CEOS CWIC Project

CWIC Data Partner's Guide

May 10, 2017

Document version 1.2

CWIC Implementation Team

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

Approvals

Approved By	Signature	Date
Yonsook K. Enloe		

Revision History

Date	Version	Brief Description	Author
	1.0	CWIC Data Partner’s Guide	Archie Warnock
12 March 2012	1.1	CWIC data partner guide	Archie Warnock Yuanzheng Shao
10 May 2017	1.2	Revised CWIC CSW data partner guide	Li Lin, Archie Warnock, Eugene G. Yu, Lingjun Kang

Table of Contents

Executive Summary 4

Recommendations for CWIC Data Partners 4

1. Before You Begin 5

1.1 CWIC Background 5

1.2 CWIC Concept and Design 5

1.3 CWIC Architecture 6

1.4 CWIC Terms and Definitions 7

1.5 CWIC Systems 8

2. CSW Query Interface 9

2.1 Introduction 9

2.2 GetRecords Operation 9

2.3 GetRecordById Operation 9

3. CWIC Metadata Model 10

3.1 CSW Core Metadata Model 10

3.2 ISO 19115-2 Metadata Model 10

3.3 WGISS Search Criteria ..... **Error! Bookmark not defined.**

4. Partner Guidelines 11

4.1 Metadata & Semantic Mapping 11

4.1.1 HTTP access 11

4.1.2 Spatial search 11

4.1.3 Temporal search 11

4.1.4 Keyword search ..... **Error! Bookmark not defined.**

4.1.5 Other search parameters ..... **Error! Bookmark not defined.**

4.1.6 Unique granule IDs 12

4.1.7 Request for granule by ID 12

4.1.8 Record counts 12

4.1.9 Search status & Error responses 12

4.2 Interaction & Services Model 12

4.2.1 Unique granule ID 12

4.2.2 Contact information 12

4.2.3 Browse URL 13

4.2.4 Order URL 13

4.3 Error Handling 13

Executive Summary

Recommendations for CWIC Data Partners

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

1. Register 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 link 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 is very time-consuming and tedious, especially when the 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, CEOS WGISS Integrated catalog (CWIC) was implemented. CWIC was expected to provide inventory search to WGISS agency catalog systems for EO data.

1.2 CWIC Concept and Design

The mediator-wrapper architecture 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 disseminate 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 particular 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, and return them to the mediator. Finally, the mediator integrates the results as a user response.

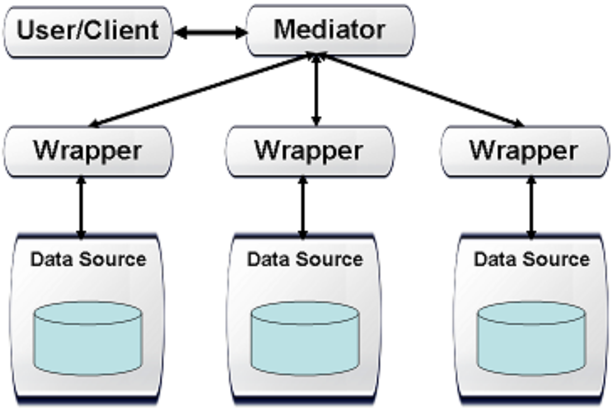

Fig. 1 The Mediator-Wrapper Architecture

The data providers connected by CWIC include NASA, USGS, NOAA GHRSST, Brazil INPE, Canada CCMEO, EUMETSAT and India ISRO (MOSDAC and NRSC). Additionally, the CWIC connector connecting Australia NCI and China AOE are under development in CWIC development server. Fig. 2 illustrates the system architecture of CWIC. Different wrappers were implemented for different data providers. The wrapper is responsible for translating and dispatching the request to different data inventories.

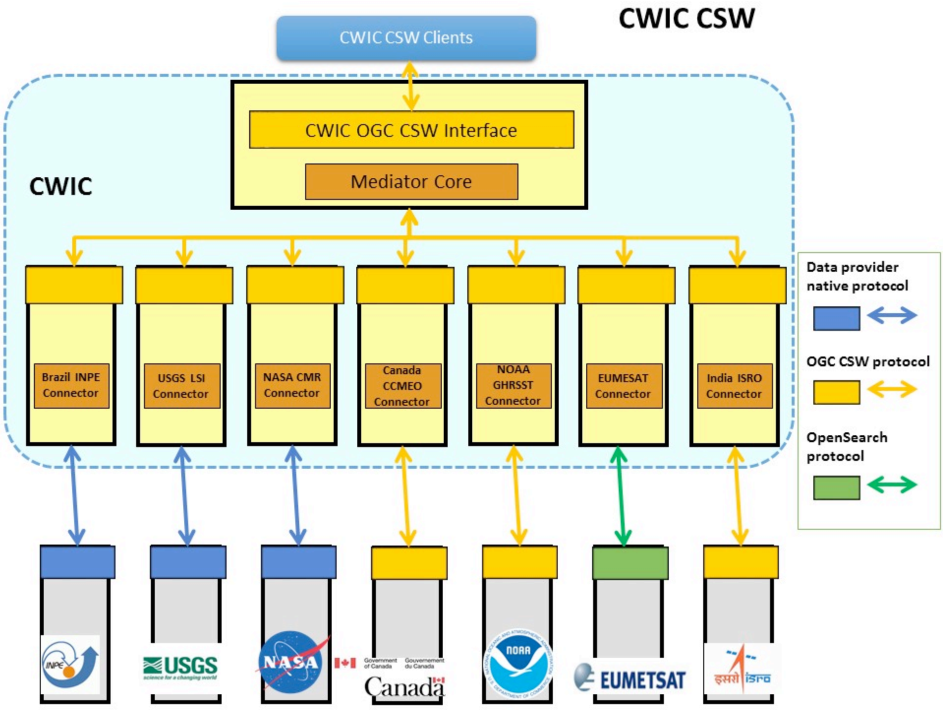

Fig. 2 The System Architecture of CWIC

1.3 CWIC Architecture

At its core, CWIC presents to End Users and Clients a standards-based CSW server. To Data Partners, it is 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 CSW server to end users and a

CSW client to the Connectors. The Connectors are custom-written proxies for the data granule inventory search systems at the individual Data Partners, accepting CSW 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 CSW 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 CSW. 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

Has the same meaning as collection, see (11)

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, a Gateway to the world of Earth Science data and services

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 CWCI and is available to all users.
  Endpoint: http://cwic.wgiss.ceos.org/
- 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/

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

4.2.1 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 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.

4.2.3 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.

4.3 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.