CEOS CWIC Project

CWIC Data Partner's Guide (OpenSearch)

Approval Date: 2017-05-09

Publication Date: 2017-05-10

Reference number of this Document: CWIC-DOC-14-001r010

Document version: V1.0

Category: CWIC Technical Document

Editors: Eugene G. Yu, Lingjun Kang, Archie Warnock, Li Lin

CWIC Implementation Team

Eugene G. Yu, George Mason University (gyu@gmu.edu)
Archie Warnock, A/WWW Enterprises (warnock@awcubed.com)
Li Lin, George Mason University (llin2@gmu.edu)

Approvals

Approved By	Signature	Date
Yonsook K. Enloe		

Document Control

Name	CWIC Data Partner's Guide (OpenSearch)
Doc. Ref. No.	CWIC-DOC-14-001r010
Document status	Under reviewing
Date of release	2017-04-26

Revision History

Date	Version	Brief Description	Author
2014-04-26	0.9	CWIC Data Partner's Guide (OpenSearch) drafting	Lingjun Kang Archie Warnock
2017-04-26	1.0	Revised release	Eugene G. Yu, Archie Warnock, Lingjun Kang, Li Lin

Table of Contents

Executive Summary

Recommendations for CWIC Data Partners

While CWIC can serve as an OpenSearch 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 GET), ideally including parameters for starting record number and number of records desired in the response
3. Support searching on spatial bounding box
4. Support searching on temporal extent, at least observation start and end dates
5. Provide search responses in well-structured text (XML, JSON, *etc.*) returning matching data granules
6. Identify each returned data granule by an identifier that is unique within the inventory system
7. Provide a capability for using the granule identifier to retrieve metadata about the granule
8. 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.

1. Before You Begin

1.1 CWIC 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.

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. 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. To realize a federated catalogue for data discovery from multiple EO data centers, the CEOS WGISS Integrated Catalog (CWIC) system was implemented. CWIC was expected to provide inventory search to WGISS agency catalog systems for EO data.

1.2 CWIC Concept and Design

CWIC uses a mediator-wrapper architecture that has been widely adopted to realize the integrated access to heterogeneous, autonomous data sources. As depicted in Fig. 1, the data source archives data and disseminates it through the Internet. The wrapper on top of the data source provides a universal query interface by encapsulating heterogeneous data models, query protocols, and access methods. The mediator interacts with the wrapper and provides the user with an integrated access through the global information schema.

Wrappers offer query interfaces hiding the underlying data model, access path, and interface technology of the partner catalog systems. Wrappers are accessed by a mediator, which offers users a front-end integrated access through its global schema. The user poses queries against the global schema of the mediator; the mediator then distributes the query to the individual systems using the appropriate wrappers. The wrappers transform the queries so they are understandable and executable by the partner catalog systems they wrap, collect the results, transform them appropriately and return them to the mediator. Finally, the mediator integrates the results as a user response.

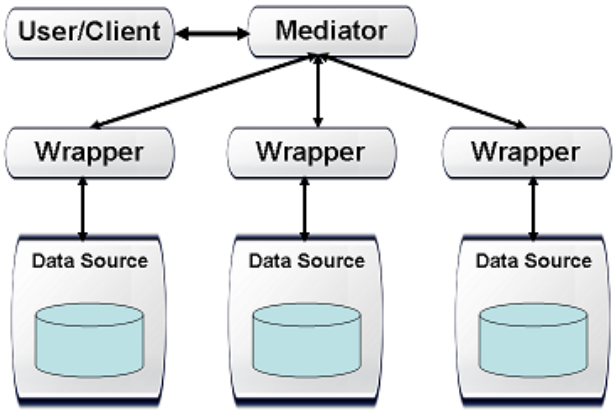

Fig. 1 The Mediator-Wrapper Architecture

Based on the mediator-wrapper architecture, current version of CWIC has been developed and operational with following data partner catalog systems: the Common Metadata Repository (CMR) of NASA, the National Centers for Environmental Information (NCEI) of NOAA, the Group for High Resolution Sea Surface Temperature (GHRSST) of NOAA, the USGS Landsat Surface Imaging (LSI) Explorer, the National Institute for Space Research (INPE) Catalog System of Brazil, the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT), the Canada Centre for Mapping and Earth Observation (CCMEO), the Meteorological and Oceanographic Satellite Data Archival Centre (MOSDAC) of the Indian Space Research Organisation (ISRO), and the National Remote Sensing Center (NRSC) of ISRO.

Different query interfaces were used to access the data partner catalog systems:

Data partner	OpenSearch	OGC CSW	Native query interface
NASA CMR	Yes	No	Yes
NOAA GHRSST	Yes	Yes	No
NOAA NCEI	Yes	Yes	No
USGS LSI	Yes	No	No
Brazil INPE	No	No	Yes
Canada CCMEO	Yes	Yes	No
ISRO MOSDAC	No	Yes	Yes
ISRO NRSC	Yes	No	Yes
EUMETSAT	Yes	No	No

In order to implement a one-stop federated catalog system, wrappers have been developed to implement CWIC OpenSearch for individual member catalogs that do not currently offer that capability.

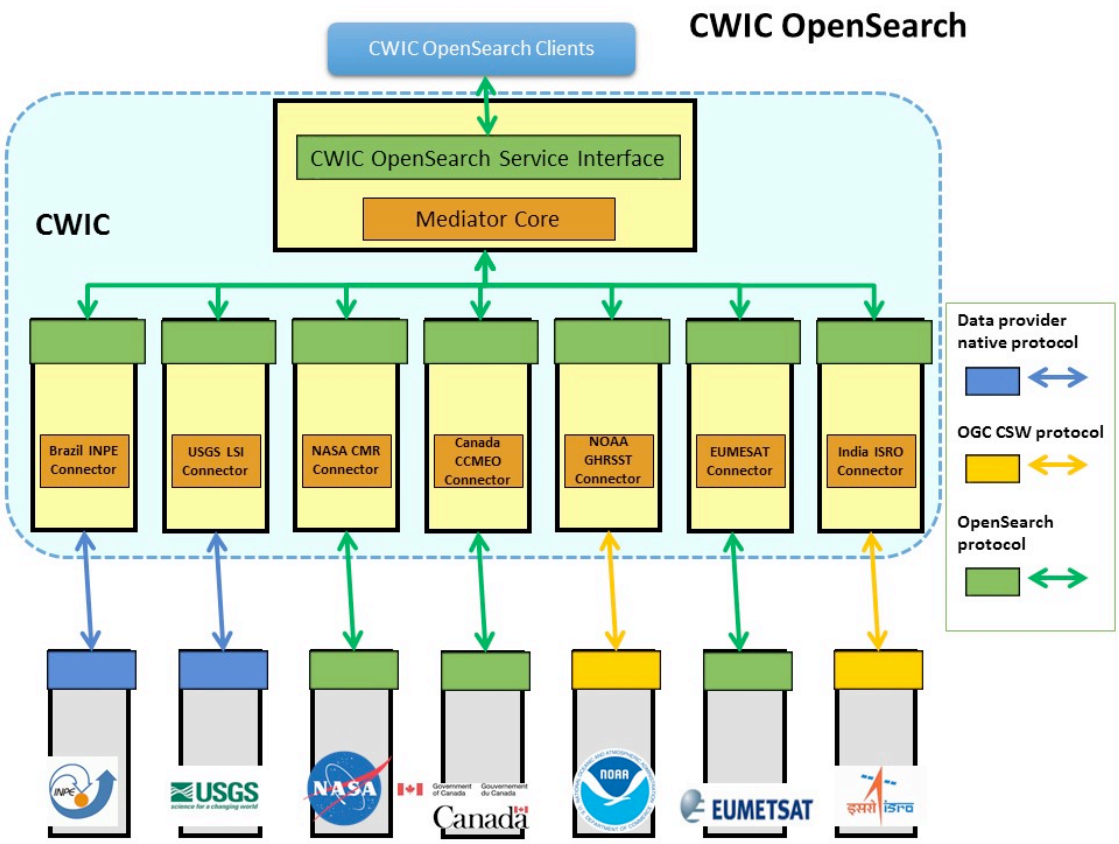

Fig. 2 The System Architecture of CWIC

Fig. 2 illustrates the system architecture of CWIC. Wrappers were implemented for different data partner catalog systems (i.e., NASA CMR, NOAA GHRSST, NOAA NCEI, USGS LSI, INPE, CCMEO, ISRO MOSDAC, and ISRO NRSC). The wrapper is responsible for translating and dispatching request to different data inventories. The mediator is in charge of dispatching the query request to the data partner inventory system and returns the response to data user.

1.3 CWIC Architecture

At its core, CWIC presents to End Users and Clients an OpenSearch server. To Data Partners, it appears to be a web-based client. It connects the two (End Users and Data Partners) through the Mediator on the front end serving as the OpenSearch server to end users and OpenSearch client to Connectors. The Connectors are custom-written proxies for the data granule inventory search systems at the individual Data Partners, accepting OpenSearch search requests from the Mediator, translating them into valid search requests for the target dataset, then parsing the results from the inventory search system and translating those into OpenSearch search responses which are passed back to the Mediator.

In this way, outside clients and, for the most part, the Mediator itself need to have no specific knowledge of the particular partner data systems and communicate only via OpenSearch. Each Data Partner will generally be accessed by a dedicated Connector called by the Mediator. The Connector handles all of the details unique to individual data partner inventory system and all of

the communications with the partner's inventory system is managed exclusively by the connector.

1.4 CWIC Terms and Definitions

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

(1) Client

A software component that can invoke an operation from a server

(2) Data Clearinghouse

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

(3) 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

(4) 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.

(5) Native ID

Dataset identifier used by CWIC to retrieve granule metadata through data provider API. This identifier is assigned by the data provider.

(6) Catalog ID

Identifiers of data provider catalogs or connections serving granule metadata.

(7) Operation

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

(8) 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

(9) Request

The invocation of an operation by a CWIC client

(10)Response

The result of an operation, returned from CWIC server to CWIC client

(11)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.

(12)Dataset

Same meaning as collection, see (8)

(13)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.

(14)IDN

The CEOS International Directory Network (IDN) is a Gateway to the world of Earth Science data.

1.5 CWIC Systems

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

- CWIC PROD – this is CWIC production instance and is available to all users.

  Location: http://cwic.wgiss.ceos.org/

- CWIC TEST – this is CWIC testing instance used by data partners and CWIC clients to perform testing before changes are made to the CWIC production instance.

  Location: http://cwictest.wgiss.ceos.org/

The production instance will provide access to only datasets which have been registered with the IDN. The testing instance may provide access to additional datasets (*e.g.*, new datasets undergoing testing and not yet registered in the IDN), and capabilities which have not yet been tested sufficiently to move to the production system.

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. CWIC OpenSearch Query Interface

2.1 Introduction

OpenSearch is a light-weight search specification[1] used by CWIC to search and return metadata related to granule-level inventory data. CWIC OpenSearch is not designed, nor is it used for returning observational data from the inventory systems, although the metadata returned should include links directly to data granules or to a data ordering system. CWIC OpenSearch 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 OpenSearch protocol itself.

The CWIC Connectors have the task of returning valid responses in Atom[2] format as per CWIC OpenSearch requests to the Mediator. These are generated on-the-fly by submitting search requests to the Data Partner's inventory system for the requested dataset, retrieving the results and translating them into syntactically valid and semantically meaningful CWIC OpenSearch 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 CWIC OpenSearch responses.

2.2 Obtaining OpenSearch Description Document (OSDD)

OpenSearch Description Documents (OSDDs) provide necessary information for clients to programmatically construct valid search requests. Specifically, clients are able to acquire both cardinality and domain of request parameters based on query template. Dataset valids (*i.e.* spatial footprint and temporal extent) are also provided through the OSDD in both machine parsable and human readable formats. With dataset valids, clients are able to construct requests yielding more accurate results.

CWIC provides both generic and dataset specific OSDDs. Clients can fetch a generic OSDD through the CWIC OSDD endpoint. The OSDD request must also include a client identifier string, as recommended by the CWIC OpenSearch Best Practices. Clients can also retrieve a dataset-specific OSDD through the OSDD endpoint by sending both client ID and dataset identifier (*i.e.* DIF Entry ID). In a dataset-specific OSDD, domain is also provided for some parameters (*i.e.* timeStart and timeEnd) in addition to the request parameter syntax.

Generic OSDD request URL example:

http://cwic.wgiss.ceos.org/opensearch/datasets/osdd.xml?clientId=foo

[1] OpenSearch specification verison 1.1 draft 5 (http://www.opensearch.org/Specifications/OpenSearch/1.1)

[2] Atom syndication format (http://tools.ietf.org/search/rfc4287)

Dataset specific OSDD request URL example:

http://cwic.wgiss.ceos.org/opensearch/datasets/C1235542031-USGS_LTA/osdd.xml?clientId=foo

2.3 Search request

The CWIC OpenSearch request operation can be used to conduct granule-level searches on the target system for a range of parameters. It also defines the start page of the result set and maximum number of results per page to be returned. CWIC OpenSearch currently supports following request parameters: unique identifier (dataset/granule ID), spatial (geo:box) and temporal search (time:start/time:end)

3. CWIC OpenSearch response

An OpenSearch response is an ATOM feed with zero or more ATOM entries. Each entry represents the metadata of a single granule pertaining to the query submitted within a set of results also defined by the query.

An ATOM response is a single ATOM feed element containing the following,

1. Information about the search conducted in terms of title, author and ID.
2. Information about the nature of the result set in terms of total number of results, number of results returned and how many the client asked for.
3. Navigation information for traversing that result set including links to the previous, next, first and last results in the set.
4. Zero or more entries pertaining to granule metadata matching the client query

An entry is an ATOM entry element containing the following,

1. Basic information about returned granule in terms of title, author, summary and ID.
2. Information about the spatial footprint of returned granule.
3. Information about the temporal extent of returned granule.
4. Collection of ATOM link elements containing endpoints of granule browse, order and metadata for the returned granule.

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.

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. For each granule, a simple bounding box, with the bounding coordinates individually identified, is the minimum required, although more complex spatial footprint geometries are possible in the future. The number of spatial footprints in data partner's response is not limited to one.

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 Request for dataset by ID

Dataset ID (cwic:datasetId) serves as a mandatory parameter in the CWIC OpenSearch query interface. CWIC will query against data partner's search system, using the dataset ID, to retrieve the matching granule level metadata. From this point, all data partners are expected to support the query based on dataset identifier. In particular, a globally unique dataset identifier (i.e. DIF Entry ID) is recommended to be implemented as query parameter on data partner side. However, other unique dataset identifiers within data partner system is also acceptable.

4.1.5 Request for granule by ID

CWIC OpenSearch supports searching for a single granule by granule ID. CWIC will query against data partner's search system with the granule ID in order to retrieve the matching granule level metadata. From this point, all data partners are recommended to support the query based on granule identifier.

4.1.6 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 the unique identifier for each granule so that the full data on individual granules may be retrieved without re-executing a (potentially time-consuming) search.

4.1.7 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.

4.1.8 Pagination

Pagination is supported (*i.e.* search by specifying start page and counts per page) in the CWIC OpenSearch query interface. The search based on pagination parameter (e.g. start page and item per page) or cursor parameter (e.g. start index and item per page) is expected to be implemented by the Data Partner.

4.1.9 Other search parameters

The Data Partner inventory systems may support search on additional parameters like LANDSAT path/row. These are not supported through CWIC OpenSearch at the current time but pose no particular issues for the CWIC connectors.

CWIC is also considering support for results sorting – by start time, by duration, by granule ID, etc. – and it would be useful, although not mandatory, to have a sort option available from the Partner inventory system.

4.1.10 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

4.2.1 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.

4.2.2 Contact information

The CWIC OpenSearch responses include contact information for each granule. These are usually the same for all data granules, and frequently the same for all datasets 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, but this mechanism is not recommended because it removes control over the contact information from the Data Partner and changes may require modifications to the Connector source code. This can lead to delays in including the correct contact information when changes are implemented by the data center.

4.2.3 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.

4.2.4 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.

4.2.5 Pagination support

As stated above, there is no strict requirement in terms of pagination mechanism or pagination parameters imposed by CWIC. Various pagination mechanism and heterogeneous parameters are expected across data partners. To provide universal pagination, the CWIC connector translates startPage and count (stipulated by CWIC OpenSearch Best Practices) to native pagination parameters for each Data Partner.

4.2.6 Spatial search

As stated above, a Data Partner is expected to support spatial search and include proper spatial component in each granule response. Native spatial components (one or many) from data partner are converted syntactically and semantically to proper spatial components (e.g. georss:box and georss:polygon) which are stipulated by CWIC OpenSearch Best Practices. Besides that, a minimum bounding rectangle (MBR) is also calculated based on native spatial components for each granule if it is not returned explicitly in the response and included in CWIC OpenSearch response.

With the knowledge of Data Partner specific limitations on spatial search, CWIC will impose validation on the spatial request parameter and return both the HTTP status code and an error message if necessary.

4.2.7 Temporal search

As stated above, a Data Partner is expected to support temporal search and include standard (ISO 8601 compliant) temporal elements in each granule response. Temporal elements with native format from data partner are converted syntactically and semantically to single temporal element (i.e. dc:date) which are stipulated by CWIC OpenSearch Best Practices.

With the knowledge of Data Partner specific limitations on temporal search, CWIC will impose validation on temporal request parameter and return both the HTTP status code and an error message if necessary.

4.3 Error Handling

Error handling in CWIC OpenSearch is based on standard HTTP status codes. The CWIC development team is investigating ways to enhance the error logging and documenting. Effort has also been dedicated to provide sensible error messages in addition to the generic HTTP status code 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.