

Award No. G15AP00102

National Geological and Geophysical Data Preservation Program

**Washington State Metadata Project:
Paper-to-Digital Conversion, Metadata, and Database Development for
Subsurface Data in Washington State**

Final Technical Report

Robert Berwick, Daniel Eungard, Ashley Cabibbo

Washington Department of Natural Resources

Division of Geology and Earth Resources

MS 47007

Olympia, WA 98504-7007

(360) 902-1431 (360) 902-1463 (360) 902-1453

(360) 902-1785 fax

robert.berwick@dnr.wa.gov

daniel.eungard@dnr.wa.gov

ashley.cabibbo@dnr.wa.gov

Award Term: July 1st, 2015 to June 30th, 2016

Submitted: June 28, 2016

Abstract

The Washington State Department of Natural Resources, Division of Geology and Earth Resources, houses many geologic and geophysical data collections, 15 of which have been inventoried in the National Digital Catalog. For our 2015 project we focused on digitizing geotechnical boring reports. Our 2015 project objective was a continuation of the work started on our geotechnical report collection in 2009, and continued in years 2010 and 2014.

This year we produced 4,855 metadata records from geotechnical reports we digitized using NGGDPP funding. This information was compiled into XML format for delivery in accordance with the National Digital Catalog specifications.

Table of Contents

Abstract.....	1
Introduction	1
Project Goals.....	2
Methodology.....	2
Results.....	4
Conclusion.....	5

Introduction

The Washington State Department of Natural Resources, Division of Geology and Earth Resources (DGER), has identified 15 collections of geologic and geophysical data, which have been inventoried in the National Digital Catalog.

In 2008, David K. Norman, Washington State Geologist and Division Manager for DGER, produced the first draft of the *Geological Data Preservation Plan of Washington*, which states that preserving subsurface records is a priority of the Division of Geology and Earth Resources. Subsequent updates to the data preservation plan retain this emphasis on preserving all available subsurface data, ensuring that it is accessible through our Geological Information Portal and the National Digital Catalog.

With these goals in mind, geotechnical reports that included subsurface data were chosen for fiscal year 2015 because they were most at risk of loss or degradation and held high value geological and geophysical data.

For year 2015 of the National Geological and Geophysical Data Preservation Program (NGGDPP), we produced and submitted metadata records for 4,855 boreholes found in 1,420 reports.

Geotechnical reports are an indispensable source of subsurface geological and geophysical information, and are produced at considerable expense. Unfortunately, geotechnical reports are generally not made available for long-term use; most local governments commonly dispose of these reports at the conclusion of the project for which the borings were drilled, and geotechnical firms typically keep their records of geotechnical investigations private. Countless geotechnical reports submitted to local governments have already been lost. For several years DGER has made an effort to collect geotechnical reports for borings drilled in Washington to preserve their long-term value. DGER has collected these reports over the years as supporting information for mapping and hazards studies, but has stepped up efforts to collect as many reports as possible for preservation and access by contacting local jurisdictions and engineering firms and requesting copies of any reports they have.

Project Goals

Our focus for this project was to continue the work begun in previous years, creating sample-level metadata for entry into the National Digital Catalog and making this information readily accessible to the public through hyperlinks to scanned documents and through an interactive mapping application.

The project covers the grant period beginning July 1st, 2015, and ending June 30th, 2016. It was anticipated that scanning the reports, compiling the data and digitizing the points would take up the majority of the time. The remaining time would be focused on QAQC of the resulting data set, adding the data to our existing subsurface database (which makes it available over our Geological Information Portal), and creating XML format metadata in accordance with the National Digital Catalog specifications.

Methodology

Geotechnical reports were scanned using an Epson Workforce Pro GT-S50 document scanner at a resolution of 300 DPI, which provides an adequate balance of image quality and file size for the scanned report. The scanned documents were saved in PDF format, and the pages that had maps were exported as TIF images. The PDF documents were then search enabled with Optical Character Recognition (OCR) using Adobe Acrobat Pro.

The TIF map images were imported into ArcMap 10.3.1, and georeferenced using a combination of parcel boundaries and aerial imagery to establish control points. From this image, points representing borehole locations were digitized into a surrogate database that mirrors the schema of Washington's Subsurface Geology Information System database.

The point feature class and document table were attributed using the most suitable method for the individual attribute. A primary key was established for the feature class, and foreign key for the document table. The primary and foreign keys are used by a relationship class to link the two datasets together.

A majority of the data was taken directly from the geotechnical reports and entered as attributes in the appropriate tables. Qualitative assessments were made for both data quality and location confidence, and values assigned for each record. Hyperlinks were created to provide a direct link to the scanned report.

Spatial attributes were calculated, derived and assigned. The latitude and longitude attributes were calculated in ArcMap using the calculate geometry tool (calculate x, calculate y). Elevations were taken directly from the report if available. When elevations were not available directly from the report, they were calculated from digital elevation models (DEM) or Light Detection and Ranging (LiDAR) when available. County attributes were derived from a layer containing county boundaries using a spatial join and field calculating the county name from the polygon to the points that were contained within the county boundary.

Quality assurance was provided by utilizing domained fields where appropriate. This limited the data to acceptable values, and provided pick lists for data entry that assured the standardization of the data as it was entered. Basic quality control was performed by examining the data to see if the data was appropriately located spatially, primary keys were unique, boreholes were linked to documents and documents were linked to boreholes. The final quality control was implemented when importing data into the XML generator template and then programmatically during the process of generating the XML metadata for the National Digital Catalog.

The tables from the proxy database that contained data needed to generate metadata for the National Digital Catalog were exported from ArcGIS. Fields from these tables were mapped to corresponding fields in the XML template. Before each field was imported into the template the data was scrutinized once more to ensure all values were expected values and that they were properly formatted.

Humanly readable, grammatically correct abstracts were generated for each record using a series string concatenations that took into account subject-verb agreement and the use of proper articles when constructing each sentence.

When the XML template was properly filled out, the XML generating script was initialized. During this procedure, the script builds the XML metadata document while running a final quality control check on the data cell-by-cell using an array of conditional statements designed to detect data issues. If an issue is detected the record is sent to an error sheet, tagged with a meaningful description of the issue and the cell address in which the problem was found so that the issue may be corrected.

Results

On June 28, 2016, we submitted NGGDPP compliant metadata for 4,855 boreholes found in the 1,420 reports we processed this year.

Estimated:	Estimated number of boreholes contained in geotechnical reports	4,000
Results:	Number of boreholes with complete metadata submitted to the National Digital Catalog	4,855

The number of metadata records is higher than we originally estimated. This was in part due to acquiring more borings locations from the geotechnical reports than we had originally estimated, and because we were able included a group of reports from one source that were scanned pro bono. This provided a time/cost savings and contributed to our ability to produce more data for this year's deliverable.

Data from this project has been added to the subsurface database and is displayed on the Washington State Geologic Information Portal, an interactive mapping site designed to address one of the core requirements of the Geologic Data Preservation Plan of Washington, which is to make data available and accessible.

At the Geologic Information Portal, a user may navigate through different themes; locate, overlay, and query data; create maps with titles, legends, and labels; create their own points and polygons; and export their maps as PDFs or georeferenced TIFFs.

The Subsurface Geology Information System is the map theme where the data from the geotechnical reports are now housed. The theme can be selected from the Map Theme drop-

down of the application. When a point is selected with the Point Identify tool, a pop-up gives the user useful information related to that point in four categories:

1. Point Data
2. Layer Data
3. Comment Data
4. Document Information

A hyperlink to a PDF version of the scanned document is provided in the Document Information section of the Identify pop-up. Subsurface data submitted to the NGGDPP is made available publically through the Geologic Information Portal at:

<https://fortress.wa.gov/dnr/protectiongis/geology/?Theme=subsurf>

Conclusion

Geotechnical reports continue to be an indispensable source of subsurface geological and geophysical information. The information they contain is used by geologists, engineers, emergency managers, planners, private-sector geoscientists, and the general public.

Throughout our partnership with the NGGDPP, we have rescued, preserved and scanned thousands of geotechnical reports, documenting well over 100,000 borehole locations. After a successful year, we have added another 4,855 boreholes to this total. We are making this data available to the thousands of visitors to the Washington Geologic Information Portal and the National Digital Catalog each month.

We hope to partner again with the NGGDPP to digitally preserve more documents and the considerable financial expenditure that they represent.