Reliable Facts from Unreliable Figures

COMPARING STATISTICAL PACKAGES IN DSPACE

BILL ANDERSON, SARA FUCHS, CHRIS HELMS
GEORGIA TECH LIBRARY
&
ANDY CARTER
UNIVERSITY OF GEORGIA

OPEN REPOSITORIES 2011
JUNE 11, 2011

Outline

- Why this project
- Georgia Tech's perspective
- UGA's perspective
- Problems with SMARTech Statistics
- Test plan
- Initial results
- Next steps

What do we want to learn?

1) What is the best way to capture statistics for a DSpace repository?

2) What statistics do we want to capture?

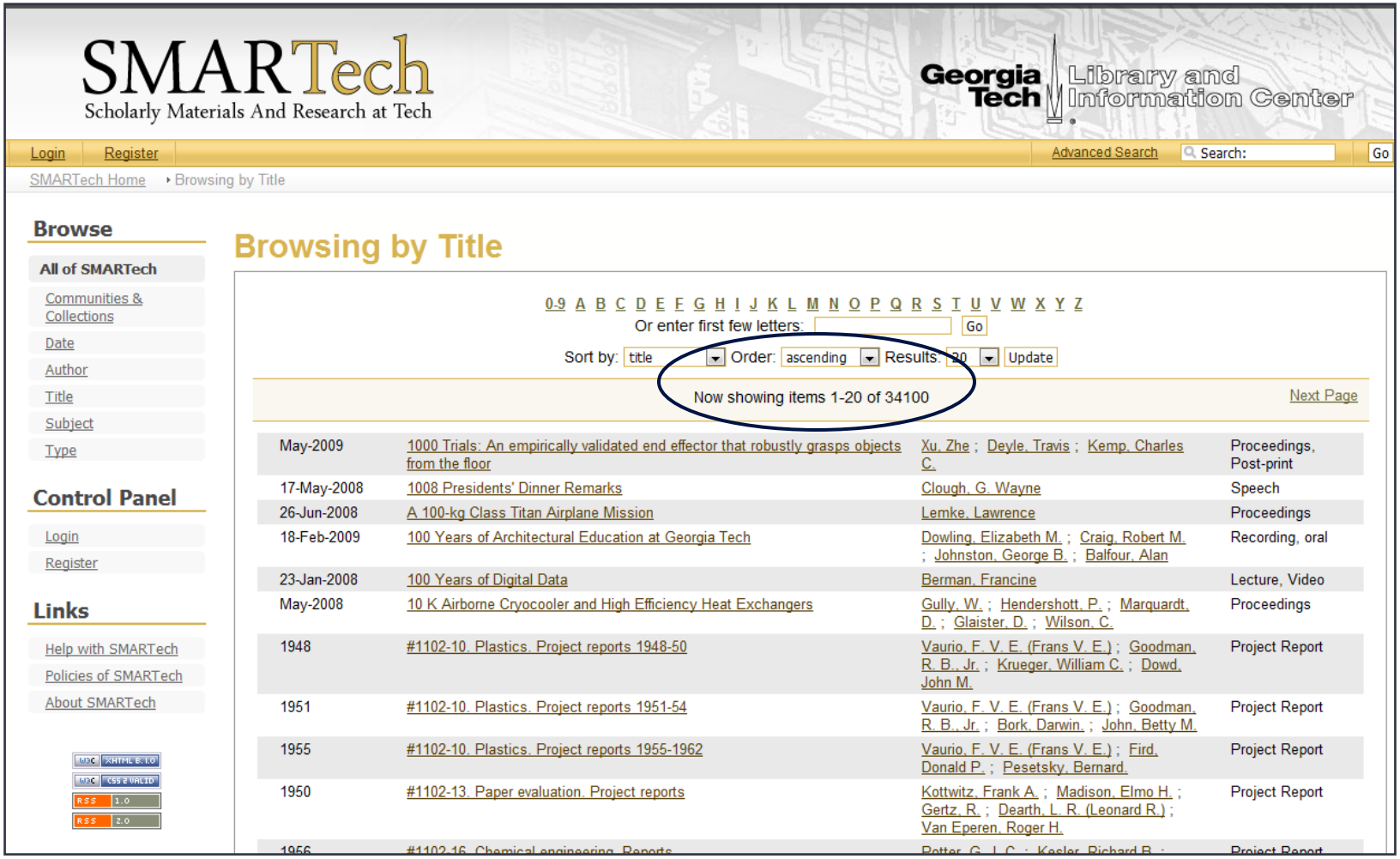
3) How do we best display these statistics to the end user?

Statistical Packages

We choose to focus on the following four:

- DSpace 1.7.1 with SOLR statistics
- DSpace statistics pre SOLR
- AWstats 7.0
- Google Analytics

SMARTech – Georgia Tech's Repository



SMARTech
Scholarly Materials And Research at Tech

Georgia Tech Library and Information Center

Login | Register | Advanced Search | Search: | Go

SMARTech Home › Browsing by Title

Browse

All of SMARTech
Communities & Collections
Date
Author
Title
Subject
Type

Control Panel

Login
Register

Links

Help with SMARTech
Policies of SMARTech
About SMARTech

Browsing by Title

0-9 A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Or enter first few letters: Go

Sort by: title Order: ascending Results: 20 Update

Now showing items 1-20 of 34100 Next Page

May-2009	1000 Trials: An empirically validated end effector that robustly grasps objects from the floor	Xu, Zhe ; Deyle, Travis ; Kemp, Charles C.	Proceedings, Post-print
17-May-2008	1008 Presidents' Dinner Remarks	Clough, G. Wayne	Speech
26-Jun-2008	A 100-kg Class Titan Airplane Mission	Lemke, Lawrence	Proceedings
18-Feb-2009	100 Years of Architectural Education at Georgia Tech	Dowling, Elizabeth M. ; Craig, Robert M. ; Johnston, George B. ; Balfour, Alan	Recording, oral
23-Jan-2008	100 Years of Digital Data	Berman, Francine	Lecture, Video
May-2008	10 K Airborne Cryocooler and High Efficiency Heat Exchangers	Gully, W. ; Hendershott, P. ; Marquardt, D. ; Glaister, D. ; Wilson, C.	Proceedings
1948	#1102-10. Plastics. Project reports 1948-50	Vaurio, F. V. E. (Frans V. E.) ; Goodman, R. B., Jr. ; Krueger, William C. ; Dowd, John M.	Project Report
1951	#1102-10. Plastics. Project reports 1951-54	Vaurio, F. V. E. (Frans V. E.) ; Goodman, R. B., Jr. ; Bork, Darwin. ; John, Betty M.	Project Report
1955	#1102-10. Plastics. Project reports 1955-1962	Vaurio, F. V. E. (Frans V. E.) ; Fird, Donald P. ; Pesetsky, Bernard.	Project Report
1950	#1102-13. Paper evaluation. Project reports	Kottwitz, Frank A. ; Madison, Elmo H. ; Gertz, R. ; Dearth, L. R. (Leonard R.) ; Van Eperen, Roger H.	Project Report
1956	#1102-16. Chemical engineering. Reports	Potter, G. L. C. ; Kesler, Richard B. ;	Project Report

Why did we initiate this project?

- Lack of trust in the numbers we were generating
- Create buy-in from submitters
- Popular content as basis of collection development decisions
- Rationale for existence of repository/future funding
- History of problems with DSpace statistics
- Solr problems meant we couldn't display stats to the author
- Lack of understanding of current numbers

Confessions of a Repository Manager

Statistics for Georgia Tech's

May 25, 2007 to Jun 2, 2011

General Overview

Items Archived	74
Bitstream Views	186,904
Item Views	213,701
Collection Views	13,295
Community Views	8,922
User Logins	116
Searches Performed	55,239
Licence Rejections	0
OAI Requests	4,630

Fiscal Year 2009-2010 Statistics	
Items viewed	2,693,150
Bitstreams viewed	4,046,314
Searches	789,327
OAI requests	42,799

Google Analytics

AWStats for May 2011	
Pages	**399,153**
Hits	**1,135,003**

Univ. of Georgia Knowledge Repository

- Launched in August of 2010
- Contains about 10,000 items

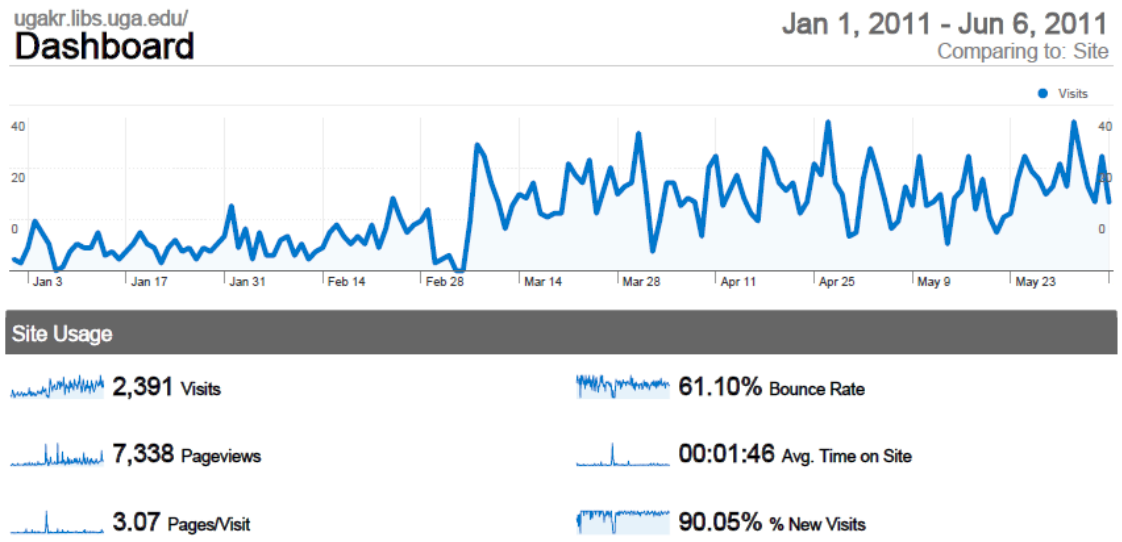
Statistics and the new repository

- Institutional context at Univ. of Georgia

- http://www.library.gatech.edu/gkr/

Stats and the new repository manager

- Do I know what I need to know? (Do I know what you need to know?)
- What do I know about what I do know?



Stats and the new repository manager

Items Viewed

(more than 20 times)

Item/Handle	Number of views
Cotton production and the boll weevil in Georgia: history, cost of control, and benefits of eradi...	94
Systematics of ladybird beetles (Ooleoptera: Coccinellidae) (Giorgi, Jose Adriano) (10724/11882)	87
http://ugakr.libs.uga.edu/handle/10724/10455	87
Quality control: a model program for the food industry (Hurst, William C. et al) (10724/12251)	85
Preventing food poisoning and food infection (Hurst, William C. et al) (10724/12254)	82
Background concentrations of trace elements in soils and rocks of the Georgia piedmont (Albright,...	73

All Actions Performed

Action	Number of times
browse	97,109
browse_by_item	75,460
Bitstream Views	63,491
OAI Requests	35,449
Item Views	25,846
browse_by_value	22,635

What Do Statistics Mean?

General Overview

Top

Items Archived	0
archive_item_excluded	22
Bitstream Views	171,758
view_bitstream_excluded	101,392
Item Views	118,448
view_item_excluded	124,682
Collection Views	11,201
view_collection_excluded	5,901
Community Views	9,253
view_community_excluded	3,158
User Logins	25
login_excluded	132
Searches Performed	44,422
Licence Rejections	0
OAI Requests	1,379

What's Wrong With This Picture?

Sep 1, 2008 to Sep 30, 2008

General Overview | Archive Information | Items Viewed | All Actions Performed | User Logins | Words Searched | Averaging Information | Log Level Information | Processing Information

General Overview

Top

Items Archived	0
Bitstream Views	0
Item Views	0
Collection Views	0
Community Views	0
User Logins	0
Searches Performed	0
Licence Rejections	0
OAI Requests	0

General Overview | Archive Information | Items Viewed | All Actions Performed | User Logins | Words Searched | Averaging Information | Log Level Information | Processing Information

The Hobgoblin of Little Minds

Item/Handle	Number of views
Belt line - Atlanta : design of infrastructure as a reflection of public policy (Gravel, Ryan Austin) (1853/7400)	616
An Evaluation of the low-income housing sector in Jamaica (Williams, Grace D.) (1853/13950)	464
Measurement of delignification diversity within kraft pulping (Boyer, Brian S.) (1853/5980)	273
List of Department/School Chairs and Bios (1853/12689)	160
DSpace 2.0 Demonstration (Bosman, Ben) (1853/28078)	155
Capitalization of Software Development Costs: A Survey of Accounting Practices in the Software Industry (Mulford, Charles W. et al) (1853/15598)	154
4th International Conference on Open Repositories Program (1853/28538)	139
Kinetics of $Ba(OH)_2$ Reaction with Na_2CO_3 and Na_2SO_4 & Particle Separation Characteristics from White Liquor (Quesada, Alexander L.) (1853/13165)	132
Alumni Association Trustee Reunion Welcome Speech (Clough, G. Wayne) (1853/12876)	128
Intuitive Revelations: The Ubiquitous Reference Model (Mathews, Brian S.) (1853/8446)	120
http://smartech.gatech.edu/handle/1853/19231	111
http://smartech.gatech.edu/handle/1853/24899	105
http://smartech.gatech.edu/handle/1853/28532	102
http://smartech.gatech.edu/handle/1853/28182	100
http://smartech.gatech.edu/handle/1853/28529	97

SMARTech
Scholarly Materials And Research at Tech

Georgia Tech Library and Information Center

Login | Register | Advanced Search | Search: | Go

SMARTech Home ▸ Community List

Browse

All of SMARTech

- Communities & Collections
- Date
- Author
- Title
- Subject
- Type

Control Panel

- Login
- Register

Statistics

- View Statistics

SMARTech Repository

If you would like to know more about SMARTech or would like to become an adopter, contact smartech@library.gatech.edu. SMARTech now contains 30,000 items, including over 15,000 Georgia Tech theses and dissertations!

Communities in SMARTech

Select a community to browse its collections.

- Center for Experimental Research in Computer Systems (CERCS)
- Center for Robotics and Intelligent Machines (RIM)
- Center for the Enhancement of Teaching and Learning (CETL)
- College of Architecture (CoA)
- College of Computing (CoC)
- College of Engineering (CoE)
- College of Liberal Arts - Ivan Allen College (IAC)
- College of Management (CoM)
- College of Sciences (CoS)


SMARTech
Scholarly Materials And Research at Tech
Georgia Tech Library and Information Center
Login
Register
Advanced Search
Search:
Go
SMARTech Home
Community List
Browse
All of SMARTech
Communities & Collections
Date
Author
Title
Subject
Type
Control Panel
Login
Register
Statistics
View Statistics
SMARTech Repository
If you would like to know more about SMARTech or would like to become an adopter, contact smartech@library.gatech.edu. SMARTech now contains 30,000 items, including over 15,000 Georgia Tech theses and dissertations.
SOLR
ATTACKS!
Communities in SMARTech
College of Architecture (CoA)
College of Computing (CoC)
College of Engineering (CoE)
College of Liberal Arts - Ivan Allen College (IAC)
College of Management (CoM)
College of Sciences (CoS)


ESPAÑA

Points to Consider

- Software can't fix wetware
- Where are visitors coming from? Are they really looking?
- Different packages count different things – changing software changes numbers
- Are we counting useful events? Are we counting them accurately?
- Spiders, harvesters, administrators, and other deadly enemies

Test Environment

- A virtual host running under ESX
- VM Setup
  - OS: Red Hat Enterprise Linux 6.0 (64-bit)
  - 2x Intel Xeon Core 2
  - 2048MB of memory
  - 30Gb of disk space
- DSpace 1.7.1, PostgreSQL 8.4.7, Java 1.6, Tomcat 6.0.32, Maven 2.2.1, Ant 1.8.2
- XMLUI with @MIRE Mirage theme
- 91 Items in archive

Configuration Notes

- Tomcat + mod_jk + Apache
- JAVA_OPTS for Tomcat

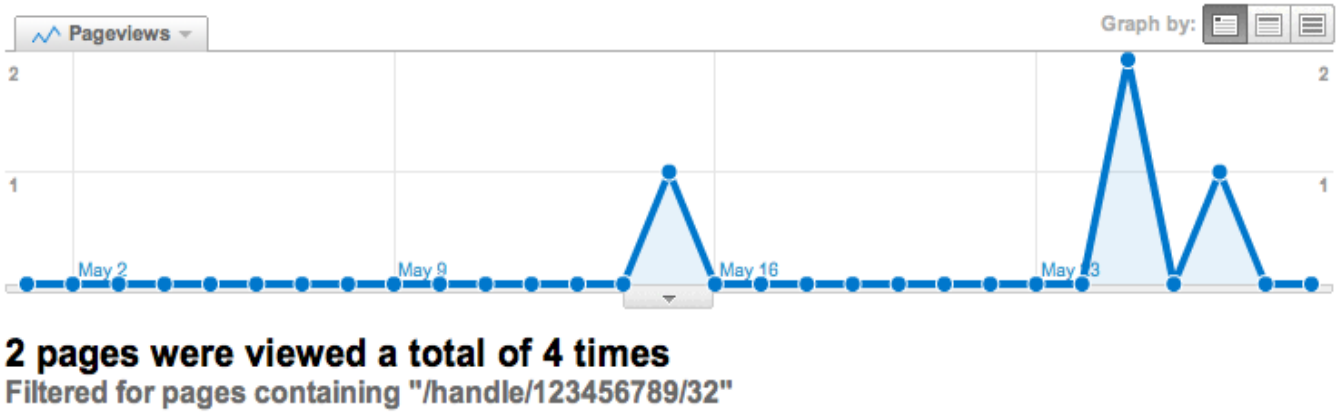
  JAVA_OPTS="-server -Xmx600M -Xms600M
  -XX:+UseParallelGC -Dfile.encoding=UTF-8
  -XX:PermSize=128M -XX:MaxPermSize=192M -d64”
- Defined xmlui.google.analytics.key within dspace.cfg
- SOLR specific settings

  solr.statistics.logBots = false

  solr.statistics.query.filter.spiderIp = false

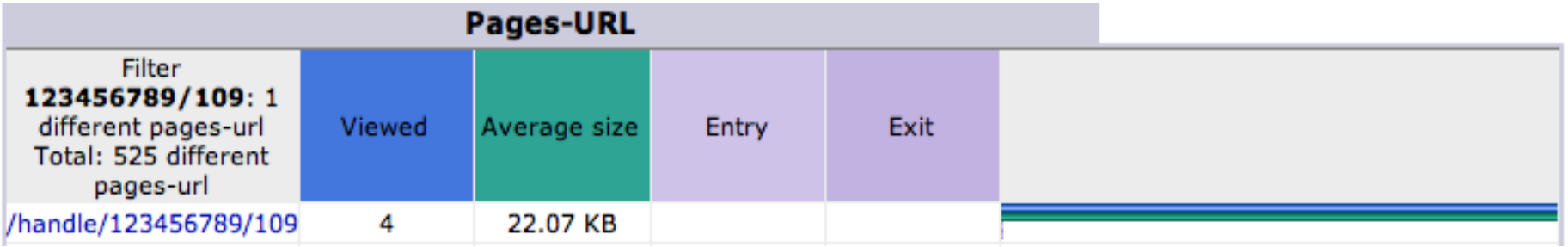
  solr.statistics.query.filter.isBot = true

Candidate I



	SOLR	AWstats	Google Analytics
Page Views	5	5	4
File Visits	104	105	N/A

Candidate II



Pages-URL					
Filter **123456789/109**: 1 different pages-url Total: 525 different pages-url	Viewed	Average size	Entry	Exit	
/handle/123456789/109	4	22.07 KB			

	SOLR	AWstats	Google Analytics
Page Views	4	4	4
File Views	33		N/A

Candidate III

Statistics for DSpace Statistics Testing Unit on gaia.library.gatech.edu

Apr 22, 2011 to Jun 7, 2011

Items Viewed

Top

(more than 20 times)

Item/Handle	Number of views
Tenth and Home site plan (Georgia Institute of Technology. Dept. of Housing) (123456789/101)	52

SOLR
Page Views: 46
File Views: 2

AWstats
Page Views: 47
File Views: N/A

Google Analytics
Page Views: 46
File Views: N/A

Moving Forward

- Outstanding issues
- Refining our reporting capabilities
- Stabilizing Solr
- Displaying statistics to users
- Usability study
- Gathering feedback

Contact

Bill Anderson
bill.anderson@library.gatech.edu

Andy Carter
cartera@uga.edu

Sara Fuchs
sara.fuchs@library.gatech.edu

Chris Helms
chris.helms@library.gatech.edu