# DSpace Statistics

DSpace 1.6 and newer versions uses the Apache SOLR application underlying the statistics. SOLR enables performant searching and adding to vast amounts of (usage) data.
Unlike previous versions, enabling statistics in DSpace does not require additional installation or customization. All the necessary software is included.



## What is exactly being logged ?

Each time a page or file gets requested, this request is being logged. The logging happens at the server side, and doesn't require a javascript like Google Analytics does, to provide usage data.

Definition of which fields are to be stored happens in the file dspace/solr/statistics/conf/schema.xml.
The fields, stored in a usage event by default are:

```
<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="continent" type="string" indexed="true" stored="true" required="false"/>
<field name="country" type="string" indexed="true" stored="true" required="false"/>
<field name="countryCode" type="string" indexed="true" stored="true" required="false"/>
<field name="city" type="string" indexed="true" stored="true" required="false"/>
<field name="longitude" type="float" indexed="true" stored="true" required="false"/>
<field name="latitude" type="float" indexed="true" stored="true" required="false"/>
<field name="owningComm" type="integer" indexed="true" stored="true" required="false" multiValued="true"/>
<field name="owningColl" type="integer" indexed="true" stored="true" required="false" multiValued="true"/>
<field name="owningItem" type="integer" indexed="true" stored="true" required="false" multiValued="true"/>
<field name="dns" type="string" indexed="true" stored="true" required="false"/>
<field name="userAgent" type="string" indexed="true" stored="true" required="false"/>
<field name="isBot" type="boolean" indexed="true" stored="true" required="false"/>
<field name="bundleName" type="string" indexed="true" stored="true" required="false" multiValued="true" />
```

The combination of type and id determine which resource (either community, collection, item page or file download) has been requested.

## Web user interface for DSpace statistics

In the XMLUI, statistics can be accessed from the lower end of the navigation menu. In the JSPUI, a view statistics button appears on the bottom of pages for which statistics are available.

If you are not seeing these links or buttons, it's likely that they are only enabled for administrators in your installation. Change the configuration parameter "statistics.item.authorization.admin" to false in order to make statistics visible for all repository visitors.

### Home page

Starting from the repository homepage, the statistics page displays the top 10 most popular items of the entire repository.

### Community home page

The following statistics are available for the community home pages:

- Total visits of the current community home page
- Visits of the community home page over a timespan of the last 7 months
- Top 10 country from where the visits originate
- Top 10 cities from where the visits originate

### Collection home page

The following statistics are available for the collection home pages:

- Total visits of the current collection home page
- Visits of the collection home over a timespan of the last 7 months
- Top 10 country from where the visits originate
- Top 10 cities from where the visits originate

### Item home page

The following statistics are available for the item home pages:

- Total visits of the item
- Total visits for the bitstreams attached to the item
- Visits of the item over a timespan of the last 7 months
- Top 10 country views from where the visits originate
- Top 10 cities from where the visits originate

## Usage Event Logging and Usage Statistics Gathering

The DSpace Statistics Implementation is a Client/Server architecture based on Solr for collecting usage events in the JSPUI and XMLUI user interface applications of DSpace.  Solr runs as a separate webapplication and an instance of Apache Http Client is utilized to allow parallel requests to log statistics events into this Solr instance.

## Configuration settings for Statistics

In the {dspace.dir}/config/modules/solr-statistics.cfg file review the following fields to make sure they are uncommented:

| | |
|---|---|
| Property: | server |
| Example Value: | server = http://127.0.0.1/solr/statistics |
| Informational Note: | Is used by the SolrLogger Client class to connect to the Solr server over http and perform updates and queries. In most cases, this can (and should) be set to localhost (or 127.0.0.1).<br><br>To determine the correct path, you can use a tool like `wget` to see where Solr is responding on your server. For example, you'd want to send a query to Solr like the following:<br><br>`wget http://127.0.0.1/solr/statistics/select?q=*:*`<br><br>Assuming you get an HTTP 200 OK response, then you should set `solr.log.server` to the '/statistics' URL of 'http://127.0.0.1/solr/statistics' (essentially removing the "/select?q=**:**" query off the end of the responding URL.) |
| Property: | spiderips.urls |

| | |
|---|---|
| Example Value: | spiderips.urls =<br>`http://iplists.com/google.txt, \`<br>`http://iplists.com/inktomi.txt, \`<br>`http://iplists.com/lycos.txt, \`<br>`http://iplists.com/infoseek.txt, \`<br>`http://iplists.com/altavista.txt, \`<br>`http://iplists.com/excite.txt, \`<br>`http://iplists.com/misc.txt, \`<br>`http://iplists.com/non_engines.txt` |
| Informational Note: | List of URLs to download spiders files into [dspace]/config/spiders. These files contain lists of known spider IPs and are utilized by the SolrLogger to flag usage events with an "isBot" field, or ignore them entirely.<br><br>The "stats-util" command can be used to force an update of spider files, regenerate "isBot" fields on indexed events, and delete spiders from the index. For usage, run:<br><br>`dspace stats-util -h`<br><br>from your [dspace]/bin directory |
| Property: | dbfile |
| Example Value: | dbfile = ${dspace.dir}/config/GeoLiteCity.dat |
| Informational Note: | The following referes to the GeoLiteCity database file utilized by the LocationUtils to calculate the location of client requests based on IP address. During the Ant build process (both fresh_install and update) this file will be downloaded from http://www.maxmind.com/app/geolitecity if a new version has been published or it is absent from your [dspace]/config directory. |
| Property: | resolver.timeout |
| Example Value: | resolver.timeout = 200 |
| Informational Note: | Timeout in milliseconds for DNS resolution of origin hosts/IPs. Setting this value too high may result in solr exhausting your connection pool. |
| Property: | useProxies |
| Example Value: | useProxies = true |
| Informational Note: | Will cause Statistics logging to look for X-Forward URI to detect clients IP that have accessed it through a Proxy service (e.g. the Apache mod_proxy). Allows detection of client IP when accessing DSpace. [Note: This setting is found in the DSpace Logging section of dspace.cfg] |
| Property: | statistics.item.authorization.admin |
| Example Value: | statistics.item.authorization.admin = true |
| Informational Note: | When set to true, only general administrators, collection and community administrators are able to access the statistics from the web user interface. As a result, the links to access statistics are hidden for non logged-in admin users. Setting this property to "false" will display the links to access statistics to anyone, making them publicly available. |
| Property: | solr.statistics.logBots |
| Example Value: | solr.statistics.logBots = true |
| Informational Note: | When this property is set to false, and IP is detected as a spider, the event is not logged.<br>When this property is set to true, the event will be logged with the "isBot" field set to true.<br>(see solr.statistics.query.filter.* for query filter options) |

| | |
|---|---|
| Property: | solr.statistics.query.filter.spiderIp |
| Example Value: | solr.statistics.query.filter.spiderIp = false |
| Informational Note: | If true, statistics queries will filter out spider IPs -- use with caution, as this often results in extremely long query strings. |
| Property: | solr.statistics.query.filter.isBot |
| Example Value: | solr.statistics.query.filter.isBot = true |
| Informational Note: | If true, statistics queries will filter out events flagged with the "isBot" field. This is the recommended method of filtering spiders from statistics. |
| Property: | query.filter.bundles |
| Example Value: | query.filter.bundles=ORIGINAL |
| Informational Note: | A comma seperated list that contains the bundles for which the file statistics will be displayed. |

### Upgrade Process for Statistics

Example of rebuild and redeploy DSpace (only if you have configured your distribution in this manner)

First approach the traditional DSpace build process for updating

```
cd [dspace-source]/dspace
 mvn package
 cd [dspace-source]/dspace/target/dspace-<version>-build.dir
 ant -Dconfig=[dspace]/config/dspace.cfg update
 cp -R [dspace]/webapps/* [TOMCAT]/webapps
```

The last step is only used if you do not follow the recommended practice of configuring *[dspace]/webapps* as location for webapps in your servlet container (Tomcat, Resin or Jetty). If you only need to build the statistics, and don't make any changes to other web applications, you can replace the copy step above with:

```
cp -R dspace/webapps/solr TOMCAT/webapps
```

*Again, only if you are not mounting [dspace]/webapps directly into your Tomcat, Resin or Jetty host (the recommended practice)*

Restart your webapps (Tomcat/Jetty/Resin)

## Older setting that are not related to the new 1.6 Statistics

The following Dspace.cfg fields are only applicable to the older statistics solution.

```
###### Statistical Report Configuration Settings ######

 # should the stats be publicly available?  should be set to false if you only
 # want administrators to access the stats, or you do not intend to generate
 # any
 report.public = false

 # directory where live reports are stored
 report.dir = ${dspace.dir}/reports/
```

These fields are not used by the new 1.6 Statistics, but are only related to the Statistics from previous DSpace releases

# Statistics Administration

## Converting older DSpace logs into SOLR usage data

If you have upgraded from a previous version of DSpace, converting older log files ensures that you carry over older usage stats from before the upgrade.

## Statistics Client Utility

The command line interface (CLI) scripts can be used to clean the usage database from additional spider traffic and other maintenance tasks.

# Statistics differences between DSpace 1.7.x and 1.8.0

### Displayed file statistics bundle configurable

In DSpace 1.6.x & 1.7.x the file download statistics were generated without regard to the bundle in which the file was located. In DSpace 1.8.0 it is possible to configure the bundles for which the file statistics are to be shown by using the **query.filter.bundles** property. If required the old file statistics can also be upgraded to include the bundle name so that the old file statistics are fixed.

Backup Your statistics data first

Applying this change will involve dumping all the old file statistics into a file and re uploading these. Therefore it is wise to create a backup of the {dspace.dir}/solr/statistics/data directory. It is best to create this backup when the Tomcat/Jetty/Resin server program isn't running.

When a backup has been made start the Tomcat/Jetty/Resin server program.
The update script has one optional command which will if given not only update the broken file statistics but also delete file statistics for files that where removed from the system (if this option isn't active these statistics will receive the "BITSTREAM_DELETED" bundle name).

```
#The -r is optional
[dspace]/bin/dspace stats-util -b -r
```

# Statistics differences between DSpace 1.6.x and 1.7.0

### SOLR optimization added

If required, the solr server can be optimized by running

```
{dspace.dir}/bin/stats-util -o
```

More information on how these solr server optimizations work can be found here: http://wiki.apache.org/solr/SolrPerformanceFactors#Optimization_Considerations.

### SOLR Autocommit

In DSpace 1.6.x, each solr event was committed to the solr server individually. For high load DSpace installations, this would result in a huge load of small solr commits resulting in a very high load on the solr server.
This has been resolved in dspace 1.7 by only committing usage events to the solr server every 15 minutes. This will result in a delay of the storage of a usage event of maximum 15 minutes. If required, this value can be altered by changing the maxTime property in the

```
{dspace.dir}/solr/statistics/conf/solrconfig.xml
```

# Web UI Statistics Modification (XMLUI Only)

### Modifying the number of months, for which statistics are displayed

Modify line 178 in the StatisticsTransformer.java file

dspace-xmlui/dspace-xmlui-api/src/main/java/org/dspace/app/xmlui/aspect/statistics/StatisticsTransformer.java

-6 is the default setting, displaying the past 6 months of statistics. When reducing this to a smaller natural number, less months are being displayed.

Related: DatasetTimeGenerator Javadoc

# Custom Reporting - Querying SOLR Directly

When the web user interface does not offer you the statistics you need, you can greatly expand the reports by querying the SOLR index directly.

## Resources

- http://www.lucidimagination.com/Community/Hear-from-the-Experts/Articles/Faceted-Search-Solr
- http://my.safaribooksonline.com/9781847195883/Cover

## Examples

### Top downloaded items by a specific user

**Query**:

```
http://localhost:8080/solr/statistics/select?indent=on&version=2.2&start=0&rows=10&fl=*%
2Cscore&qt=standard&wt=standard&explainOther=&hl.fl=&facet=true&facet.field=epersonid&q=type:0
```

**Explained**:

facet.field=epersonid — You want to group by epersonid, which is the user id.
type:0 — Interested in bitstreams only

```
<lst name="facet_counts">
    <lst name="facet_fields">
        <lst name="epersonid">
            <int name="66">1167</int>

<int name="117">251</int>

<int name="52">42</int>

<int name="19">36</int>

<int name="88">20</int>

<int name="112">18</int>

<int name="110">9</int>

<int name="96">0</int>

</lst>
    </lst>
</lst>
```