

**CREATION OF DIGITAL METADATA
FOR A WELL-RELATED DATA COLLECTION
OF THE OKLAHOMA GEOLOGICAL SURVEY**

GRANT NO. DOI – USGS 09HQPA0009

FINAL TECHNICAL REPORT

Co-Principal Investigators:

G. Randy Keller, OGS

Email: grkeller@ou.edu

405-325-7968

Jane L. Weber, OGS

Email: jlweber@ou.edu

405-325-7331

Oklahoma Geological Survey, University of Oklahoma
100 E. Boyd Street, Room N-131
Norman, Oklahoma 73019-0628
Phone: 405-325-3031
Fax: 405-325-7069

Grant Initiation Date: July 1, 2009

Grant Completion Date: June 30, 2010

Submittal Date of Final Technical Report: June 21, 2010

TABLE OF CONTENTS

TABLE OF CONTENTS	2
ABSTRACT	3
INTRODUCTION	4
APPROACH TO WORK	5
Preparation of Data	5
Mapping of Data to Metadata Elements.....	5
Uploading of Metadata to National Digital Catalog	7
Examples of Metadata Records.....	7
ACCOMPLISHMENTS RELATIVE TO PROJECT GOAL AND LONG-TERM GOALS	9

ABSTRACT

The Oklahoma Geological Survey (OGS) created a file of digital sample-specific metadata for 8910 individual rock core samples from a collection that previously had been identified in the on-line survey component of the National Digital Catalog. The rock cores were cataloged, identified, and described in an Excel file.

Metadata were created in an Excel spreadsheet template that contained all 7 elements required by the National Geological and Geophysical Data Preservation Program (NGGDPP) plus 3 elements listed as optional. The metadata file was populated by mapping OGS rock core properties to one of the elements in the metadata spreadsheet using Copy and Paste functions. The resulting metadata were then extracted to a Comma Separated Value (CSV) file format using the pipe character (|) as a record delimiter.

The CSV file was uploaded through the NGGDPP web interface at <http://my.usgs.gov/csc/nggdpp/upload> , successfully validated on the first try, and loaded into the National Catalog.

The process used in this project can serve as a model for preparing other sample-specific metadata, particularly metadata for a data collection related to drill holes.

CREATION OF DIGITAL METADATA FOR A WELL-RELATED DATA COLLECTION OF THE OKLAHOMA GEOLOGICAL SURVEY

INTRODUCTION

The Oklahoma Geological Survey (OGS) maintains a repository of diverse geological and geophysical data related to the State's non-biological natural resources and has a responsibility to disseminate information about those data to the public. In FY2007, as part of the National Geological and Geophysical Data Preservation Program (NGGDPP), OGS identified 26 data collections to enter into the on-line survey component of the National Digital Catalog (NDC). Only a few of those collections have been individually cataloged. Of the remaining collections, some are sorted to enable manual searches while others have yet to be organized into a useable form. Cataloging and computerizing the estimated several million items in the collections for the benefit of users is a long-term goal of OGS.

To fulfill requirements of its FY2009 NGGDPP grant, OGS developed sample-specific metadata for one of its datasets that previously had been listed in the on-line survey, a collection of about 9000 rock core samples. Metadata were created in an Excel spreadsheet and extracted to a Comma Separated Value (CSV) file format using the pipe character (|) as a record delimiter. Included in the file were the 7 required metadata elements plus 3 optional elements, all listed as acceptable by the National Catalog. The resulting metadata file was successfully uploaded to the National Catalog via the NGGDPP web interface at <http://my.usgs.gov/csc/nggdpp/upload>.

APPROACH TO WORK

Preparation of data: The OGS rock core database had been assembled over a long period of time in an Excel spreadsheet, with data entry accomplished by many different employees working under different circumstances. Consequently, the first step in this project was to review the data, looking for obvious data or typographic errors, inconsistencies, and data omissions.

The clean-up process involved:

- Cross-checking County Name vs. API Number to make sure the county code portion of the API Number corresponded to the County Name listed.
- Cross-checking Township-Range designations vs. County Name to make sure the listed Township could be in the listed County. The presence of two meridians in Oklahoma can lead to confusion for 1-6N, 1-27E locations if the appropriate meridian is not specified.
- Filtering depth data for negative or zero values.
- Adding missing API Numbers.
- Adding missing Latitude and Longitude values.

Sources of information consulted in resolving data conflicts or data omissions were IHS Energy data; Natural Resources Information Systems (NRIS) data from Oil-Law; and, for spatial coordinate determinations, the Spatial Calculator at the University of Oklahoma's Center for Spatial Analysis. Coordinates thus determined were calculated using the NAD 83 datum standard and Topographic Mapping Company's land grid. It is recognized that in many cases latitude and longitude values already in the database were obtained using a different datum and/or land grid. The resulting database contained 18 fields of information for 9266 Oklahoma rock core samples representing 4964 wells.

Mapping of data to metadata elements: For this project, we chose to provide a CSV file of metadata. To preclude possible problems associated with commas used in descriptive fields, the default comma record delimiter (List Separator) on the work computer was replaced with the pipe character (|) before work was begun.

The NGGDPP program named 7 required metadata elements and 6 optional elements. After studying the examples described in the 5-15-09 email "Preparing Metadata for the National Digital Catalog" as well as Alaska's example sent to us, we were able to decide which of our fields of sample properties to use in populating the required elements plus 3 additional optional elements in the CSV metadata template. Our intent was to offer as much

information about the resource as possible, given available time and personnel to devote to the task. The match-ups we made were:

<u>NDC Metadata Element</u>	<u>OGS Database Fields Used</u>
<i>collectionID</i> (req.)	(Assigned by NCGDPP. Same for all records.)
<i>title</i> (req.)	Lease, Well Number
<i>abstract</i> (req.)	Operator, County, Formation, Top, Bottom, Sample Description
<i>dataType</i> (req.)	(Controlled by NCGDPP. Same for all records.)
<i>supplementalInformation</i> (req.)	Not derived from dataset. Same for all records.
<i>coordinates</i> (req.)	Latitude, Longitude
<i>datasetReferenceDate</i> (req.)	Month of data review. Same for all records.
<i>alternateTitle</i> (opt.)	File Number, API Number
<i>alternateGeometry</i> (opt.)	Township Number and Direction, Range Number and Direction, Section, Quarter Section Name of meridian included to prevent confusion between samples east of Indian Meridian and those east of Cimarron Meridian
<i>verticalExtent</i> (opt.)	Top, Bottom Also included under <i>abstract</i> to enhance description of resource

There was no one-to-one mapping between OGS data properties and permissible metadata elements. To combine more than one column of sample data into one metadata element, a formula with script was written for the first sample record and copied to all subsequent records. The resulting column of metadata was then copied but only its “values” were pasted into the Excel metadata template file.

Certain OGS sample records were excluded during the mapping process. If location information did not include at least a Quarter Section specification, the sample was omitted from the mapping process due to the uncertainty in the location. A couple of hundred samples were excluded on this basis. If depth values were missing from the dataset, those samples were also excluded, since depth is a critical piece of information for a rock core sample. More than 100 samples were excluded for this reason. The majority of samples omitted were of whole core. In a number of these cases, the same well is represented by other records in the metadata file --- records for another sample type (chips, slab, etc.), another depth, or another operator. It is not uncommon to encounter “skips” in depth for core material. Hence, there can be one record listing whole core from a depth of 2439 -2512 feet (including “skips”) plus another record from the same well listing chips from 2439-2448 feet in depth.

A test file of the first 30 samples completed was submitted to Rick Brown of the USGS Central Region Geospatial Information Office to ascertain that our procedure and resulting metadata were acceptable. They were acceptable.

Uploading of metadata to National Digital Catalog: The first two records of the completed metadata template file are shown here, divided at the end of the line for display purposes. The first record contains the names of the metadata elements, required and optional, corresponding to the data presented in the second record.

collectionID	title	abstract	alternateTitle	verticalExtent	coordinates
1089230	Core sample from well: PROTHRO 1-C	This sample is Type: Whole Core Top: 3598 ft. Bottom: 5200 ft. Formation: Morrowan Operator: Cities Service State: Oklahoma County: Cimarron	File Number: 3129 API Number: 35025350770000	ft, 5200,3598	-102.38444, 36.53949

alternateGeometry	supplementalInformation	datasetReferenceDate	dataType
Public Land Survey System Township: 1 North Range: 6 East of the Cimarron Meridian Section: 23 Quarter-quarter: SENW	Contact the manager of the Oklahoma Geological Survey's Oklahoma Petroleum Information Center at 405-325-3031 to access the core. Pulling/restocking fees for in-house or out-of-house viewing apply. Additional information is available at http://www.ogs.ou.edu/OPIC.php .	2009-02	Rock Core

After all OGS samples were mapped to the Excel metadata template, the metadata file was saved as a CSV file, also known as an ASCII or flat file. The first two records of the CSV file follow. As described above, the first record contains the names of the metadata elements corresponding to the data presented in subsequent lines. Note that the pipe symbol (|) is being used as a delimiter to separate metadata elements. Bolded items were not bolded in the uploaded file but are bolded here to illustrate which information was derived from our database.

collectionID|title|abstract|alternateTitle|verticalExtent|coordinates|alternateGeometry|supplementalInformation|datasetReferenceDate|dataType
 1089230|Core sample from well: **PROTHRO 1-C** |This sample is Type: **Whole Core** Top: **3598** ft. Bottom: **5200** ft. Formation: **Morrowan** Operator: **Cities Service** State: Oklahoma County: **Cimarron** |File Number: **3129** API Number: **35025350770000**|ft, **5200,3598**|-**102.38444**, **36.53949**|Public Land Survey System Township: **1 North** Range: **6 East** of the Cimarron Meridian Section: **23** Quarter-quarter: **SE****NW** |Contact the manager of the Oklahoma Geological Survey's Oklahoma Petroleum Information Center at 405-325-3031 to access the core. Pulling/restocking fees for in-house or out-of-house viewing apply. Additional information is available at <http://www.ogs.ou.edu/OPIC.php>.|2009-02|Rock Core

We acquired a login account from myusgs@usgs.gov and uploaded a file of 8910 site-specific metadata records for Oklahoma rock core samples through the interface at <http://my.usgs.gov/csc/nggdpp/upload>. There were no web-based validation errors but Rick Brown noted 98 errors involving special non-ASCII characters (line-feed and carriage-return) at the end of coordinate strings that were not reported in the web-based validation process. Inspection of our data suggested these characters were likely introduced when we combined datasets containing differently formatted coordinate data, left-justified "text"

coordinates with right-justified “number” coordinates, and then combined the latitude and longitude coordinate values into the metadata element format accepted by the NDC. Mr. Brown chose to update the NDC loader code to make it more robust for this type of problem; we did not have to modify our data or our file. We verified the presence of our data at <http://myusgs.gov/csc/nggdpp/state/ok>. Several weeks later we noted a discrepancy between the date information we submitted for the datasetReferenceDate element and that which appeared in the NDC. Mr. Brown determined this problem had been introduced when he was seeking a remedy to fix a date problem with the loader and immediately updated our data in the Catalog to show the correct date.

ACCOMPLISHMENTS RELATIVE TO PROJECT GOAL AND LONG-TERM GOALS

The primary goal of this project, to provide *sample-specific* metadata to the National Digital Catalog for a major geological sample collection maintained by OGS, was accomplished. Starting with an in-house catalog of 9266 rock core samples representing 4964 wells, OGS developed and successfully submitted metadata to the Catalog for 8910 of those rock cores. Missing or incomplete data prevented some samples from being included.

A secondary but highly desirable goal was to develop a set of procedures that could serve as a framework or model when preparing sample-specific metadata, particularly metadata for a data collection related to drill holes. Achieving that goal would be important to long-term data preservation plans of OGS. Prior to this project our long-term goals for data preservation did not include the preparation of sample-specific metadata. Completing this project has introduced OGS to this aspect of data preservation. As we had hoped in proposing this particular project, with little modification, the process and techniques employed in this initial effort can be applied to developing metadata for the National Digital Catalog of other datasets related to drill holes maintained by OGS. These collections include: mud logs, strip logs, electric logs, drill cuttings, well completions, and core analyses among others. The knowledge and experience gained in fulfilling this grant, especially discovering the amount of time required for the tasks involved, will be important factors to consider when allocating scarce OGS resources in the future.

OGS recognizes the need for and the value of metadata. The major stumbling block for OGS continues to be the lack of resources necessary to catalog individual items within a collection. Without that type of itemization, *sample-specific* metadata cannot be created for the collection. Meanwhile, available resources must continue to be devoted to other data preservation efforts, such as filing and storing material, preparing material for public access, improving physical examination areas and data handling procedures, building or adding to digital databases, reviewing legacy databases, and pursuing leveraging opportunities for data preservation with external entities.