Web Service and Web Accessible Folder access to the NSF EarthCube Paleogeoscience Research Coordination Network "Catalogs of Resources", a primer

Author: Wendy S. Gross, NOAA's National Centers for Environmental Information, World Data Center for Paleoclimatology
Contact: wendy.gross@noaa.gov
Date last modified: January 15, 2015
Online resource: ftp://ftp.ncdc.noaa.gov/pub/data/paleo/webservices/doc/ogc-cs-webservice-primer-ec-cp4-rcn-catalogs.pdf

I. Introduction

Metadata in catalogues represent resource characteristics that can be queried and presented for evaluation and further processing by both humans and software. The EarthCube Paleogeoscience (CP4) provides three catalogues: Software, Repositories, and Databases (CP4 Catalogues). NOAA's National Centers for Environmental Information (NCEI), World Data Center for Paleoclimatology (WDC Paleo) provides for the harvesting of these catalogues programmatically in two ways:

(1) A web service, specifically an instance of the Open Geospatial Consortium (OGC) Catalogue Service, and

(2) Web accessible folders (WAFs), which make accessible a Dublin Core (dc) Metadata record, and NASA Directory Interchange Format (DIF) Metadata record for every dataset/record of the CP4 Cataloging project. (Note: WDC Paleo has collaborators that are using the Unix command "wget" to programmatically access metadata records from these web accessible folders).

The remainder of this document contains information about both of these harvesting methodologies.

II. Open Geospatial Consortium Catalogue Service for the Web (CSW)

The Open Geospatial Consortium (OGC), Catalogue Service for the Web [a.k.a. Catalogue Service] (CSW) is provided through NOAA/NCEI's instance/installation of the ESRI Geoportal Toolkit. The Catalogue Service supports the ability to publish and search collections of descriptive information (metadata) for data, services, and related information objects.

Following is an introduction and examples of how a client would use operations provided by a CSW Server, in this case, the ESRI GeoPortal implementation, to:

- Retrieve metadata about operations available
- Search and retrieve collections of metadata

from the CP4 Catalogues. Note that all three CP4 Catalogues are actually stored in one CSW catalogue and, as will be described, are differentiated into types of Software, Repositories, and Databases by tags of each catalogue record's XML metadata record.

This document does not discuss all operations and capabilities of the CSW. For full documentation on the CSW standard, visit: http://www.opengeospatial.org/standards/cat

A. GetCapabilities Operation

The getCapabilities operation can be thought of as the Rosetta Stone. It allows CSW clients to retrieve service metadata describing a Catalogue Service instance. The response to a GetCapabilities request is an XML document containing information about the servers capabilities. For example, the XML document returned by the GetCapabilities operations includes the following information: metadata about the server, metadata about operations provided by this server (e.g., the URL to GetRecords operation), the query language supported. For more information on the Get Capabilities operation see the OpenGIS Catalogue Services Specification version 2.0.2 (http://www.opengeospatial.org/standards/cat)

The HTTP Get operation for GetCapabilities for the CP4 Catalog is:

http://gis.ncdc.noaa.gov/gptpaleo/csw?service=CSW&version=2.0.2&request=GetCapabilities

B. GetRecords Operation – Querying by subject of the record

GetRecords performs the resource discovery operations (searching and presentation or records) for the CSW. The Query element is used for all searching. For more information on the GetRecords operation, see the OpenGIS Catalogue Services Specification version 2.0.2. Examples of this request follow. For each of the examples, click on the link or cut and past the URL into your browsers URL bar.

Examples 1, 2, and 3: Obtain the first fifty records for Software, Repository, and Database CP4 Catalog resources for Software, Repositories, and Database, respectively. The following links result in the first 50 records for the given resource type:

Example 1 – Obtain first fifty records of Software resources:

http://gis.ncdc.noaa.gov/gptpaleo/csw?service=CSW&version=2.0.2&request=GetRecords&typeNames=csw:Record&resultType=results&startPostion=1&maxRecords=50&CONSTRAINTLANGUAGE=Filter&Constraint=<Filter><PropertyIsEqualTo><PropertyName>subject</PropertyName><Literal>earth science/paleoclimate/software</Literal></PropertyIsEqualTo></Filter>&constraint_language_version=1.1.0

Example 2 – Obtain first fifty records of Repository resources:

http://gis.ncdc.noaa.gov/gptpaleo/csw?service=CSW&version=2.0.2&request=GetRecords&typeNames=csw:Record&resultType=results&startPostion=1&maxRecords=50&CONSTRAINTLANGUAGE=Filter&Constraint=<Filter><PropertyIsEqualTo><PropertyName>subject</PropertyName><Literal>earth science/paleoclimate/repository</Literal></PropertyIsEqualTo></Filter>&constraint_language_version=1.1.0

Example 3 – Obtain first fifty records of Database resources:

http://gis.ncdc.noaa.gov/gptpaleo/csw?service=CSW&version=2.0.2&request=GetRecords&typeNames=csw:Record&resultType=results&startPostion=1&maxRecords=50&CONSTRAINTLANGUAGE=Filter&Constraint=<Filter><PropertyIsEqualTo><PropertyName>subject</PropertyName><Literal>earth science/paleoclimate/database</Literal></PropertyIsEqualTo></Filter>&constraint_language_version=1.1.0

As an explanation of the previous three examples, following is the dissection of the third. (Note: for a nice description of all GetRecords parameters, visit http://reference.mapinfo.com/software/mapinfo_developer/english/1_0/csw/csw/postget/postgetgetrecords.html):

"http://" – defines the Internet Protocol as Hypertext Transfer Protocol
"gis.ncdc.noaa.gov/" - Internet Domain of the CSW service
"gptpaleo/csw?" - path of to the CSW service followed by the question mark as the path terminator

The remaining text of the URL is a series of parameters which are passed to the CSW service, each separated with an ampersand character , '&':
"service=CSW" – defines this to be a call to the CSW service
"request=GetRecords" – defines the operation to be performed to be GetRecords
"typeNames=csw:Record" – XML schema type of csw:Record, where csw is the namespace
"resultType=result" – Specifies the detail type of the response. The resultType determines whether the service returns a summary of the result set (hits), one or more records from the result set (results), or validates the request message and processes it asynchronously (validates).
"startPostion=1" - Specifies which record position the catalogue should start generating the output. Default value is 1, or the first recond in the result set.
"maxRecords=50" - Specifies the maximum number of records returned in the result set of a query. The default value is 10.
"CONSTRAINTLANGUAGE=Filter"
"constraint_language_version=1.1.0" – the CONSTRAINT LANGUAGE, and constraint_language_version allow catalogue clients to specify the predicate (logical) language to be used to constrain operations, thus constraining the set of records returned. These two parameter will be discussed as examples in the next section about Predicate Languages.

C. Predicate Constraint Language

The implementation of CSW used for CP4 Catalogs allows catalogue clients to constrain the results returned. It does so by using the FILTER predicate language.

FILTER Is an XML encoding format, and all CSW implementations are required to support this filter syntax. For details visit:
http://www.opengeospatial.org/standards/filter for the complete reference and,

http://webhelp.esri.com/geoportal_extension/9.3.1/index.htm#geoportal_csw_cmpnts.htm Table 3 - Core Queryable Description and Mapping, for ESRI specific implementation features.

Now going back once again to ***Example 3*** – Obtain first fifty records of Database resources:

http://gis.ncdc.noaa.gov/gptpaleo/csw?service=CSW&version=2.0.2&request=GetRecords&typeNames=csw:Record&resultType=results&startPostion=1&maxRecords=50&CONSTRAINTLANGUAGE=Filter&Constraint=<Filter><PropertyIsEqualTo><PropertyName>subject</PropertyName><Literal>earth science/paleoclimate/database</Literal></PropertyIsEqualTo></Filter>&constraint_language_version=1.1.0

"CONSTRAINTLANGUAGE=Filter" – specifies that the constraint language that will be used is **"Filter"**

Now to dissect:

"Constraint=<Filter><PropertyIsEqualTo><PropertyName>subject</PropertyName><Literal>earth science/paleoclimate/database</Literal></PropertyIsEqualTo></Filter>"

Here, what is being asked for is all of the catalog records where tag "<subject>" is equal to "earth science/paleoclimate/database".

Lastly, the parameter **"constraint_language_version"** is set to 1.1.0

D. GetRecords Operation – Using Logical "Or" operator

Logical operators may be used to combine spatial operators and comparison operators in one filter expression, as shown in Example 4, below. Encoding of Logical Operators is defined in section 7. 10.2 of the OGC Filter Encoding 2.0 Encoding Standard (http://docs.opengeospatial.org/is/09-026r2/09-026r2.html) .

Example 4 – Obtain records that are "software" OR "database":

http://gis.ncdc.noaa.gov/gptpaleo/csw?service=CSW&version=2.0.2&request=GetRecords&typeNames=csw:Record&resultType=results&startPostion=1&maxRecords=50&CONSTRAINTLANGUAGE=Filter&Constraint=<Filter><Or><PropertyIsEqualTo><PropertyName>subject</PropertyName><Literal>earth science/paleoclimate/software</Literal></PropertyIsEqualTo><PropertyIsEqualTo><PropertyName>subject</PropertyName><Literal>earth science/paleoclimate/database</Literal></PropertyIsEqualTo></Or></Filter>&constraint_language_version=1.1.0

E. GetRecords Operation – Querying by Modification Date

Example 5 – Obtain software records that have been modified since the beginning of 2013.

http://gis.ncdc.noaa.gov/gptpaleo/csw?service=CSW&version=2.0.2&request=GetRecords&typeNames=csw:Record&resultType=results&startPostion=1&maxRecords=50&CONSTRAINTLANGUAGE=Filter&Constraint=<Filter><And><PropertyIsEqualTo><PropertyName>subject</PropertyName><Literal>earth science/paleoclimate/software</Literal></PropertyIsEqualTo><PropertyIsGreaterThan><PropertyName>Modified</PropertyName><Literal>2013-01-01</Literal></PropertyIsGreaterThan></And></Filter>&constraint_language_version=1.1.0

Returns: ... numberOfRecordsMatched="247" ...
There have been 247 Software records Modified since the beginning of 2013

F. GetRecords Operation – Combining Logical Operators

Example 6 – Obtain software, repository, or database records that have been modified since the beginning of 2013

http://gis.ncdc.noaa.gov/gptpaleo/csw?service=CSW&version=2.0.2&request=GetRecords&typeNames=csw:Record&resultType=results&startPostion=1&maxRecords=50&CONSTRAINTLANGUAGE=Filter&Constraint=<Filter><And><Or><Or><PropertyIsEqualTo><PropertyName>subject</PropertyName><Literal>earth science/paleoclimate/software</Literal></PropertyIsEqualTo><PropertyIsEqualTo><PropertyName>subject</PropertyName><Literal>earth science/paleoclimate/repository</Literal></PropertyIsEqualTo></Or><PropertyIsEqualTo><PropertyName>subject</PropertyName><Literal>earth science/paleoclimate/database</Literal></PropertyIsEqualTo></Or><PropertyIsGreaterThan><PropertyName>Modified</PropertyName><Literal>2014-01-01</Literal></PropertyIsGreaterThan></And></Filter>&constraint_language_version=1.1.0

Returns: ... numberOfRecordsMatched="665" ...
There have been 247 Software records Modified since the beginning of 2013

G. GetRecordById Operation

This operation is a subset of the GetRecords operation, and is included as a convenient short form for retrieving and linking to records in a catalogue. The key parameter is Id, which is a comma-separated list of anyURI, allowing you to request more than one record from the catalog.

http://gis.ncdc.noaa.gov/gptpaleo/csw?service=CSW&version=2.0.2&request=GetRecordById&Id=http://www.manfredmudelsee.com/soft/2samples/index.htm

III. Web Accessible Folders, for accessing CP4 Catalog records

A Web Accessible Folder (WAF) is a simple *directory* of files on a *web.* A Dublin Core (dc) Metadata record which exists for every cataloged dataset of the CP4 project, is used as the source for the information searched and returned by the CP4 Catalog CSW service. All of the Dublin Core metadata records, accessible via this CSW Service, are also available via a web accessible folder (WAF).

The web accessible folder is available from the following two synonymous URLS:
http://www1.ncdc.noaa.gov/pub/data/metadata/published/paleo/dc/xml
ftp://ftp.ncdc.noaa.gov/pub/data/metadata/published/paleo/dc/xml

The specification of each filename in this folder includes:
(a.) the CP4 Catalog resource type (software, repository, other [for other database/collection], or Paleo proxy [coral, tree ring, pollen, etc.]) and
(b.) a unique numeric dataset identifier.
For example the file:
http://www1.ncdc.noaa.gov/pub/data/metadata/published/paleo/dc/xml/noaa-repository-1002694.xml
is the metadata record for the: Bangor University Marine Bivalve Shell Repository.
It is made up as follows:
"noaa-", followed by the resource type ("repository")
followed by the "-" (dash character), followed by the numeric
unique dataset id ("1002694"), followed by ".xml"

Note 1: The WAF filename can be used to harvest all Software and Repository XML metadata records. However, the WAF filename alone cannot be used to harvest Database XML metadata records, the <dc:subject> tag of each record must also be used. Details follow.

The content of all Software and Repository records <dc:subject> tag is:
earth science/paleoclimate/software and earth science/paleoclimate/repository
respectively, and the word “software” or “repository” is included in the WAF filename.

Similarly, the content of all Database records <dc_subject> tag is:
earth science/paleoclimate/database
HOWEVER, unlike Software and Repository records, for database records the string "other" is included in the WAF filename, NOT “database”. Therefore, the filename alone cannot be used to harvest the "database" metadata catalogue records, the <dc:subject> tag must also be used.

Note 2: NOAA/WDC/PALEO also provides a web accessible folder that contains a DIF metadata record for every CP4 Catalog metadata record, which is named analogous to the specification for Dublin Core (above). the URLs are:
http://www1.ncdc.noaa.gov/pub/data/metadata/published/paleo/dif/xml
ftp://ftp.ncdc.noaa.gov/pub/data/metadata/published/paleo/dif/xml

These DIF records are NOT used by the CP4 Catalog CSW.

References:

OGC Network, Easy Catalogue Services for the Web http://www.ogcnetwork.net/node/630

OGC Catalogue Service, Overview and Downloads of Catalogue Service Specification documentation http://www.opengeospatial.org/standards/cat

MapInfo, CSW Service Reference, http://reference.mapinfo.com/software/mapinfo_developer/english/1_0/csw/csw/postget/postgetgetrecords.html

OGC Filter Encoding, http://www.opengeospatial.org/standards/filter

ESRI Geoportal Extension Catalog Service, http://webhelp.esri.com/geoportal_extension/9.3.1/index.htm#geoportal_csw_cmpnts.htm

OGC Filter Encoding 2.0 Encoding Standard, http://docs.opengeospatial.org/is/09-026r2/09-026r2.html