# NGDS FY18 Q2 – Design Test and Validation Report

## Activity Details for Sandia National Laboratory

*April 20, 2018*
*Gary Hudman*
*Arizona Geological Survey*

This quarter has been very active, so we are providing a detailed account of AZGS activity. The NGDS now is current in all harvests, with approximately 93% of the resource links active and working ( +/- 3%). Search, data, and harvest services are fully functional, but the metadata editing and data correction functions still need improvement, with fixes either in progress or planned. Since Forge datasets are of particular interest to the geothermal community, frequent harvest to the GDR site have been occurring, with noticeable recent activity in new Forge datasets. With the flurry of recent harvest activity, recent tests show approximately 9500 duplicate datasets, metadata duplicated in different harvest source. These records have different identifiers, indicating a systemic problem. These corrections are also in process.

Additional staff members were participating this quarter (in addition to Laura and myself) which provided needed expertise and additional resources.

Megha Agarwal – graduate programmer, has been working on python code fixes, Github updates, and documentation.

Dr. Steve Richard - has been working extensively with data, contacting organizations, programming guidance, and planning for system designs improvements.

**Accomplished**

Fixes to the GDR harvest – the GDR moved to the SSL (secured HTTP) protocol, and issues a redirect from non-secure http to the secure address. Our web service request were not able to successfully follow this redirect, causing harvest failure.

Completed re-harvest of all sources. - Seven active harvest sources completed successfully. The USGIN source, an aggregator site that supports all of the AASG data sets (repository.stategeothermaldata.org), is a multi-tier network of geoportals that allows local curation, bulk data correction, and data hosting. We reviewed the process and re-harvested all relevant endpoints.

Schema Fixes in harvest code – the harvest jobs on all data sources generated some faults in the code that processed metadata schema. These were corrected in the code base and reprocessed.

Completed dataset resource validation and inventory - Approximately 60K datasets representing approximately 72K resources were validated for access and data quality. These were performed by script and by manual inspection. A summary report is included at the end of this document.

Working with data host organizations for data corrections - for sustainable use of data, we are collaborating with organizations to identify and correct metadata and data access, within the NGDS & USGIN repository and at the hosting organization.

Github NGDS CKAN codebase updates - the NGDS and USGIN repositories have been successfully updated to current working code base, with references to REI repositories removed. This will allow normal development lifecycle updates and builds.

New Github repository for Data Quality – provides documentation, code, and detailed data reports for the inventory of NGDS resource data. This will be used for ongoing updates on data issues, with detailed reports included in the reports folder:

https://github.com/ngds/data-quality-tools

New NGDS development and debug tools - New development platforms (virtual machines) were built and utilized to improve the debugging process. These will improve the ability to efficiently maintain NGDS servers and data.

AZGS Security Audit – performed in January, including vulnerability testing of NGDS servers. Identified potential areas of improvement and fixes are planned.

**In Process**

System Modification Design Document – This specification will provide a detailed description of proposed changes and improvements, is in process and will be available by the end of the quarter (April 30).

Data correction and correction – will be performed on an ongoing basis. Data quality reports will be updated each quarter.

Data traffic tracking tools for analytics – data traffic tracking processes have been added that log filtered statistics to the database, this process is in test and evaluation stage.

# Data Quality Validation - Summary Statistics

Comprehensive data validation has been in process since April 2, after the current round of harvests. This has been both an automated and manual effort, with results documented in detail in the NGDS data-quality-tools repository.

**Total Resource links – 71495 Working 64450 Errors – 3290 No response - 1235**

| Top 10 Resource providers | |
|---|---|
| https://www.geothermal-library.org/ | 11941 |
| http://www.geothermal-energy.org/ | 8164 |
| http://www.osti.gov/ | 7982 |
| http://ngds.egi.utah.edu/ | 7172 |
| https://gdr.openei.org/ | 5473 |
| http://www.geothermal-library.org/ | 5055 |
| https://pangea.stanford.edu/ | 3754 |
| https://www.sciencebase.gov/ | 1852 |
| http://digitallib.oit.edu/ | 1505 |
| http://geothermal.smu.edu/ | 1372 |

| | |
|---|---|
| http://repository.stategeothermaldata.org/ | 1353 |

**Principle Sources of Errors Identified**

- OSTI.gov – approximately 8000 records were incorrect, due to changes on the OSTI server. Most of these have been corrected in NGDS via script.
- SMU – approximately 800 records return error messages indicating no distribution rights.
- Nevada Geologic Survey – 525 records refer to an ftp site that no longer exists, see below
- Utah – 306 records EGI – files paths have changed
- Iowa – Map server doesn't exist
- Indiana – Map server doesn't exist
- North Carolina – repository no longer accessible - nc-maps.stores.yahoo.net

**Total Map Server Links – 2029 Active – 1895 Error – 106 No Response - 32**

Map Server links or web services that support WFS & WMS were also identified and validated with additional scripts. The table below lists the top map servers and the link counts for each:

| | | |
|---|---|---|
| USGS | https://www.sciencebase.gov/ | 866 |
| Illinois | http://geothermal.isgs.illinois.edu/ | 161 |
| Arizona | http://services.azgs.az.gov/ | 105 |
| Kentucky | http://kgs.uky.edu/ | 104 |
| Nevada | http://web2.nbmg.unr.edu/ | 76 |
| NGDS | http://geothermaldata.org/ | 36 |
| Delaware | http://maps.dgs.udel.edu/ | 21 |
| SMU | http://geothermal.smu.edu:9000/ | 20 |
| Washington | http://ec2-50-18-49-187.us-west-1.compute.amazonaws.com/ | 15 |
| Oregon | http://www.oregongeology.org/ | 13 |
| New Hampshire | http://www4.des.state.nh.us/ | 13 |

Map Servers with errors, and error counts are noted below:

| Agency | Map Server Link Error URL | Error Count |
|---|---|---|
| Illinois | http://geothermal.isgs.illinois.edu/ | 71 |
| USGIN | http://data.usgin.org/ | 2 |
| Washington | http://ec2-50-18-49-187.us-west-1.compute.amazonaws.com/ | 29 |
| Kentucky | http://services.kgs.ku.edu/ | 2 |
| Kentucky | http://kgs.uky.edu/ | 2 |
| SMU | http://geothermal.smu.edu/ | 2 |
| Minnesota | http://mgsweb2.mngs.umn.edu/ | 10 |

| | | |
|---|---|---|
| Arizona | http://services.azgs.az.gov/ | 10 |
| Oregon | http://www.oregongeology.org/ | 1 |
| Wisconsin | http://gis.wgnhs.org/ | 8 |
| USGS | https://www.sciencebase.gov/ | 1 |

**Redundant Datasets**

The SMU and USGIN harvest data sources have many of the same records, with different identifiers, which allows them to be all loaded during harvests. Correcting this issue will significantly reduce the number of data link errors noted above.

**Activities Coordinated with Other Agencies**

GDR – Working with Jon Weers for harvest connectivity and metadata concerns. No current issues. Forge datasets have been noted in recent harvests.

Indiana – The Indiana map server is no longer operating, we are waiting response from Gary Motz to determine the best way to proceed with these records.

Iowa – The Iowa map server is no longer operating, we are waiting response from Mary Howes to determine the best way to proceed with these records.

USGS – Science Base – We identified metadata issue and worked with Drew Ignizio, corrected our harvest code.

OSTI – The OSTI modified their services, rendering URL's invalid. We corrected approximately 6400 resource records by script, and notified OSTI librarian Sara Studwell, who indicated that they will also be resolving on their servers.

Nevada Geological Survey – ongoing activity with Bridget Ayling and Jennifer Mauldin. Nevada had moved to a new data server, deprecating the ftp used during the NGDS build. In NGDS, there are 525 resource links still directed at the ftp site. Jennifer had updated metadata records on the AASG site, which was successfully re-harvested into NGDS. However erroneous links remain in AASG and NGDS, we are in the process of isolating the source of error and correcting.

Southern Methodist University – We have been working with Cather Chickering and Maria Richards on several issues. Cathy had noted errors on SMU and other datasets in NGDS, which have been corrected. Reviewing for similar errors, we found a number and have corrected. In addition, as part of the data validation suite, a significant number of errors (approximately 800) are caused by links that on the SMU site that respond with:

"SMU does not have the right to distribute the document you seek."

We are waiting confirmation of a proposed fix. Our proposed updates for those is to remove the online linkage all together with a description in the metadata that says 'not available online'.