Solr statistics migration after AIP export/import

Problem

After the entire site is exported and imported via AIP, internal identifiers of DSpace objects (bitstreams, items, collections and communities) will have changed. This is because the AIP format doesn't persist these internal identifiers and AIP import will generate new internal identifiers. Only handles will be persisted.

This poses a problem if you want to also migrate usage events (Solr statistics) to the new site, because usage events have been tied to DSpace objects via the aforementioned internal identifiers (this is true up to and including DSpace 5). This can be observed as numbers (item_ids) appearing in the "statistics-home" pages instead of item titles.

The procedure described on this page will allow you to export old usage events, convert the old internal identifiers to new identifiers and import the usage events to the new site. The usage events will then match objects in the new site.

This can be used when migrating from Oracle to Postgres because AIP export/import is the only easy way to achieve such migration.

Pre-requisites

- database of the old site is available (handle and bitstream tables)
- database of the new site is available (handle and bitstream tables)
- Solr statistics from the old site are available (can be exported to CSV starting from DSpace 5.3)
- python 2.7
- migrate_solr_statistics.py
- vim (for regex edits to .csv files, but other tools like sed or perl can be used as well)

If you're upgrading from older DSpace than 5.3 which didn't have the `solr-export-statistics` command, you'll need to migrate the Solr statistics core to DSpace 5.3 first and export it to CSV.

Procedure

You will need to prepare the following files: handle-old.csv, handle-new.csv, bitstream-old.csv, bitstream-new.csv, solr-in.csv

handle-old.csv, handle-new.csv, bitstream-old.csv, bitstream-new.csv - these need to be exported from the old and new database, respectively.

Postgres makes this easy:

handle-XXX.csv, bitstream-XXX.csv

```
$ psql
\copy (SELECT handle,resource_type_id,resource_id FROM handle) TO '/tmp/handle-XXX.csv' WITH CSV HEADER;
\copy (SELECT bitstream_id,checksum FROM bitstream WHERE checksum IS NOT NULL) TO '/tmp/bitstream-XXX.csv' WITH
CSV HEADER;
exit
vim /tmp/handle-XXX.txt
:%g/[^0-9]$/d
:wq
```

Oracle sqlplus has no native way of exporting CSVs, so here's a workaround:

handle-XXX.csv

```
sqlplus
set colsep ,
set pagesize 0
set trimspool on
set headsep off
set linesize 300
set numwidth 10
spool on
spool /tmp/handle-XXX.csv
SELECT handle,resource_type_id,resource_id FROM handle;
spool off;
exit

vim /tmp/handle-XXX.csv
# Remove the first few lines and the last few lines manually. Then:
:%s/ *//g
:%g/[^0-9]$/d
:wq
```

bitstream-XXX.csv

```
sqlplus
set colsep ,
set pagesize 0
set trimspool on
set headsep off
set linesize 300
set numwidth 10
spool on
spool /tmp/bitstream-XXX.txt
SELECT bitstream_id,checksum FROM bitstream WHERE checksum IS NOT NULL;
spool off;
exit

vim /tmp/bitstream-XXX.csv
# Remove the first few lines and the last few lines manually. Then:
:%s/ *//g
:%g/[^0-9]$/d
:wq
```

solr-in.csv can be exported from DSpace 5.3 and newer using this procedure:

```
[dspace]/bin/dspace solr-export-statistics
cat [dspace]/solr-export/*.csv > /tmp/solr-in.csv
# remove duplicated headers
vim /tmp/solr-in.csv
yy
:%g/^uid,rpp,userAgent/d
P
:wq
```

To convert the identifiers in solr-in.csv and write the statistics to **solr-out.csv**, run:

```
cd /tmp
curl -s -O -J "https://wiki.duraspace.org/download/attachments/70584988/migrate_solr_statistics.py?
version=1&modificationDate=1443973577575&api=v2"
python migrate_solr_statistics.py handle-old.csv handle-new.csv bitstream-old.csv bitstream-new.csv solr-in.csv
solr-out.csv
```

Please note that this will leave out any usage events pertaining to items, collections or communities that were deleted, so there will likely be fewer lines in solr-out.csv than in solr-in.csv.

Then you can import the statistics to your new site, **replacing all data that is already present** there:

```
rm -rf /dspace/solr-export/*.csv
mv /tmp/solr-out.csv /dspace/solr-export/statistics_export-all.csv
/dspace/bin/dspace solr-import-statistics --clear
```

Conclusion

This procedure is a workaround for a problem that currently doesn't have a solution in DSpace. As DSpace 6 will contain work replacing internal identifiers with UUIDs and Solr statistics and AIP export will have to be changed to accomodate that, there's hope that this procedure will soon be obsolete.