CEOS Working Group
on Information Systems and Services

WGISS/CWIC Data Partner Guide (Catalog Services for the Web)

Publication Date: 2019-04-04

Updated: 2019-04-09

Document version: V1.0

Category: WGISS Technical Document

Editors: WGISS CDA System-Level Team

Executive Summary

Recommendations for CWIC Data Partners

While CWIC can serve as a CSW proxy to search almost any Internet-accessible inventory system for Data Partners, there are a small number of recommendations for Data Partners that will make the job vastly easier.

1. Register each distinct, searchable data set in the IDN
2. Provide a search interface accessible via a simple URL (i.e. HTTP/HTTPS GET), ideally including parameters for starting record number and number of records desired in the response. OGC CSW v2.0.2 or OpenSearch (CEOS compliant) are preferred but not mandatory.
3. Support searching on spatial bounding box
4. Support searching on temporal extent, at least observation start and end dates
5. Identify and, ideally, filter for limitations on search extent (spatial and temporal) to prevent search timeouts
6. Provide search responses in well-structured text (XML, JSON, etc.) returning matching data granules
7. Identify each returned data granule by an identifier that is unique within the inventory system
8. Provide a capability for using the granule identifier to retrieve metadata about the granule
9. Return URLs for browse data and direct access to granule-level data (or to a data ordering system) in the search response

In general, these are common and widely implemented capabilities in almost any granule search system and should not represent an impediment to joining CWIC as a Data Partner. If any of these capabilities are not implemented, it is still possible to become a CWIC Data Partner – contact any of the CWIC team for details.

Table of Contents

1. Before You Begin

This chapter introduces the background, concepts and architecture of IDN, CWIC, and FedEO, within the scope of the WGISS Connected Data Assets. The related skills you will need as a data partner are also discussed.

1.1 CWIC Connected Data Asset Background

For scientists who conduct multi-disciplinary research, there may be a need to search multiple catalogs in order to find the data they need. Such work can be very time-consuming and tedious, especially when different catalogs may use different metadata models and catalog interface protocols. It would be desirable, therefore, for those catalogs to be integrated into a catalog federation which will present a well-known and documented metadata model and interface protocol to users and hide the complexity and diversity of the affiliated catalogs behind the interface. With such a federation, users only need to work with the federated catalog through the public interface or API to find the data they need instead of working with various catalogs individually.

The Committee on Earth Observation Satellite (CEOS) addresses coordination of the satellite Earth Observation (EO) programs of the world's government agencies, along with agencies that receive and process data acquired remotely from space. The Working Group on Information Systems and Services (WGISS) is a subgroup of CEOS, which aims to promote collaboration in the development of systems and services that manage and supply EO data to users world-wide.

NASA's contributions to the CEOS International Directory Network (IDN) provides access to more than 34,000 Earth science data set and service descriptions (stored in the NASA Common Metadata Repository [CMR]) which cover subject areas within Earth and environmental sciences. The IDN's mission is to assist researchers, policy makers, and the public in the discovery of and access to data and related services relevant to Earth science research.

To aid in the search and discovery effort, Global Change Master Directory (GCMD) controlled keywords have been developed and are regularly being refined and expanded. These keywords are also used in other applications within the broader scientific community. Users may perform searches through the IDN website and OpenSearch API using the controlled keywords, free-text searches, map/date searches, or any combination of the above; and may also search or refine a search by data center, instrument, platform, project, or temporal/spatial resolution.

The IDN also supports docBUILDER, a web-based metadata authoring tool that allows metadata authors to add (or modify) data set descriptions (DIFs) that comply with the CMR Unified Metadata Model for Collections (UMM-C). The tool also allows metadata

authors to validate and submit their DIF-10 records directly to the CMR for discovery in the IDN.

To realize a federated catalogue for data discovery from multiple EO data centers, the CEOS WGISS Integrated Catalog (CWIC) system has been implemented. CWIC was initiated and supported by NASA, NOAA, and USGS as a contribution to CEOS. CWIC provides inventory search to WGISS agency catalog systems for EO data by distributing search requests to the appropriate server and sending search responses back to the requesting client. CWIC will provide translation from the CSW search request to the native protocol used by the data partner server if the data partner system does not implement CSW.

FedEO (Federated Earth Observation Gateway) provides a unique entry point to a growing number of scientific catalogues and services for, but not limited to, EO European and Canadian missions. FedEO is deployed with ESA (European Space Agency) infrastructure as a gateway to provide brokered discovery, access and ordering capability to European/Canadian EO missions data based on HMA (Heterogeneous Missions Accessibility) interfaces.

WGISS is now coordinating efforts to connect CWIC and FedEO system with IDN through a common registration of metadata records to seamlessly provide search results for relevant data sets regardless of which system is used to access the granule level data.

1.2 Search Concept and Design

A two-step collection/granule search process, which separates discovery of data collections from searching within relevant data collections to retrieve specific data granules, has been adopted to realize the integrated access to heterogeneous, autonomous data sources.

The WGISS Connected Data Assets system is an implementation of this two-step process. The IDN provides a CSW front end to the collection search. The response from the collection search includes links to the Capabilities Documents at one of several Granule Gateways, providing search capability for granules at the relevant data providers. Current WGISS Connected Data Assets Granule Gateways include the CWIC and FedEO systems. Each of these systems provides access to different data archive systems using the same protocol. Spatial and temporal metadata are the only attributes guaranteed to be supported at all data providers.

1.2.1 Collection Search Criteria

Catalog Services for the Web (CSW) is used as the IDN's collections search implementation based on the OGC CSW v2.0.2 specification. CSW allows clients to formulate CSW-compliant queries against the IDN collections and specify the desired

search results format as XML. The IDN CSW API has implementation the following search fields for users' queries:

- AnyText
- ArchiveCenter
- BoundingBox
- Instrument
- Modified
- Platform
- ScienceKeywords
- TempExtent_begin
- TempExtent_end
- Title

Also, client developers are able to query with specific tags: isCeos, isCwic, isGeoss, and isFedEO. Tagging allows arbitrary sets of collections to be grouped under a single namespace value. The sets of collections can be recalled later when searching by tag fields.

IDN query examples:

- GET request for the IDN capabilities document:
  https://cmr.earthdata.nasa.gov/csw/collections?request=GetCapabilities&service=CSW&version=2.0.2
- POST request for the first 10 IDN collections:

```
<?xml version="1.0" encoding="UTF-8"?>
<csw:GetRecords
  xmlns="http://www.opengis.net/cat/csw/2.0.2"
  xmlns:csw="http://www.opengis.net/cat/csw/2.0.2"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:ogc="http://www.opengis.net/ogc"
  xmlns:gml="http://www.opengis.net/gml"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.opengis.net/cat/csw/2.0.2
http://schemas.opengis.net/csw/2.0.2/CSW-discovery.xsd"
  service="CSW"
  version="2.0.2"
  resultType="results"
  outputSchema="http://www.isotc211.org/2005/gmd"
  startPosition="1"
  maxRecords="10"  >
</csw:GetRecords>
```

- POST request for the first 10 IDN collections containing the GCMD instrument keyword MODIS:

```
<?xml version="1.0" encoding="UTF-8"?>
<csw:GetRecords
  xmlns="http://www.opengis.net/cat/csw/2.0.2"
  xmlns:csw="http://www.opengis.net/cat/csw/2.0.2"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:ogc="http://www.opengis.net/ogc"
  xmlns:gml="http://www.opengis.net/gml"
```

```
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.opengis.net/cat/csw/2.0.2
http://schemas.opengis.net/csw/2.0.2/CSW-discovery.xsd"
  service="CSW"
  version="2.0.2"
  resultType="results"
  outputSchema="http://www.isotc211.org/2005/gmd"
  startPosition="1"
  maxRecords="10"  >
  <csw:Query typeNames="csw:Record">
    <csw:ElementSetName>full</csw:ElementSetName>
    <csw:Constraint version="1.1.0">
      <Filter xmlns="http://www.opengis.net/ogc">
        <PropertyIsEqualTo>
          <PropertyName>instrument</PropertyName>
          <Literal>MODIS</Literal>
        </PropertyIsEqualTo>
      </Filter>
    </csw:Constraint>
  </csw:Query>
</csw:GetRecords>
```

- POST request for the first 10 CWIC IDN collections containing the GCMD instrument keyword MODIS:

```
<?xml version="1.0" encoding="UTF-8"?>
<csw:GetRecords
  xmlns="http://www.opengis.net/cat/csw/2.0.2"
  xmlns:csw="http://www.opengis.net/cat/csw/2.0.2"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:ogc="http://www.opengis.net/ogc"
  xmlns:gml="http://www.opengis.net/gml"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.opengis.net/cat/csw/2.0.2
http://schemas.opengis.net/csw/2.0.2/CSW-discovery.xsd"
  service="CSW"
  version="2.0.2"
  resultType="results"
  outputSchema="http://www.isotc211.org/2005/gmd"
  startPosition="1"
  maxRecords="10"  >
  <csw:Query typeNames="csw:Record">
    <csw:ElementSetName>full</csw:ElementSetName>
    <csw:Constraint version="1.1.0">
      <Filter xmlns="http://www.opengis.net/ogc">
        <And>
          <PropertyIsEqualTo>
            <PropertyName>instrument</PropertyName>
            <Literal>MODIS</Literal>
          </PropertyIsEqualTo>
          <PropertyIsEqualTo>
            <PropertyName>isCWIC</PropertyName>
            <Literal>True</Literal>
          </PropertyIsEqualTo>
        </And>
      </Filter>
    </csw:Constraint>
  </csw:Query>
```

```
</csw:GetRecords>
```

1.3 CWIC Terms and Definitions

For the purposes of this document, the following terms and definitions apply:

(1) catalog ID

Identifiers of data provider serving granule metadata

(2) **client**

A software component that can invoke an operation from a server

(3) CMR

The Common Metadata Repository (CMR) of National Aeronautics and Space Administration (NASA) is a high-performance, high-quality, continuously evolving metadata system that catalogs Earth Science data and associated service metadata records. See IDN definition for its relation to CWIC.

(4) **collection**

A grouping of granules that all come from the same source, such as a modeling group or institution. Collections have information that is common across all the granules they "own" and a template for describing additional attributes not already part of the metadata model.

(5) Catalog Services for the Web (CSW)

Seehttp://www.opengeospatial.org/standards/cat

(6) **data clearinghouse**

The collection of institutions providing digital data, which can be searched through a single interface using a common metadata standard

(7) **dataset**

Has the same meaning as collection, see (8)

(8) GCMD

The Global Change Master Directory (GCMD) is a comprehensive directory of information about Earth science data.

(9) **granule**

The smallest aggregation of data that can be independently managed (described, inventoried, and retrieved). Granules have their own metadata model and support values associated with the additional attributes defined by the owning collection.

(10) **Granule Gateway**

A CEOS server providing access to remote data partner inventory systems.

(11) **granule ID**

A character string that uniquely identifies a single granule to a granule gateway

(12) **identifier**

A character string that may be composed of numbers and characters that is exchanged between the client and the server with respect to a specific identity of a resource

(13) **IDN**

The CEOS International Directory Network (IDN) is a gateway to earth science data and services.

(14) IDN dataset ID

Unique dataset identifier in IDN, returned from the IDN in response to the OSDD request. This identifier is assigned by the IDN CMR database and may be referred to as the "conceptID" in CMR-specific discussions.

(15) native ID

Dataset identifier used by CWIC and FedEO to retrieve granule metadata through data provider API. This identifier is assigned by the data provider but may be the same as the IDN dataset ID.

(16) **operation**

The specification of a transformation or query that an object may be called to execute

(17) **profile**

A set of one or more base standards and - where applicable - the identification of chosen clauses, classes, subsets, options and parameters of those base standards that are necessary for accomplishing a particular function

(18) **request**

The invocation of an operation by a client

(19) **response**

The result of an operation, returned from server to client

1.4 IDN Systems

The IDN, CMR CSW (for IDN), and the GCMD's Keyword Management Service (KMS) only have operational systems which end-users can access.

- IDN CSW server is available to all users.
  Location:
  https://cmr.earthdata.nasa.gov/csw/collections?request=GetCapabilities&service=CSW&version=2.0.2

- KMS - production instance is available to all users.
  Location:
  https://wiki.earthdata.nasa.gov/display/gcmdkey/Keyword+Management+Service+Application+Program+Interface#KeywordManagementServiceApplicationProgramInterface-1Introduction

The IDN site search interface and the CMR CSW production instances will provide access to all datasets which have been registered in the IDN. The KMS production instance will provide access to all approved GCMD keywords registered by IDN providers.

1.5 CWIC Systems

There are two operational CWIC systems to which end-users have access.

- CWIC Operations. This is the current operational system for CWIC and is available to all users.
  Endpoint: http://cwic.wgiss.ceos.org/
- CWIC CSW Capabilities Document
  Endpoint:
  http://cwic.wgiss.ceos.org/discovery?service=CSW&request=GetCapabilities&version=2.0.2
- CWIC Partner Test. This is a test system area used by partners and CWIC developers to test before changes to the CWIC system go operational.
  Endpoint: http://cwictest.wgiss.ceos.org/

1.6 Contact Information

All the documents and information about CWIC are available at WGISS CWIC page at

http://wgiss.ceos.org/cwic

Any questions regarding to CWIC, please send the email to

cwic-help@wgiss.ceos.org

2. CSW Query Interface

2.1 Introduction

The CSW protocol is a catalog service search specification and is used by CWIC to search and return metadata related to granule-level inventory data. CSW is not designed, nor is it used for returning observational data from the inventory systems, although the metadata returned might include links directly to data granules or to a data ordering system. CWIC is intended to take the end user as close to actual data as possible within the constraints of the data partner inventory systems and the limits of the CSW protocol itself.

The CWIC Connectors have the task of returning valid responses to CSW GetRecords and GetRecordById requests to the Mediator. These are generated on-the-fly by submitting search requests to the Data Partner inventory system for the requested dataset, retrieving the results and translating them into syntactically valid and semantically meaningful CSW responses. The Connector implementer will work with the Data Partner's support team to define the mappings between quantities contained in the inventory system response and the associated elements in the CSW responses.

2.2 GetRecords Operation

The CSW GetRecords operation can be used for geospatial catalog searches on the target system with a wide range of parameters. The search parameters supported by CWIC include dataset identifier, spatial (bounding box) and temporal search (start/end date and time).

GetRecords requests can specify one of two types of responses – "hits" or "results." The "hits" request returns only a count of results, no results are actually returned. It turns out that not all inventory systems can easily predict the number of responses to a query without actually processing the query and building the full result set in order to count the records. This can be quite costly in terms of CPU usage and bandwidth, so the CWIC team discourages the use of this request. The "results" request returns actual results, but also includes the total number of matching records, as well as the starting record number and count of the records returned.

GetRecords requests also can specify the result set or type of results returned, i.e., how much information to include in the response for each record.

2.3 GetRecordById Operation

The CSW GetRecordById request is intended to allow the user to request a single specific record from the target system, generally as a follow-up to a broader GetRecords request. No search filter is specified – only the unique identifier for the specific record is

required. The response is identical to the GetRecords response, except that only a single record will be returned.

3. CWIC Metadata Model

3.1 CSW Core Metadata Model

The CSW Core metadata is a small set of metadata elements, essentially the Dublin Core metadata, intended to provide a minimal set of interoperability for CSW servers and clients. Table 1 & 2 of the CWIC Client Guide provide the minimal list of supported search and response elements. For CWIC purposes, the core metadata specification provides definitions for granule identifier, spatial and temporal components as well as the basic required elements for CSW requests and responses (i.e., response type, element set, attributes for result set paging, etc.) and XML representation of the model.

3.2 ISO 19115-2 Metadata Model

The ISO 19115 part 2 metadata is a more extensive set of metadata elements with more complete response models. It is the primary metadata schema currently supported by CWIC. Table 3 of the CWIC Client Guide shows the additional elements, in addition to those in the CSW Core, available to be returned by CWIC in search responses. Many of these may already be included in the responses from the Data Partner's inventory search system, although CWIC will omit any optional elements which cannot be populated from the inventory response and will return empty elements for any mandatory elements which cannot be populated unless information is available from some other source (e.g. contact information).

4. Partner Guidelines

4.1 Metadata & Semantic Mapping

4.1.1 HTTP access

Although CWIC will attempt to use any mechanism available for connecting to Data Partners' data management systems in order to access the available inventory search, there are a few specifics which make the process simpler and more robust.

The use of HTTP for accessing the inventory search engine is strongly preferred. This is widely used already, as web browsers are nearly universal and provide an effective user interface for both human and automated access. While other protocols may be used (Z39.50, for example), HTTP is the preferred mechanism for the CWIC connectors.

Similarly for results, CWIC will attempt to extract the relevant results from any responses the partner data system returns. However, structured text of some sort – XML, for example – is strongly preferred. The ability to easily and definitively parse results makes the process of mapping the metadata returned in the search response simpler and less error-prone. Other structured formats like comma-delimited tables or JSON are acceptable.

4.1.2 Spatial search

All CWIC data partners are expected to support some level of spatial search since all of the inventory data are anticipated to have a spatial component. Simple bounding box, with the bounding coordinates individually identified is the minimum required, although more complex spatial footprint geometries are possible in the future.

It is desirable to have the API also support a dynamic call to return the limits of the spatial search, although not necessary. The presence of such a service can help CWIC avoid invalid or inappropriate search requests, such as those outside the spatial boundaries for specific data collections.

4.1.3 Temporal search

Similar to spatial search described in the previous section, all CWIC data partners are expected to support some level of temporal search since all of the inventory data is anticipated to have a temporal component. Simple temporal extent, with the start and end times individually identified is the minimum required, although more complex temporal relations are anticipated in the future. It is best to support some minimal subset of the ISO 8601 time specification for syntax – YYYY-MM-DD, at least.

It is desirable to have the API also support a dynamic call to return the limits of the temporal extent search, although not necessary. The presence of such a service can help

CWIC avoid invalid or inappropriate search requests, such as those outside the existing temporal extent for specific data collections.

4.1.4 Unique granule IDs

Each data granule returned in a search response should have an identifier associated with it which is unique within the dataset. It is important that the search response include a unique identifier for each granule so that the full data on individual granules may be retrieved without re-executing a (potentially time-consuming) search.

4.1.5 Request for granule by ID

CWIC supports the CSW GetRecordById request and so Connectors expect to be able to submit to the search system a request to return information on a single granule specified by its unique identifier. Generally, this will be so that the Connector can return to the Mediator the full metadata record for that data granule, including links to browse data and to the data granule for download or order.

4.1.6 Record counts

As part of the search response from the inventory system, it is highly desirable to have the total count of matching granules returned, even if the metadata for the granules is not contained in the search response. This parameter, coupled with the ability to specify the starting record number and number of desired records from the inventory system, will allow clients to implement results paging and reducing the load on both the CWIC system and on the data partners.

4.1.7 Search status & Error responses

Useful status and error messages help the Connector manage client sessions effectively. Any limitations on submitted search requests to the inventory systems should be noted in the response (e.g., “too many records requested”, “search timed out”) so that predictable error-handling can be managed by the Connector.

4.2 Interaction & Services Model

4.1.8 Unique granule ID

As described above, each data granule should have a unique identifier which a) is passed back to the client as part of the search response and b) can be used as a key with which to retrieve that specific granule. The CWIC components will manage the task of associating the identifier with the correct dataset and data center.

4.1.9 Contact information

The CSW GetRecords and GetRecordById responses include several blocks of contact information – for distributor, point of contact and metadata contact. These are usually the same for all data granules, and frequently the same within a single data center.

There is no need for this information to be returned with each search response or each data granule, although it might be. The CWIC Connector can cache this information in the CWIC runtime environment, so coordination with the CWIC development team to ensure the accuracy and currency of the contact information is essential.

4.1.10 Browse URL

If browse images of the data granule are available, a valid URL to display the browse image should be included in the search response for each granule so that the client can display it as a link. While it is possible for the CWIC connector to build the URL based on some pre-defined, fixed pattern, this mechanism is not recommended because it removes control over the form of the URL from the Data Partner and changes may require modifications to the Connector source code. This can lead to delays in the deployment of the correct URL when changes are implemented by the data center.

4.1.11 Order URL

The CWIC team recommends that, when the granule data can be downloaded from the data center directly, a valid URL to retrieve the data be included in the search response for the granule.

Alternatively, the search response may contain a URL directing the user to a web site for ordering the data if this is the only option permitted by the data center. This is often necessary even for freely available data if, for example, data center policies require user registration before downloads are made available. In such cases, the CWIC team strongly recommends that the granule ID requested be cached at the ordering system so that whenever the data center requirements for downloading data are met, the user will be able to retrieve the data without re-entering the granule ID.

4.3 Error Handling

The CSW protocol itself has relatively limited capabilities for documenting errors which may arise during a transaction. The CWIC development team is investigating ways to enhance this functionality to provide better information to the end user or client. In order to support this eventuality, it would be useful for the inventory search system to attempt to return sensible and relevant http status codes (where applicable) if something goes wrong with the search or, perhaps even better, a small, descriptive response document (in XML or JSON or whatever the default format might be) providing error codes and error text. In this way, the CWIC connectors can distinguish the type of error arising at the inventory system from those arising elsewhere and take appropriate action. There are no specific recommendations at this time but this should be part of ongoing discussions between the Connector developers and the Data Partner's support staff.

Search timeouts are particularly difficult to identify and manage because they are unpredictable. In many cases, the remote server loses is unable to pass a status code to the client because the connection simply disappears. The underlying search engine or database might simply take too long to build a very large set of results, resulting in the connection from the client timing out. Limitations on spatial or temporal ranges might be imposed by the search engine or database. Best practices for handling such unpredictable situations are under discussion by the WDCA System Level team.