Elasticsearch Usage Statistics

Elasticsearch Usage Statistics feature is deprecated in 6.0

As of the DSpace 6.0 release, the Elasticsearch Usage Statistics feature is deprecated, and no new enhancements/improvements will be added. We recommend instead using the default SOLR Statistics engine and/or DSpace's integration with Google Analytics.

If you are already using Elasticsearch Usage Statistics, you may continue to use them until they are fully removed in a future release. See DS-2897 for more information.

Added in DSpace 3.0 is an optional statistics engine using Elasticsearch, which may be enabled as an alternative to the default SOLR Statistics engine (based on Apache SOLR). The motivation for adding Elasticsearch was to find an alternative statistics processing engine that would handle the workload of a large amount of statistics data. Additionally, the Elasticsearch statistics display offers another method for creating statistical queries against your data. Elasticsearch Usage Statistics has been contributed by Peter Dietz of Ohio State University's Knowledge Bank. The data source for Elasticsearch Statistics is DSpace Usage Events, where Usage Event is a view or download of a DSpace Object (Bitstream, Item Page, Collection Page, Community Page). Elasticsearch Statistics is bundled with DSpace, and requires no additional installation of software, it just needs to be enabled. Elasticsearch is only available for use with XMLUI.

What data is being recorded?

The default information below is what DSpace will record about a Usage Event. In DSpace 3.0 the fields of data collected is not configurable through a configuration setting.

Information about the User Requesting the Content

- IP Address
- Time of Request
- DNS / Hostname
- User Agent
- isBot, a flag that DSpace thinks that user is a robot or not
- Geographical Information about where the user is located:
    - Continent
    - Country
    - Country Code
    - City
    - Geographical Latitude/Longitude

Information about the DSpace Resource that was used

- DSpace Object ID
- DSpace Object Type: (Item, Bitstream, Collection, or Community)
- If it is relevant, we also store the hierarchy of where this object exists within DSpace
    - Owning Community
    - Owning Collection
    - Owning Item

Enabling Elasticsearch Statistics

Elasticsearch Statistics is disabled by default in DSpace 3.0, the following steps will enable Elasticsearch so that you can collect data, and present statistics reports.

Modify dspace/config/xmlui.xconf, and uncomment the aspect, Statistics Elasticsearch.

Enable Elastic Search Statistics in dspace/config/xmlui.xconf

```
<!--
    If you prefer to use "Elasticsearch" Statistics, you can uncomment the below
    aspect and COMMENT OUT the default "Statistics" aspect above.
    You must also enable the ElasticSearchLoggerEventListener.
-->
<!-- <aspect name="Statistics - Elasticsearch" path="resource://aspects/StatisticsElasticSearch/" /> -->
```

Modify dspace-xmlui/src/main/webapp/WEB-INF/spring/applicationContext.xml and uncomment the following code block for ElasticSearchLoggerEventListener

Enable ElasticSearchLoggerEventListener

```
<!-- Elasticsearch -->
<!--<bean class="org.dspace.statistics.ElasticSearchLoggerEventListener">
        <property name="eventService">
            <ref bean="dspace.eventService" />
        </property>
    </bean>-->
```

After making these two changes, you will then need to rebuild and restart DSpace.

Importing Legacy Data into Elasticsearch Statistics

Once Elasticsearch Statistics has been enabled, it will begin adding all new Usage Events to its data store. To import your legacy data, you will need to import the data from the dspace.log files. There is no tool yet that converts SOLR statistics data to Elasticsearch statistics data.

From the (Windows / Linux) terminal, you will need to use the DSpace Command Launcher to convert the dspace.log files to a statistics log format. Then you will need to import the statistics log format files into DSpace Statistics.

The Log Converter program converts log files from dspace.log into an intermediate format that can be inserted into Elasticsearch Statistics.

Command used:	`[dspace]/bin/dspace stats-log-converter`
Java class:	org.dspace.statistics.util.ClassicDSpaceLogConverter
Arguments short and long forms):	Description
`-i or --in`	Input file
`-o or --out`	Output file
`-m or --multiple`	Adds a wildcard at the end of input and output, so it would mean if `-i dspace.log -m` was specified, dspace.log* would be converted. (i.e. all of the following: *dspace.log, dspace.log.1, dspace.log.2, dspace.log.3,* etc.)
`-n or --newformat`	If the log files have been created with DSpace 1.6 or newer
`-v or --verbose`	Display verbose output (helpful for debugging)
`-h or --help`	Help

An example form of this command would be `[dspace]/bin/dspace stats-log-converter -i dspace.log -o statistics.log -m -n`

The Log Importer program takes the intermediate format data produced in the previous step, and imports it into Elasticsearch Statistics.

Command used:	`[dspace]/bin/dspace stats-log-importer-elasticsearch`
Java class:	org.dspace.statistics.util.StatisticsImporterElasticSearch
Arguments short and long forms):	Description
`-i or --in`	Input file

`-m or --multiple`	Adds a wildcard at the end of input and output, so it would mean if `-i statistics.log -m` was specified, statistics.log* would be imported. (i.e. all of the following: *statistics.log, statistics.log.1, statistics.log.2, statistics.log.3,* etc.)
`-s or --skipdns`	To skip the reverse DNS lookups that work out where a user is from. (The DNS lookup finds the information about the host from its IP address, such as geographical location, etc. This can be slow, and wouldn't work on a server not connected to the internet.)
`-v or --verbose`	Display verbose output (helpful for debugging)
`-h or --help`	Help

An example form of this command would be `[dspace]/bin/dspace stats-log-importer-elasticsearch -i statistics.log -m`

Viewing Data in Elasticsearch Statistics

In XMLUI, while logged in as an administrator, the Context Panel will have an additional "View Statistics" link when you browse to a Community, Collection, or Item.

The Statistics Report includes:

- Bitstreams with Most Downloads, for all time.
- Bitstreams with Most Downloads, previous month.
- Total Number of Downloads to Bitstreams within this container, broken down by month.
- Number of hits per Country

This data is presented as either a Table or Line Graph, and requires JavaScript to draw the graphics.