@MIRE

DSpace 1.6 usage statistics: How does it work?

Ben Bosman - @mire
1 - Introduction
2 - Technical Overview
3 - User Interface Additions
4 - Advanced Use Cases
5 - High Load Installations
6 - Content & Usage Analysis Add-on Module
1 - Introduction

- Community Survey
  - Highest rated feature request
  - 36% of requests in survey
- @mire Contribution to DSpace 1.6
  - Core of @mire’s Content & Usage Analysis
  - Logging usage events
    - in search index
  - Querying usage events to provide statistics
    - On-the-fly queries instead of predefined reports
  - For XMLUI & JSPUI
1 - Introduction

- In-house storage of usage data:
  - No dependency on external services
    - availability, long-term support, ...
  - No privacy issues
  - Create your own Back-ups

- Storing original usage events
  - No limitations on views of the data
  - View usage data in detail
  - Full history available
  - Not only aggregated (e.g. per month)
2 - Technical Overview

- **Usage events:**
  - Community homepage visits
  - Collection homepage visits
  - Item visits
  - Bitstream downloads

- **Data per usage event:**
  - Timestamp
  - IP address
  - Location: continent, country, city
  - and much more ...
2 - Technical Overview

- Usage event logging
  - Apache Solr
    - Open source enterprise search platform from the Apache Lucene project
    - New web application added to DSpace
  - Performance
    - Fast logging in search index
    - Can easily be deployed on a separate server
    - Advanced solutions for fast querying based on caching
How to store statistics data

- Define the fields to store:
  dspace/solr/statistics/conf/schema.xml

- Actual storage:
  org.dspace.statistics.SolrLogger.post()
  doc.addField("fieldname", fieldvalue);
2 - Technical Overview

<field name="type" type="integer" indexed="true" stored="true" required="true" />
<field name="id" type="integer" indexed="true" stored="true" required="true" />
<field name="ip" type="string" indexed="true" stored="true" required="false" />
<field name="time" type="date" indexed="true" stored="true" required="true" />
<field name="epersonid" type="integer" indexed="true" stored="true" required="false" />
<field name="country" type="string" indexed="true" stored="true" required="false" />
<field name="city" type="string" indexed="true" stored="true" required="false" />
<field name="owningComm" type="integer" indexed="true" stored="true" required="false" multiValued="true" />
2 - Technical Overview

Display of statistics data
- Depends on the User interface
- StatisticsTransformer for XMLUI
- DisplayStatisticsServlet for JSPUI
Display Statistics - XMLUI

- StatisticsTransformer generates the DRI with the statistics information
- Uses multiple StatisticsDisplay objects
  - StatisticsListing for a two column table
  - StatisticsTable for a multiple column table
## Statistics

### Total Visits

<table>
<thead>
<tr>
<th>DSF S XML UI Project Technical Overview</th>
<th>Views</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>71</td>
</tr>
</tbody>
</table>

### Total Visits Per Month

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DSF S XML UI Project Technical Overview</td>
<td>0</td>
<td>40</td>
<td>18</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>5</td>
</tr>
</tbody>
</table>

### File Visits

<table>
<thead>
<tr>
<th>File Name</th>
<th>Views</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSF S XML UI Project Technical Overview.pdf</td>
<td>19</td>
</tr>
<tr>
<td>license.txt</td>
<td>7</td>
</tr>
</tbody>
</table>

### Top country views

<table>
<thead>
<tr>
<th>Country</th>
<th>Views</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States of America</td>
<td>30</td>
</tr>
<tr>
<td>Japan</td>
<td>6</td>
</tr>
<tr>
<td>China</td>
<td>5</td>
</tr>
<tr>
<td>Belgium</td>
<td>4</td>
</tr>
<tr>
<td>Germany</td>
<td>4</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>3</td>
</tr>
<tr>
<td>Canada</td>
<td>2</td>
</tr>
<tr>
<td>Spain</td>
<td>2</td>
</tr>
<tr>
<td>France</td>
<td>2</td>
</tr>
<tr>
<td>India</td>
<td>2</td>
</tr>
</tbody>
</table>

### Top cities views

<table>
<thead>
<tr>
<th>City</th>
<th>Views</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beijing</td>
<td>4</td>
</tr>
<tr>
<td>Dortmund</td>
<td>4</td>
</tr>
<tr>
<td>Seattle</td>
<td>4</td>
</tr>
<tr>
<td>Brussels</td>
<td>2</td>
</tr>
<tr>
<td>Madrid</td>
<td>2</td>
</tr>
</tbody>
</table>
2 - Technical Overview

- Display Statistics - JSPUI
  - DisplayStatisticsServlet
  - Uses StatisticsBean objects to store the information
  - display-statistics.jsp builds tables from StatisticsBean objects
## Technical Overview

### Statistics

#### Total Visits

<table>
<thead>
<tr>
<th>DSpace XML UI Project Technical Overview</th>
<th>Views</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>70</td>
</tr>
</tbody>
</table>

#### Total Visits per Month

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>40</td>
<td>18</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

### File Downloads

<table>
<thead>
<tr>
<th>DSpace XML UI Project Technical Overview.pdf</th>
<th>Views</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>19</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>licence.txt</th>
<th>Views</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>7</td>
</tr>
</tbody>
</table>

### Top Country Views

<table>
<thead>
<tr>
<th>Country</th>
<th>Views</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States of America</td>
<td>30</td>
</tr>
<tr>
<td>Japan</td>
<td>6</td>
</tr>
<tr>
<td>China</td>
<td>5</td>
</tr>
<tr>
<td>Germany</td>
<td>4</td>
</tr>
<tr>
<td>Belgium</td>
<td>3</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>3</td>
</tr>
<tr>
<td>Canada</td>
<td>2</td>
</tr>
<tr>
<td>Spain</td>
<td>2</td>
</tr>
<tr>
<td>France</td>
<td>2</td>
</tr>
<tr>
<td>India</td>
<td>2</td>
</tr>
</tbody>
</table>

### Top City Views

<table>
<thead>
<tr>
<th>City</th>
<th>Views</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beijing</td>
<td>4</td>
</tr>
<tr>
<td>Dortmund</td>
<td>4</td>
</tr>
<tr>
<td>Seattle</td>
<td>4</td>
</tr>
<tr>
<td>Madrid</td>
<td>2</td>
</tr>
<tr>
<td>New Delhi</td>
<td>2</td>
</tr>
<tr>
<td>New Haven</td>
<td>2</td>
</tr>
<tr>
<td>Quebec</td>
<td>2</td>
</tr>
<tr>
<td>San Jose</td>
<td>2</td>
</tr>
<tr>
<td>State College</td>
<td>2</td>
</tr>
<tr>
<td>Adelaide</td>
<td>1</td>
</tr>
</tbody>
</table>
3 - User Interface Additions

- 3 small changes in detail
  - Create repository wide statistics
  - Specify a timespan
  - Separate downloads and page visits
3 - User Interface Additions

- Create repository wide statistics
- Add a Transformer to create overview of e.g. top 3 items.

```java
StatisticsListing statListing = new StatisticsListing(new StatisticsDataVisits(dso));
statListing.setTitle("Top 3 items");
statListing.setId("list-top-items");
DatasetDSpaceObjectGenerator dsoAxis = new DatasetDSpaceObjectGenerator();
dsoAxis.addDsoChild(Contents.ITEM, 3, false, -1);
statListing.addDatasetGenerator(dsoAxis);
addDisplayListing(division, statListing);
```
3 - User Interface Additions

- Specify a timespan
  - Limit displayed statistics data on e.g. item display page to specific timespan.

```java
StatisticsSolrDateFilter dateFilter = new StatisticsSolrDateFilter();
dateFilter.setStartDate(startDate);
dateFilter.setEndDate(endDate);
dateFilter.setTypeStr("month");
statListing.addFilter(dateFilter);
```

- Determine timespan in User Interface
Separate downloads and page visits

Split up item views and file downloads in item display statistics

```java
StatisticsListing statListing = new StatisticsListing(new StatisticsDataVisits(dso));
DatasetDSpaceObjectGenerator dsoAxis = new DatasetDSpaceObjectGenerator();
dsoAxis.addDsoChild(Constants.BITSTREAM, 10, false, -1);
statsList.addDatasetGenerator(dsoAxis);
```
4 - Advanced Use Cases

- Extend the data being stored
- Harvest usage data
- Store popular searches
- Recommendations
4 - Advanced Use Cases

- Extend the data being stored
  - Referrer
    - Store referrer to visualize incoming links from other websites, and internal navigation
    - Referrer URL retrieved from the browser
4 - Advanced Use Cases

- Extend the data being stored
- Fields to be stored are defined
  
dspace/solr/statistics/conf/schema.xml

```xml
<field name="referrer" type="string"
indexed="true" stored="true" required="false" />
```
4 - Advanced Use Cases

- Extend the data being stored
  - Storage is handled by post() method in the SolrLogger class
    ```java
doc1.addField("referrer", request.getHeader("referrer"));
```
  - Extend the user interface to use this new data
4 - Advanced Use Cases

- Harvest Usage Data
  - Goals
    - Mining usage data from partner institutions
    - Compare usage data amongst different institutions
    - Construct cross-institution recommendations based on usage data
  - Example set-ups
    - NEEO
    - PIRUS
4 - Advanced Use Cases

- Store popular searches

  - Goals
    - Display top searches within your repository
    - Display relevant additional search terms to be included or excluded
    - Rank search results based on usage by other visitors

  - Requirements
    - Store executed search terms
    - Store relation amongst search terms
    - Store relation between search terms and opened items
4 - Advanced Use Cases

http://getcloudlet.com/
4 - Advanced Use Cases

- Recommendations
  - Build recommendations solution based on the concept of ‘users who visited this item, also visited …’
5 - High Load Installations

- High load installations
  - Generate large amounts of usage data
  - Slower query execution
  - Response time increases
  - Request time outs if response time is too long
5 - High Load Installations

_solution
- Optimization using Solr server features
- Autocommit system
- Query warmup system
Autocommit feature optimizes storage of usage events
Out of the box, synchronous commits of usage events are used
The autocommit feature enables asynchronous commits of these events
5 - High Load Installations

Autocommit

- Remove `solr.commit()` from `SolrLogger`
- Add the AutoCommit code to the `solrconfig.xml`:
  
  ```xml
  <autoCommit>
    <maxDocs>10000</maxDocs>
    <maxTime>900000</maxTime>
  </autoCommit>
  ```
Query warmup is used to optimize the query execution time

Preheated queries are cached by the Solr server based on the filter query

Queries are being warmed up for the current month

Query warmup required at:
- Server startup
- End of each month
5 - High Load Installations

Query warmup

- Server startup:
  - <listener event="firstSearcher" class="solr.QuerySenderListener"/>

- End of month:
  - Execute all expected queries for next month
5 - High Load Installations

More detailed information about how to improve your performance

- Performance improvements tips page:
  - Register at http://atmire.com/contact.php to be notified when the explanation has been completed
  - Or email info@atmire.com
CUA module

- Designed and developed by @mire
- Module core contributed to DSpace 1.6
- Same data logging
- Improved interface
6 - Content and Usage Analysis Module

- **Statlets**
  - Configurable display in repository
    - Determine the displayed data
    - Separate configuration for item, collection, community, repository homepage

- **Graphs**
  - Generate various types of graphs, and integrate them in the display pages
6 - Content and Usage Analysis

Statistics

Top 5 Countries

<table>
<thead>
<tr>
<th>countryCode</th>
<th>Views</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States of America</td>
<td>152</td>
</tr>
<tr>
<td>Japan</td>
<td>4</td>
</tr>
<tr>
<td>Korea</td>
<td>4</td>
</tr>
<tr>
<td>Germany</td>
<td>2</td>
</tr>
<tr>
<td>France</td>
<td>2</td>
</tr>
</tbody>
</table>

Item Visits

File Downloads

2006 2007 2008 2009 2010

2006 2007 2008 2009 2010
6 - Content and Usage Analysis Module

- Administrator interface
  - Wide range of reports
    - Created instantaneous in the web interface
    - Configure type of report to be requested
    - Fast access to results to verify the configuration
    - Generate data in a few clicks
  - View report as:
    - Data Table
    - Downloadable Spreadsheet
    - Various Graph types
# 6 - Content and Usage Analysis Module

<table>
<thead>
<tr>
<th>Title</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Comparison of Work Motivation in Bulgaria Hungary and the Netherlands Test of a Model</td>
<td>Roe, Robert A, Zinovieva, Irina L, Dienes, Elizabeth, Horn, Laurens A Ten</td>
</tr>
<tr>
<td>Aanbod van arbeid en werklozen per bedrijfsklasse in Nederland, 1965-1981</td>
<td>Muysken, J., Nijkamp, M</td>
</tr>
</tbody>
</table>

Showing 1 to 3 of 3 items

Included usage data:  
- Item Page views  
- Bitstream Downloads  

[Generate Data]
6 - Content and Usage Analysis Module

Show usage statistics for: [Select]

Show 10 items

Title
- A Comparison of Work Motivation in Bulgaria, Hungary and the Netherlands: Test of a Model
- A response to ‘Nanotechnology and the need for risk governance’ O. Renn & M.C. Roco 2006. J. Nanoparticle Research 8(2) 153-191
- Aanbod van arbeid en werklozen per bedrijfklasse in Nederland, 1965-1981

Graph Type: Bars
Orientation: Vertical
Legend: South

Item Hits / Month

A Comparison of Work Motivation in Bulgaria, Hungary and the Netherlands: Test of a Model
A response to ‘Nanotechnology and the need for risk governance’ O. Renn & M.C. Roco 2006. J. Nanoparticle Research 8(2) 153-191
Aanbod van arbeid en werklozen per bedrijfklasse in Nederland, 1965-1981
6 - Content and Usage Analysis Module

- Content analysis
  - Visualize repository growth
    - Display amount of records per year
    - Compare growth amongst various communities
  - Visualize distribution
    - Display amount of records per type, language, ...
6 - Content and Usage Analysis Module

Data Source Selection

Primary Dataset: Time
Start: 2000
End: 2010

Secondary Dataset: Type
- All
- None
- Book
- Article
- BookPart
- Lecture
- ObjectFile
- WorkingPaper
- DoctoralThesis
- BookReview
- All
- None

Get Data
6 - Content and Usage Analysis

Data Source Selection

Primary Dataset: Time
Start: 2000, End: 2010

Secondary Dataset: Type
All, None
- Book
- Article
- BookPart
- Lecture
- ObjectFile
- WorkingPaper
- DoctoralThesis
- BookReview

Get Data

Preferences

Graph Type: Bars
Orientation: Vertical
Legend: South
Show more options

Graph Settings

Graph Title

Records

Books, Articles, Book Parts, Lectures, Object Files, Working Papers, Doctoral Theses, Book Reviews

Charts for 2000 to 2010
Thank you

Questions?

ben@atmire.com