10000	Filter	Filter
1	1928/12178	79925	12200	Padre Sol\, Madre Luna :
2	1928/14546	17557	14733	New Mexico roots ltd : a c
3	1928/11748	17350	11788	Antropología indigenista
4	1928/11782	12825	11826	Investigación científica
5	1928/10563	11412	10430	La danza de los signos : n

Statistical Report for Padre Sol, Madre Luna : cuentos del desarrollo de base pluricultural

Report Generator

Used to generate reports with an arbitrary date range.

Generate Report

Showing Data (All Data Available)

Back to Summary Statistics for Padre Sol, Madre Luna : cuentos del desarrollo de base pluricultural

Print This Report
Download Data as .csv

Top Downloads (all time)

Title	Creator	Publisher	Date	Count
Padre Sol, Madre Luna : cuentos del desarrollo de base pluricultural		Abya-Yala	2000	260134

Top Downloads for February 2017

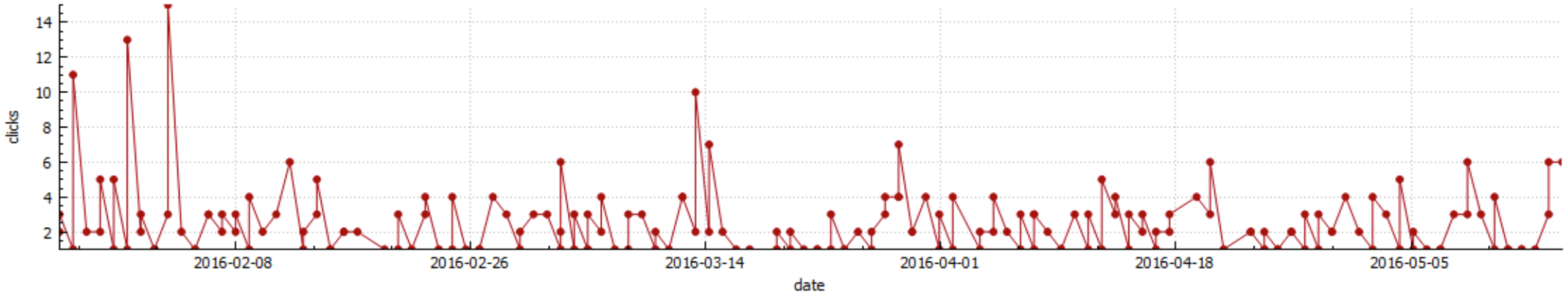
Title	Creator	Publisher	Date	Count
Padre Sol, Madre Luna : cuentos del desarrollo de base pluricultural		Abya-Yala	2000	195

Overview: 12178

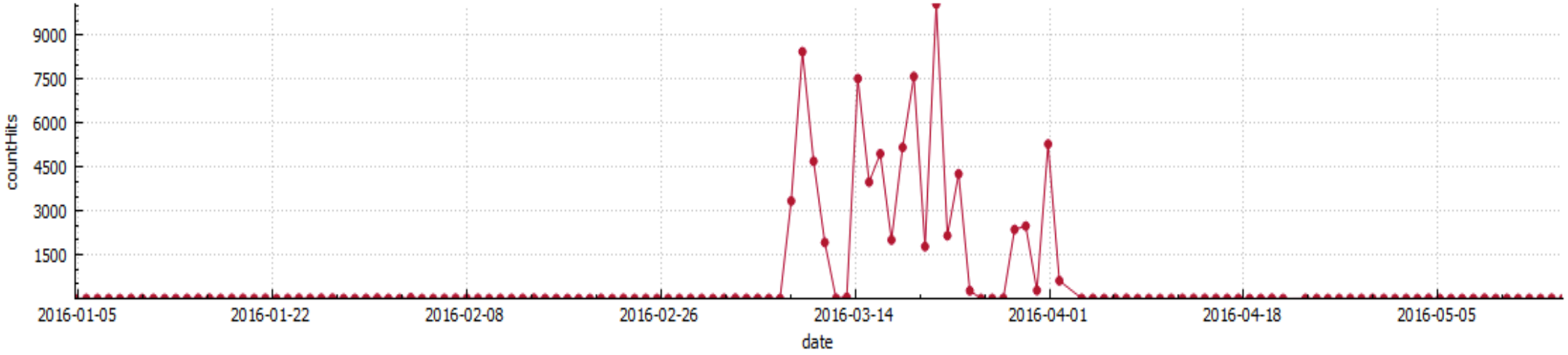
Table	Solr CCD	Google CCD
Joined Solr statistics and search core data	79,925	
GA/GSC with Handles		462
Solr events with corresponding GA/GSC events	74,612	462
Solr events without corresponding GA/GSC events	5313	

Comparison of Logged Events

GA/GSC

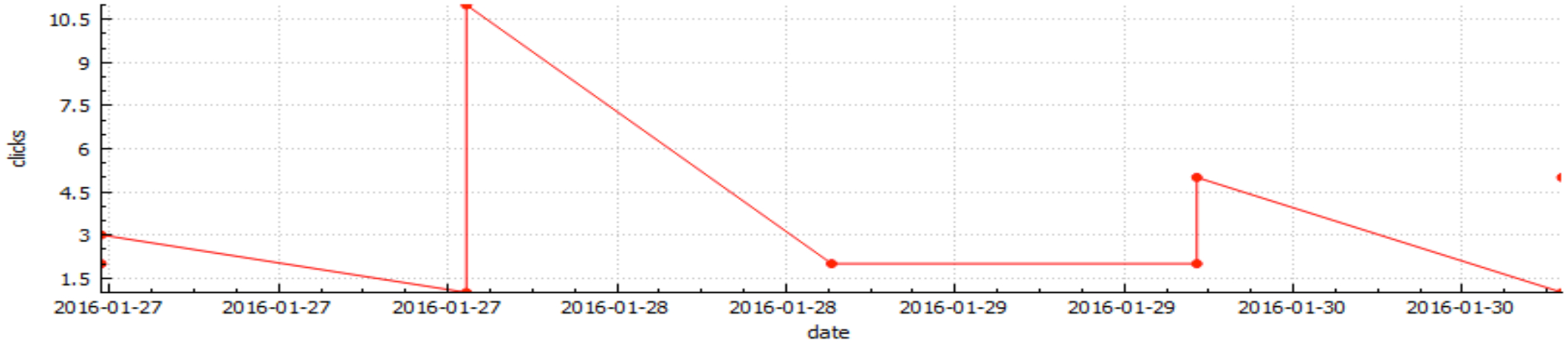


Solr

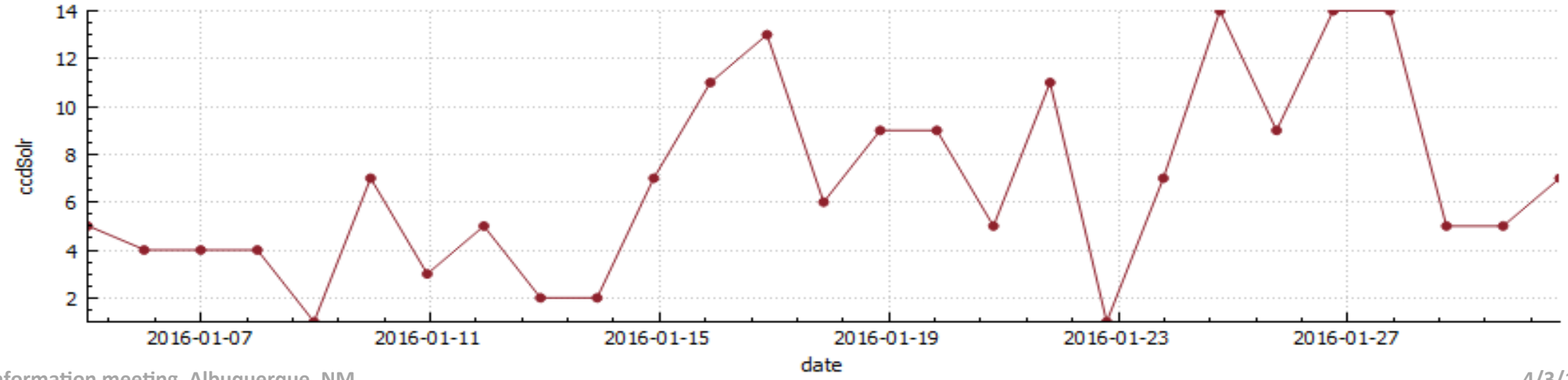


January Logged Events

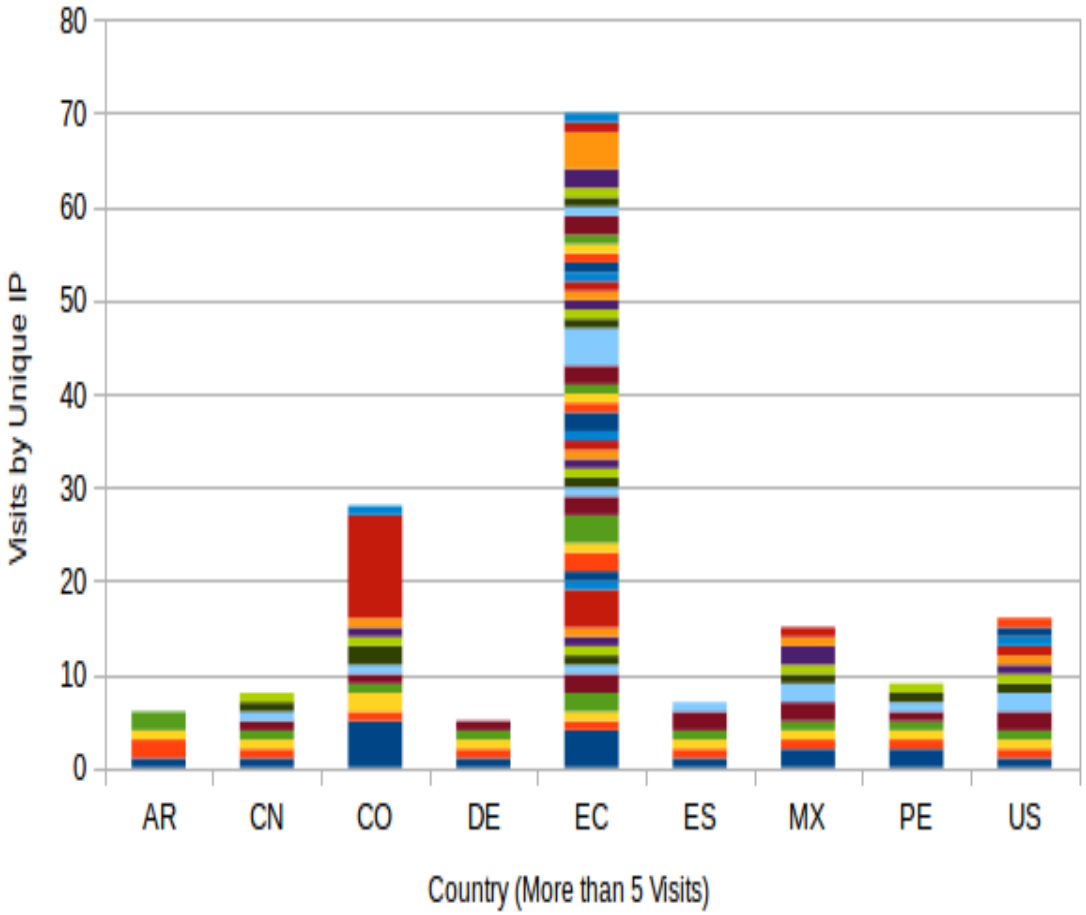
GA/GSC

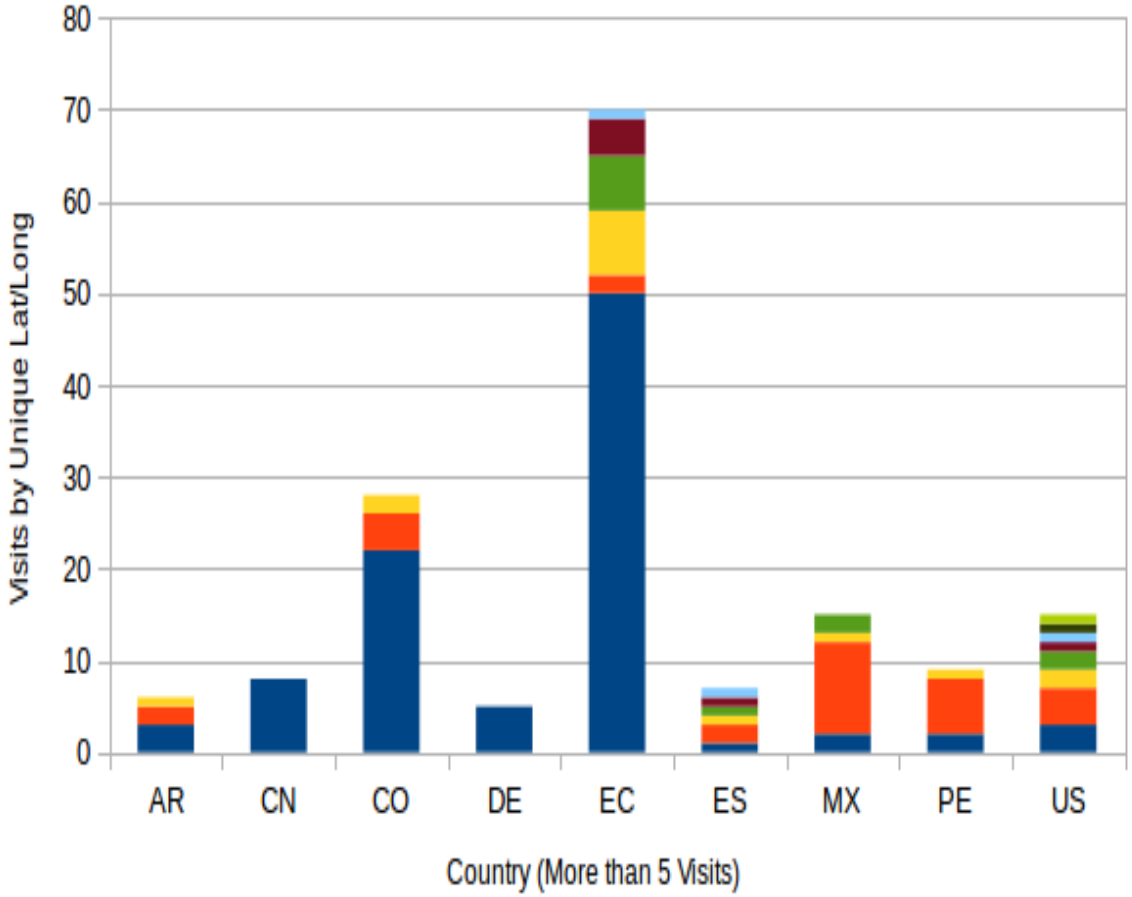


Solr



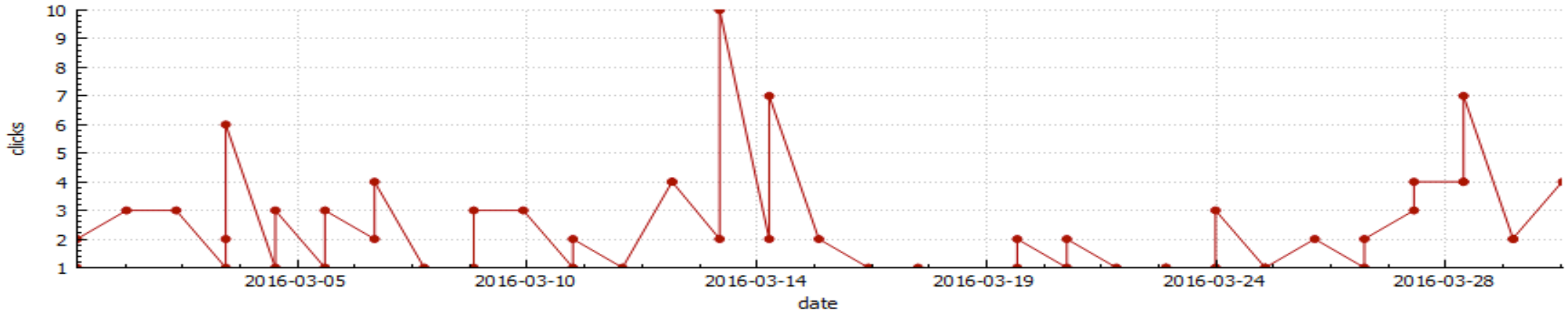
January Solr Unique IP and Unique Lat/Long



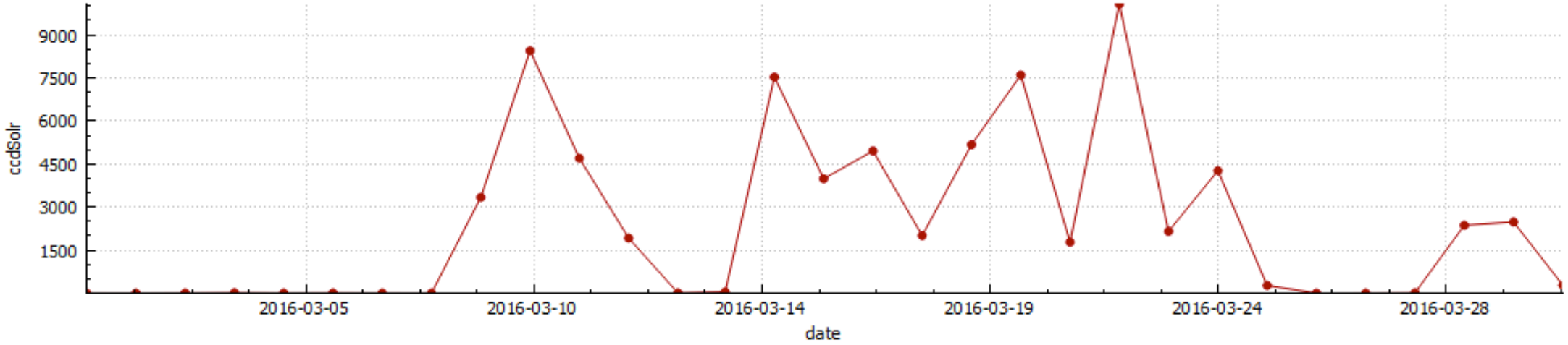


March Logged Events

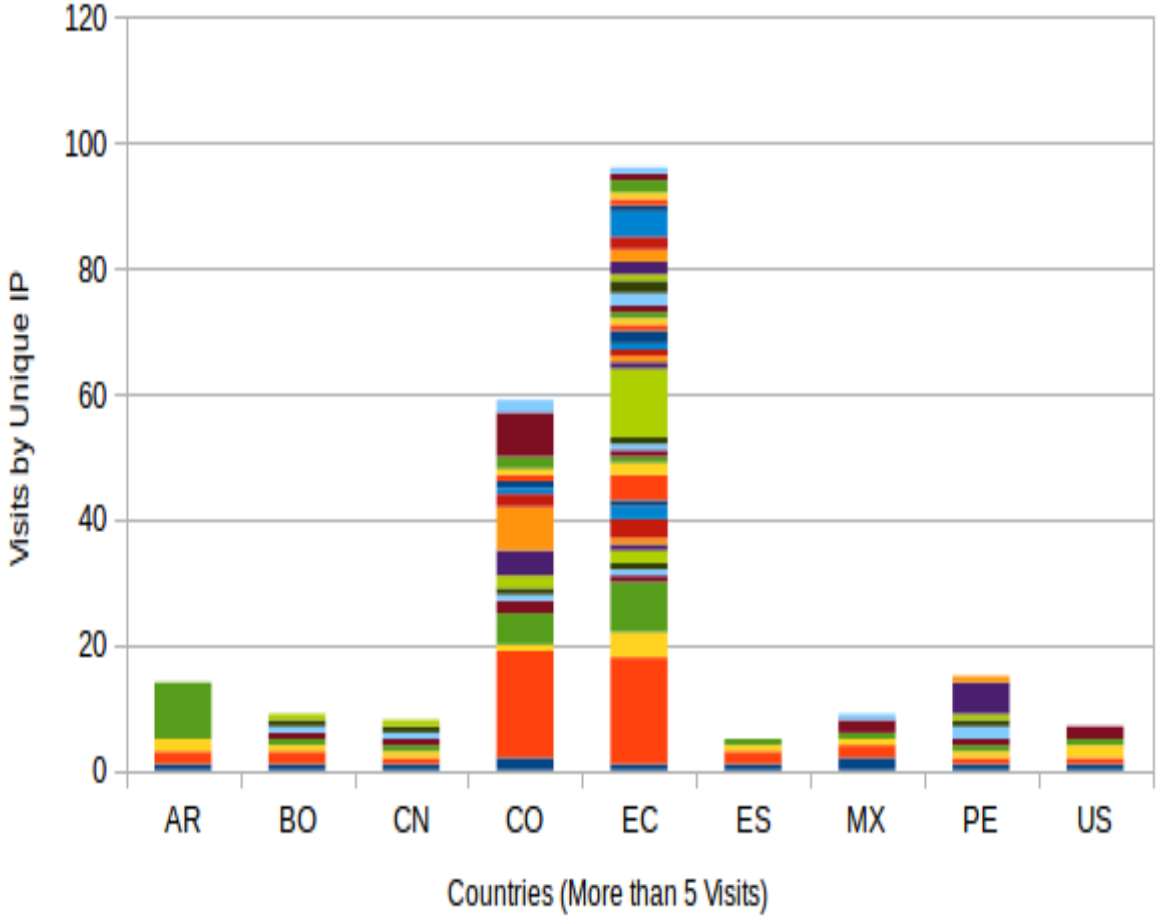
GA/GSC

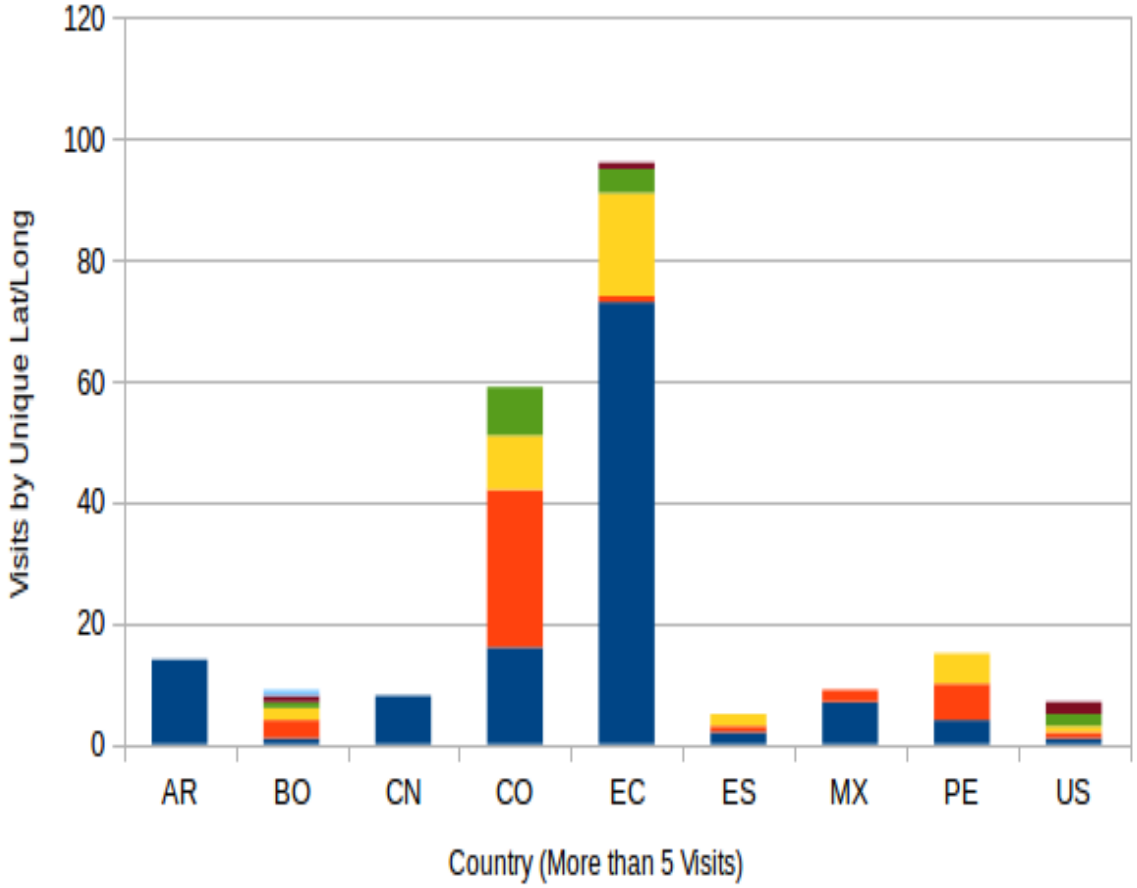


Solr



March Solr Unique IP & Unique Lat/Long





Considerations

- Human vs. Bot Behavior
  - What about organizations?
- Returning Users & Bookmarked Content
  - Visits and devices
  - Difficult to lump events by day because events may be separated by hours
- Pageviews vs. Downloads
- Access vs. Use
  - When is access not use?

Closing

- Variation among IR platforms and configurations complicates benchmarking internally and across institutions
- Log data over-count usage
  - Google does a much better job of filtering bots
- Page tagging services undercount usage
  - Item level analysis suggests that undercounting results from assumptions which rule out legitimate user interactions

Publications

Published:

Patrick OBrien, Kenning Arlitsch, Jeff Mixter, Jonathan Wheeler, Leila Sterman. “RAMP: Repository Analytics and Metrics Portal: A Prototype Web Service that Accurately Counts Item Downloads from Institutional Repositories,” *Library Hi Tech*, vol. 35, no. 1, March 2017

Patrick OBrien, Kenning Arlitsch, Leila Sterman, Jeff Mixter, Jonathan Wheeler, and Susan Borda. “Undercounting File Downloads from Institutional Repositories,” *Journal of Library Administration*, vol. 56, no. 7, 2016

Proposal funded by IMLS:

”Measuring Up: Assessing Accuracy of Reported Use and Impact of Digital Repositories” - scholarworks.montana.edu/xmlui/handle/1/8924

Jonathan Wheeler, Data Curation Librarian, University of New Mexico
jwheel01@unm.edu

Kenning Arlitsch, Dean of the Library, Montana State University
kenning.arlitsch@montana.edu
@kenning_msu

Thank you!