

Statistics Import Export Issues

- Update 2017-02-08
- Data Corruption Issue (DSpace 6)
- Shard Testing Issues
- Statistics Import/Export Tool Issues
- Manual Repair of Corrupted Export Files
- Testing Solr
 - Testing CSV Export
 - Testing CSV Import
 - Creating a Shard in the Admin Console

There are a number of issues with the Statistics Sharding process in DSpace.

Update 2017-02-08

- JIRA issu
 -
 -
 -
 - Unable to locate Jira server for this macro. It may be due to Application Link configuration.
- Document
 - DSpace 6x : [SOLR Statistics Maintenance](#)
 - The warning section on this page should be referenced in the 6.1 and 7.0 release notes
 - Additional testing documentation for shards: [Testing Solr Shards](#)
 - DSpace 5x : [SOLR Statistics Maintenance](#)
 - The warning section on this page should be referenced in the 5.7 release notes
- PR's needing a Merge
 - Allow Shard Overwrite
 -
 - Unable to locate Jira server for this macro. It may be due to Application Link configuration.
 - Prevent r
 - 5x: <https://github.com/DSpace/DSpace/pull/1645>
 - master: <https://github.com/DSpace/DSpace/pull/1644>

Data Corruption Issue (DSpace 6)

1. In DSPac
 - a. Unable to locate Jira server for this macro. It may be due to Application Link configuration.
 - b. try to reproduce this error.
 - c. **How to test**
 - i. Back up your solr directory
 - ii. If you have statistics from 2016 or earlier in your statistics repository
 1. Run stats-util -s to shard last years records into a shard
 - iii. If you do not have statistics from 2016 in your repository
 1. See the instructions related to testing PRs 1623/1624 to force old stats records (from a prior year) into your statistics repository
 2. Run stats-util -s
 - iv. Restart tomcat
 - v. Tomcat will not restart properly. The tomcat process will not respond to a stop request.
 - vi. To resolve this issue
 1. kill -9 your tomcat process
 2. delete the statistics shard directories
 3. restart tomcat
2. In DSPac
 - a. Unable to locate Jira server for this macro. It may be due to Application Link configuration.
 - b. the other issues listed here have been discovered.
 - c. Ultimately, we need the ability to repair statistics shards that contain a corrupted owningComm (perhaps via the solr-reindex-statistics command)

While attempting to resolve this issue, a number of long standing challenges with the sharding process have become evident.

Shard Testing Issues

1. The shard process requires statistics records from a prior calendar year to be present.
 - a. Proposal: Ensure that the statistics import/export tools allow for the creation of records from a prior year.
 - i. See "Statistics Import/Export Tool Issues"

2. Once the shard process has been run for records from a calendar year, the process cannot be re-run

a. Proposal

- i.
- ii.

Unable to locate Jira server for this macro. It may be due to Application Link configuration.

1. run stats-util -s to create shards

2. import old records (from a prior year where a shard already exists) into the statistics repository

3. run stats-util -s again

- a. Without this PR, the action should fail because the shard exists
- b. With this PR the action should succeed

b. Pull Requests

- i. DSpace 5x PR: <https://github.com/DSpace/DSpace/pull/1625>
- ii. DSpace 6x PR: <https://github.com/DSpace/DSpace/pull/1633>
- iii. DSpace master PR: <https://github.com/DSpace/DSpace/pull/1634>

Statistics Import/Export Tool Issues

Make solr-import-statistics, solr-export-statistics, and solr-reindex-statistics easier to use

Unable to locate Jira server for this macro. It may be due to Application Link configuration.

1. Issues

- a. The import and export tool always assume that the main statistics repo is being processed making it difficult to successfully process an individual shard.
- b. The import tool often fails when attempting to import records due to `_version` issues.
- c. Error messages are confusing from these tools.
- d. The export and re-index tools often fail due to the presence of existing export files
- e. The reindex process fails on a

i.

as configuration.

Unable to locate Jira server for this macro. It may be due to Application Link

f. The reindex process fails on multi-value fields like `owningComm`

g. Shard names are on by one character

i.

as configuration.

Unable to locate Jira server for this macro. It may be due to Application Link

2. Proposed Change

- a. Proposal 1: Do not force the inclusion of a `"-i statistics"` parameter to the function. Rather, set `"-i statistics"` as a default when no `"-i"` parameter is found.

i. How to test (solr-export-statistics)

- 1. You will need a shard. If you do not have one, See Proposal 2 to facilitate the creation of a shard.
- 2. Clear the solr-export directory
- 3. Run `solr-export-statistics -i statistics-xxxx`
- 4. Without this PR, you will notice that both statistics and statistics-xxxx are exported
- 5. With this PR, you will notice that only statistics-xxxx is exported

ii. How to test (solr-export-statistics)

- 1. You will need a shard. If you do not have one, See Proposal 2 to facilitate the creation of a shard.
- 2. Clear the solr-export directory
- 3. Run `solr-export-statistics -i statistics-xxxx -i statistics`
- 4. Both statistics and statistics-xxxx are exported
- 5. Without this PR

a. Run `solr-import-statistics -i statistics-xxxx`

b. You will notice that both statistics and statistics-xxxx are imported (or attempted to be imported)

6. With this PR,

a. Run `solr-import-statistics -i statistics-xxxx -f`

b. You will notice that only statistics-xxxx is imported

iii. How to test (solr-reindex-statistics)

1. See Proposals 5 and 6 for testing instructions

b. Proposal 2: Make the import process more tolerant during record ingest

i. How to test

- 1. Clear the solr-export directory
- 2. run `"solr-export-statistics -i statistics"`
- 3. extract the top 3-5 lines from the export file saving it to a new file matching the naming convention (for instance make the output file for `statistics_export_2017-01.csv` be `statistics_export_2017-02.csv`)
 - a. Edit the identifier on each record (it is a uuid, so just edit with an alphanumeric character)
- 4. run `"solr-import-statistics -i statistics"`
 - a. Note that the process fails with a `"_version_"` error
- 5. Install the PR and run `"solr-import-statistics -i statistics"`
 - a. The records should import successfully
- 6. NOTE: to force records from a prior year, repeat this process modifying the record date to use a prior year

c. Proposal 3: Make import/export failure messages more explicit. Include the repository, the export file, and the reason for failure in error and log messages.

i. How to test

- 1. Clear the solr-export directory
- 2. run `"solr-export-statistics -i statistics"`
- 3. run `"solr-export-statistics -i statistics"`

- a. The second time this command is run, you will see an error message warning
 - b. Without the PR, the error message will be unclear
 - c. With this PR, the error message will clearly indicate that the export file cannot be overwritten
 - d. Proposal 4: Add a command line option allowing export files to be overwritten on export.
 - i. How to test
 - 1. Clear the solr-export directory
 - 2. run "solr-export-statistics -i statistics"
 - 3. run "solr-export-statistics -i statistics"
 - a. The process will fail
 - 4. run "solr-export-statistics -i statistics -f"
 - a. The export file will be overwritten
 - e. Proposal 5: Add a command line option allowing export files to be overwritten on re-index
 - i. How to test
 - 1. Clear the solr-export directory
 - 2. run "solr-reindex-statistics -i statistics"
 - 3. run "solr-reindex-statistics -i statistics"
 - a. The process will fail due to the existence of an export file
 - 4. run "solr-reindex-statistics -i statistics -f"
 - a. The export file will be overwritten
 - f. Proposal 6: Set the correct "instanceDir" for statistics shards (since the config files reside in the "statistics" directory)
 - i. How to test
 - 1. Clear the solr-export directory
 - 2. run "solr-reindex-statistics -i statistics-xxxx"
 - g. Proposal 7: Correctly re-index multi-value fields such as owningComm
 - i. How to test
 - 1. View an item with multiple owning communities in DSpace
 - 2. Find the item view record in the Solr Admin console
 - 3. Notice that owningComm is an array
 - 4. run "solr-reindex-statistics -i statistics"
 - 5. Find the item view record in the Solr Admin console
 - 6. owningComm should still be an array with multiple values
 - a. Without the fix, owningComm is a string separated by commas
 - h. Proposal 8: Repair multi-value fields in a shard that were corrupted by prior sharding or prior reindex operations
 - i. How to test
 - 1. In the Solr Admin Console, look for owningComm fields containing either "," or "\"
 - a. Note the id's or other identifying information for the records
 - 2. run "solr-reindex-statistics -i statistics-xxxx"
 - 3. Find the records again in the Solr Admin Console
 - 4. If problems exist, run
 - a. solr-export-statistics -i statistics-xxxx -f
 - b. for file in *; do sed -E -e "s/[\\]+/,/g" -i \$file; done
 - c. solr-import-statistics -i statistics-xxxx
 - 5. The owningComm fields should be an array
 - i. Proposal 9: Consistently use UTC from statistics records to determine shard name
 - i. How to test
 - 1. If not in UTC, create a statistic record for a shard that does not exist
 - 2. run shard process without the PR
 - a. Note the shard name is off by one year
 - b. Test results may vary based on your time zone relative to UTC
 - 3. Repeat the process with the PR in place
 - a. Note that the shard name matches the year of the records
3. Pull Requests
- a. DSpace 5x PR: <https://github.com/DSpace/DSpace/pull/1623/files>
 - b. DSpace 6x PR: <https://github.com/DSpace/DSpace/pull/1624/files>
 - c. DSpace master PR: <https://github.com/DSpace/DSpace/pull/1635>

Manual Repair of Corrupted Export Files

- Use solr-export-statistics to export a repo
- Run the following to repair records

for file in *; do sed -E -e "s/[\\]+/,/g" -i \$file; done

- Run solr-import-statistics to import the fixed records

Testing Solr

Testing CSV Export

The SOLR Admin Console provides a mechanism to test the CSV Export Process and Parameters

Apache Solr

Dashboard
Logging
Core Admin
Java Properties
Thread Dump
statistics-2006
Overview
Analysis
Dataimport
Documents
Files
Ping
Plugins / Stats
Query
Replication
Schema Browser

Request-Handler (qt)
/select

common

q

fq

sort

start, rows
0 10

fl
uid,time,owningComm

df

Raw Query Parameters
key1=val1&key2=val2

wt
csv

☒ indent
☐ debugQuery
☐ dismax
☐ edismax
☐ hl
☐ facet
☐ spatial
☐ spellcheck

Execute Query

https://solr/statistics-2006/select?q=%3A*&fl=uid%2Ctime%2CowningComm&wt=csv

CSV Output is returned

uid,time,owningComm
32c0ac9b-c30d-482e-ae57-e73132800ab4,2007-04-20T14:08:03.789Z,"4,1,1"

3 fields chosen to simplify display

Triggers CSV output

Multi-value field is returned with a comma separator (default) and a double quote encapsulator (default when separator is present)

Testing CSV Import

The SOLR Admin Console provides a mechanism to access the CSV Upload process. Unfortunately, it does not all parameters to be provided.

Apache Solr

Dashboard
Logging
Core Admin
Java Properties
Thread Dump
statistics-2006
Overview
Analysis
Dataimport
Documents
Files
Ping
Plugins / Stats
Query
Replication
Schema Browser

Request-Handler (qt)
/update

Document Type
CSV

Document(s)
uid,time,owningComm
32c0ac9b-c30d-482e-ae57-e73132800ab4,2007-04-20T14:08:03.789Z,"4,1,1"

Commit Within
1000

Overwrite
true

Submit Document

Status: success
Response:
{
 "responseHeader": {
 "status": 0,
 "QTime": 32
 }
}

Trigger CSV import

Note that there is no way to set import params

Documentation Issue Tracker IRC Channel Community forum Solr Query Syntax

Note that the multi-value field is corrupted if you import by this manner.

- Dashboard
- Logging
- Core Admin
- Java Properties
- Thread Dump
- statistics-2006
 - Overview
 - Analysis
 - Dataimport
 - Documents
 - Files
 - Ping
 - Plugins / Stats
 - Query
 - Replication
 - Schema Browser

fq

sort

start, rows

0 10

fl

df

Raw Query Parameters

key1=val1&key2=val2

wt

json

☒ indent

☐ debugQuery

☐ dismax

☐ edismax

☐ hl

☐ facet

☐ spatial

☐ spellcheck

Execute Query

```

"q": "",
"_:": "1485464107796",
"wt": "json"
}
},
"response": {
  "numFound": 1,
  "start": 0,
  "docs": [
    {
      "uid": "32c0ac9b-c39d-482e-ae57-e73132800ab4",
      "time": "2007-04-20T14:08:03.789Z",
      "owningComm": [
        "4,1,1"
      ],
      "_version_": 1557621909527462000,
      "statistics_type": "view"
    }
  ]
}

```

Note that owningComm is no longer an array

[Documentation](#)
[Issue Tracker](#)
[IRC Channel](#)
[Community forum](#)
[Solr Query Syntax](#)

It is possible to csv import parameters using curl.

Running CSV Upload with curl

```
curl -F "data=@statistics-2006_export_2007-04.csv" "http://localhost/solr/statistics-2006/update/csv?skip=_version_&csv.mv.escape=%5C&f.owningColl.split=true&f.owningColl.separator=%7C&f.owningComm.split=true&f.owningComm.separator=,&f.owningItem.split=true&f.owningItem.separator=%7C&f.bundleName.split=true&f.bundleName.separator=%7C&stream.contentType=text%2Fcsv%3Bcharset%3Dutf-8&commit=true&softCommit=false&waitSearcher=true&wt=java&version=2"
```

Creating a Shard in the Admin Console

While this is probably not necessary, it is possible to create an empty shard in the Solr Admin console.

Note that existing shards use the statistics directory as an "instance" directory.

- Dashboard
- Logging
- Core Admin
- Java Properties
- Thread Dump
- Core Selector

+ Add Core

authority

oai

search

statistics

statistics-2006

statistics-2007

statistics-2010

statistics-2011

statistics-2012

statistics-2013

statistics-2015

✖ Unload

📄 Rename

🔄 Swap

🔄 Reload

🔧 Optimize

📁 Core

startTime: 41 minutes ago

instanceDir: /opt/dspace/solr/statistics/

dataDir: /opt/dspace/solr/statistics-2006/data/

📊 Index

lastModified: 8 months ago

version: 115

numDocs: 1

maxDoc: 1

deletedDocs: -

optimized: ✓

current: ✓

Manually create a new shard



Dashboard

Logging

Core Admin

Java Properties

Thread Dump

Core Selector

+ Add Core

✖ Unload

🔄 Rename

name: statistics-1976

instanceDir: statistics

dataDir: [dspace-install]/solr/statistics-1976

config: solrconfig.xml

schema: schema.xml

ⓘ instanceDir and dataDir need to exist before you can create the core

✔ Add Core

✖ Cancel

statistics-2015

optimized:

The new shard can be queried like the other ones



Dashboard

Logging

Core Admin

Java Properties

Thread Dump

statistics-1976

Overview

Analysis

Dataimport

Documents

Files

Ping

Plugins / Stats

Query

Replication

Schema Browser

/select

— common —

q

:

fq

sort

start, rows

0

10

fl

df

Raw Query Parameters

key1=val1&key2=val2

wt

json

☒ indent

☐ debugQuery

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 1,
    "params": {
      "indent": "true",
      "q": "*:*",
      "_": "1485466236243",
      "wt": "json"
    }
  },
  "response": {
    "numFound": 0,
    "start": 0,
    "docs": []
  }
}
```