DISTRIBUTED CATALOGUE SEARCH OF EARTH OBSERVATION DATA

by

Huilin Wang
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Earth Systems and Geoinformation Sciences

Committee:

_______________________________ Dr. Liping Di, Dissertation Director

_______________________________ Dr. David Wong, Committee Member

_______________________________ Dr. Chaowei Yang, Committee Member

_______________________________ Dr. Jeff Offutt, Committee Member

_______________________________ Dr. Peggy Agouris, Department Chairperson

_______________________________ Dr. Timothy L. Born, Associate Dean for Student and Academic Affairs, College of Science

_______________________________ Dr. Vikas Chandhoke, Dean, College of Science

Date: _________________________ Spring Semester 2013
George Mason University
Fairfax, VA

Distributed Catalogue Search of Earth Observation Data

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

by

Huilin Wang
Bachelor of Science
Wuhan University, 2006

Director: Liping Di, Professor
Department of Geography and Geoinformation Science

Spring Semester 2013
George Mason University
Fairfax, VA

DEDICATION

This is dedicated to my parents and my sister.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Liping Di for his support and guidance throughout my study and stay at George Mason University.

I would like to thank all committee members for their guidance, patience and time on the dissertation.

I would like to thank all the members of the Center for Spatial Information Science and Systems (CSISS) at George Mason University for their help and friendship in my research and life.

I would like to thank Yuanzheng Shao, Lingjun Kang, Ranjay Shrestha and Genong Yu for the help and suggestions on the dissertation.

I would like to thank Yuqi Bai for his mentorship during his stay in CSISS.

Finally, I would like to thank my family for their support and love.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Academy of Optic-Electronic	AOE
Application Programming Interface	API
Alternative Query Language	AQL
Canada Centre for Remote Sensing	CCRS
Catalogue Federation Service	CFS
CEOS WGISS Integrated Catalog	CWIC
Committee on Earth Observation Satellites	CEOS
Common Object Request Broker Architecture	CORBA
Common Query Language	CQL
Comprehensive Large Array-Data Stewardship System	CLASS
Content Standard for Digital Geospatial Metadata	CSDGM
Database Management system	DBMS
Dublin Core Metadata Initiative	DCMI
Earth Observation	EO
Earth Observation System Data and Information System	EOSDIS
Earth Observing System	EOS
Earth Observing System Clearing House	ECHO
EOSDIS Core System	ECS
Electronic Business Registry Information Model	ebRIM
Electronic Business using eXtensible Markup Language	ebXML
eXtensible Markup Language	XML
Federal Geographic Data Committee	FGDC
Geographic Information System	GIS
Global Change Master Directory	GCMD
Graphical User Interface	GUI
Group for High Resolution Sea Surface Temperature	GHRSST
Hypertext Transfer Protocol	HTTP
Identifier	ID
Integrated Catalogue Framework	ICF
Interface Definition language	IDL
International Organization for Standardization	ISO
Internet Inter-ORB Protocol	IIOP
Japan Aerospace Exploration Agency	JAXA
Key Value Pair	KVP
Metadata Catalogue Service	MCS
National Aeronautics and Space Administration	NASA

National Institute for Space Research	INPE
National Oceanic and Atmospheric Administration	NOAA
National Oceanographic Data Center	NODC
Open Geospatial Consortium	OGC
OpenGIS Catalogue Service for Web	CSW
Organization for the Advancement of Structured Information Standards	OASIS
Spatial Data Infrastructure	SDI
Structure Preserving Semantic Matching	SPSM
Term Frequency Inverse Document Frequency	TF-IDF
U.S. Geological Survey	USGS
Uniform Resource Identifier	URI
Web Coverage Service	WCS
Web Ontology Language	OWL
Working Group on Information Systems and Services	WGISS

ABSTRACT

DISTRIBUTED CATALOGUE SEARCH OF EARTH OBSERVATION DATA

Huilin Wang, Ph.D.

George Mason University, 2013

Thesis/Dissertation/Project Director: Dr. Liping Di

Catalogues in different organizations use different metadata information models, catalogue interfaces, and transport protocols to support discovery, access and use of Earth Observation (EO) data. The heterogeneity of these catalogues makes it difficult for scientists to find required data. There have been various efforts to integrate distributed catalogues to facilitate EO data search, however, an integrated framework that supports EO data discovery on the dataset level and the data granule level simultaneously is still lacking. This dissertation addresses that gap and designs an integrated catalogue framework (ICF) to integrate distributed catalogues for EO data products. The ICF presented in this study will enhance EO data search by unifying different models and interfaces used in distributed catalogues and releasing users from the heterogeneity of these geospatial catalogues services. This framework not only offers a harmonized interface for users to access distributed catalogues but also supports two levels of search granularities: 1) keywords-based dataset search and 2) data granule search.

To achieve interoperability for the proposed ICF, two levels of mapping are required for each search granularity: One is the mapping of query language to hide users from the various heterogeneous query interfaces; the other is the metadata information model mapping. These two levels of mapping offer users not only a consistent mechanism to specify search criteria but also unified and integrated search results. The framework adopted the OpenGIS Catalogue Service for Web (CSW) specification as the core-underlying standard and leveraged the concept of mediator and wrappers for resources integration. Search requests are translated to CSW Query Language (CQL) and dispatched to distributed catalogues by a mediator. Search results from these catalogues are mapped by wrappers to a format compatible with the CSW 2.0.2 - ISO Metadata Application Profile (OGC 07-045) and the Core Profile derived from the Dublin Core Metadata Element Set. A detailed use case for integrating two different catalogues is described to show the capabilities of this framework to support distributed Earth Observation data search on two levels.

CHAPTER ONE INTRODUCTION

1.1 Background

As large volumes of Earth Observation (EO) data are created and collected, it is crucial to find a solution for sharing and utilizing geographic information. Early efforts have been made to facilitate data sharing by upgrading the traditional models of stand-alone systems to distributed Geographic Information System (GIS) web services. Currently numerous catalogues available online from different agencies are used for geospatial data discovery and retrieval. Both their virtual and physical locations are widely distributed. The strategies for data storage and updating frequency of those catalogues vary among to different satellites, earth observation data, metadata models, and a variety of other variables. As the number of services and catalogs available in an environment grows, there will be an increasing need for more sophisticated search-engine-like tools that can consolidate, organize and present information retrieved from various sources (Alameh 2003).

In the field of spatial data infrastructures, the international standards of the Open Geospatial Consortium (OGC) and the International Organization for Standardization (ISO) form the basis of most existing catalogue interface implementations (Senkler et al., 2004). OGC is an international industry consortium of more than 230 companies, government agencies, and universities aiming at growing the interoperability of technologies involving spatial information and location. Its mission is to promote the development and use of advanced open system standards and techniques in the area of geo-processing and related information technologies by delivering spatial interface specifications that are openly available for global use (Nogueras, 2005). OGC web services specifications allow seamless access to geospatial data in a distributed environment, regardless of the format, projection, resolution, and the archive location (Wei et al., 2005). One of those specifications is the OGC CSW, which defines standard interfaces for data discovery, metadata query, and other services. These interfaces have been widely used in developing web services, such as the grid-enabled web services by Chen et al. (2007), and the ontology-based search for interactive digital maps by Hubner et al. (2004).

Interoperability of geospatial web services is achieved by using standards, mainly from the Federal Geographic Data Committee (FGDC), the ISO and the OGC (Yang et al., 2010). There are six levels of interoperability between two or more spatially distributed independent GISs (Bishr, 1998). Based on Bishr's definition of interoperability levels, the lowest level of interoperability is network protocol, followed by hardware & OS, spatial data files, Database Management system (DBMS), data model, and application semantics. Interoperability can occur at any of those levels. In general, interoperability means both data level and program level (Laurini, 1998). A common issue with standardization is the development and creation of metadata for EO products. Data providers use metadata to describe their data collections. Different abstract information models are adopted to create those metadata. Some of the metadata may be created based on standard specifications. However, most of the metadata are customized by spatial agencies due to the heterogeneity of the EO data collected. Further complications arise from the catalogues which are developed to support discovery, search, access and use of EO data.

With the development of new technologies and growing investment in interoperability, great efforts and progress have been made to enhance the interoperability of data sharing and searching abilities for EO data. One method is to utilize the mediator-wrapper architecture to standardize distributed catalogues. Proposed by Wiederhold in 1992 (Widerhold, 1992), the mediator-wrapper architecture has been widely used in integrated access to multiple data and information sources. For example, Naumann et al. (1999) adopted this architecture to do quality-driven integration of heterogeneous information systems. Chang et al. (2010) built a metadata classification assisted scientific data extraction architecture based on Wiederhold’s mediator-wrapper. The advantage of adopting the mediator-wrapper architecture is that heterogeneous online catalogues can be wrapped in a standard way, thus facilitating the integration of those catalogues. With the commonly recognized OGC specifications available, it is possible to develop a wrapper for each of the distributed catalogues based on CSW. The wrapper then serves as a plug-and-play function for the mediator. The mediator takes searching criteria from users, analyzes the requests, and then decides to which wrapper each query should go. The responses from wrappers are also processed by the mediator before they are returned to users. Multiple ways of dispatching strategies are described in detail in this dissertation.

With the integration of different catalogues, the scope of EO data search is greatly extended. Users can access distinct catalogues through a unified interface. However, distributed EO data search is a 2-dimensional search. Not only should the scope of THE search be considered, but also the granularity. The top level for the granularity of EO data search is the dataset level. In the National Aeronautics and Space Administration's (NASA) community, a dataset is defined as a collection of data granules that usually share the same information model. The NASA Global Change Master Directory (GCMD) holds metadata for more than 28,000 Earth science datasets. With the aid of GCMD, metadata for datasets can be easily discovered. The second level for the granularity of EO data search is the granule level search. Granule is the smallest aggregation of data that can be independently managed. Most online catalogues sit at this level. As an example, the Comprehensive Large Array-Data Stewardship System (CLASS) of the National Oceanic and Atmospheric Administration (NOAA) holds 74 datasets. When users choose one dataset and set searching criteria such as spatial and temporal extents, information for corresponding granules is returned. The third level is the data coverage level. Coverage is digital geospatial information representing space/time-varying phenomena. At this level, users already have the granule information and are searching for the coverage information. Once the coverage is acquired, it can be processed by Web Coverage Services (WCSs) that allow further processing, such as reformatting, reprojection, and subsetting operations. Figure 1 illustrates the scope and granularity of distributed EO data search.

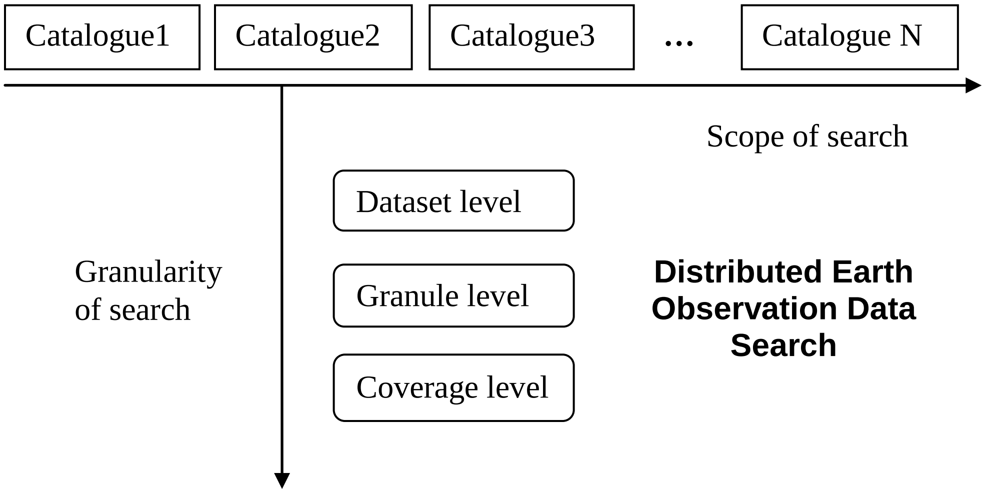


Figure 1 Scope and granularity of Earth Observation data search

When integrating a heterogeneous distributed catalogue search for distributed EO data, two dimensions need to be taken into consideration. Since the granule is the smallest downloadable unit, most users are interested in this level. The following part of this dissertation focuses on EO data search on dataset level and granule level. An ideal concept for integrating catalogues is that numerous catalogues could be embedded into one that supports a two-level granularity search, so that users are better able to search and obtain the information for the EO data they request.

1.2 Motivation

The motivation for this dissertation is the interest in interoperability for online systems and services for Earth Observation data. With large volumes of satellite data collected every day, online systems and services to manage the discovery, search, and access of EO data are increasingly popular. Countless supporting Web services have been developed. In recent decades, traditional Web services for data sharing have evolved into catalogue services. A catalogue service facilitates sharing, discovery, retrieval, management of, and access to large volumes of distributed geospatial resources. Chen et al. (2010) show examples of these resources: data, services, applications, and their replicas on the Internet. Each space agency may develop its own catalogue, requiring users to understand the query languages, information models for data, and query interfaces related to the specific EO data catalogue search. This leads to the new requirement for an integrated catalogue which will offer a harmonized interface and commonly recognized information model for EO data.

Most catalogues support data search on the granule level, which means even though these catalogues are integrated, users need to have basic knowledge about the target data granule and to which dataset it belongs. From this perspective, the granularity of EO data search needs to be extended to enhance the interoperability for data sharing.

1.3 Objective and Contribution

The primary objective is to fill the gap where multiple levels of distributed EO data discoveries are missing and enhance the interoperability of EO data search across distributed heterogeneous catalogues. This objective is accomplished by designing an integrated framework which will be capable of extending both the scope and granularity of EO data search and providing a consistent search mechanism to release users from the complexity of heterogeneous geospatial catalogue services. For users without much domain knowledge, the framework should also support keyword search. Given the requirements for integrating catalogues and extending granularity of EO data searching, the framework is designed to offer a harmonized interface for users to access distributed catalogues and support two levels of searching granularities: keyword-based search to obtain dataset level information and data granule search obtain granule level information.

To fulfill the requirements for integrating catalogues, mediator-wrapper architecture is adopted. The basic concept of the mediator-wrapper architecture is to create a wrapper for each of the data sources and for each of the catalogues. The wrapper serves as a plug-and-play function for the mediator. The mediator is developed as a dispatching center which takes queries from users, analyzes the requests, and then decides to which wrapper each query should go. Each wrapper interacts with a system connector that directly communicates with a corresponding catalogue through the catalogue’s specific protocol. The wrapper provides two types of mapping: the input and output of its corresponding system connector. The objective of the input wrapping is to hide users from the various heterogeneous query interfaces; the output wrapping offers users a consistent structured output.

A set of commonly recognized specifications is applied in the process of requests and results conversion. This will be discussed in Chapter 3. The strategies and core modules for integrating catalogues will be discussed in Chapter 4.

To fulfill the objective of extending the granularity of EO data searching, searching for the dataset level is added for all datasets in each catalogue that is integrated into the framework. The dataset level search allows users to use keywords to get the dataset level information. The basic concept is to search against a directory of datasets in catalogues using keywords to get the metadata for datasets. This research exploits two ways in which the keywords search is applied at the dataset level. One is to build a new dataset directory with keywords. The other is to harvest dataset metadata from GCMD and then save the dataset metadata in GeoNetwork, which is an open source catalog application. The strategies will be discussed in detail in Chapter 4.

To illustrate the abilities of the framework to integrate distributed catalogues and support two-level distributed Earth Observation data, a use case is demonstrated in Chapter 5 with two catalogues used as data resources. The NOAA CLASS system offers users an interface protocol to search their available data from 74 types of data archives while the Earth Observing System (EOS) Clearing House (ECHO) system of NASA manages 11 topics of data: agriculture, atmosphere, biosphere, climate indicators, human dimensions, hydrosphere, land surface, oceans, solid earth, spectral/engineering, sun-earth interactions and terrestrial hydrosphere. Updating frequencies vary especially among the data archives of the same catalogue. In addition the updating frequencies are different according to temporal scales of the satellite instruments. The two catalogues also offer different search interfaces. The former offers an online Graphical User Interface (GUI) while the latter offers a set of Application Programming Interface (API) for searching metadata. As mentioned in the abstract, the two catalogues will be standardized and integrated into one integrated catalogue according to the strategies provided in the designed framework. A dataset level search and a data granule search for these two catalogues will be developed.

This research contributes to the distributed catalogue search of Earth Observation data field in the following ways. First, a framework which supports two levels of data searching to enhance the interoperability for distributed catalogued search of EO data is developed. Second, a solution for integrating distributed catalogues is provided to extend the scope of EO data search. Third, strategies for developing keywords search function is provided to extend the granularity of EO data search.

1.4 Dissertation Organization

The rest of the dissertation is organized as follows. In Chapter 2, the author reviews the current situation for EO data management, presents the current research effort to integrate catalogues for EO data search, and discusses some of the major weaknesses in current integrated catalogues. An overview of the technology crucial for developing the framework is also provided. In Chapter 3, the author describes the specifications which are applied in this research to enhance interoperability for both data level and program level, including the abstract information model for metadata standardization and OGC catalogue services for developing an integrated catalogue. In Chapter 4, the development of the integrated catalogue framework is described in detail. The core modules for each level of search granularity are explained in detail. In Chapter 5, a detailed use case for integrating two distributed catalogues demonstrates the ability of the framework to support a two-level distributed Earth Observation data search. In Chapter 6, the author presents and discusses the result from the use case. In Chapter 7, conclusions and future developments of this framework are presented.

CHAPTER TWO LITERATURE REVIEW

2.1 Earth Observation Data Management

The amount of Earth Observation digital data is growing at a rapid rate. In addition to the increase in volume, the number of sites making information available is also increasing (Miller and Nusser, 2003). In early days, most GIS applications were stand-alone systems and access to geospatial data was quite limited due to the lack of data discovery and sharing technologies. With the generalization of the Internet and development of Web service technology, data sharing has improved dramatically. Organizations and clients have made their own online systems to facilitate management and distribution of their data. In more recent years, users began to realize the inefficiency and redundancy of the data provided by a batch-oriented approach. With the rapid development of information systems and distributed database paradigms, GIS users realized the need for interoperable geographical information systems (Bishr 1998). To manage these large datasets efficiently, metadata or descriptive information about the data needs to be accommodated accordingly (Singh, et al., 2003). Metadata catalogue services are developed so that users can easily obtain metadata information about EO data necessary for educational or research needs. Since metadata is quite an ambiguous term, it can be in any format and contain any content about the described data. Many information models are developed as constraints for metadata. Different data providers and catalogue services use different metadata standards to describe specific data products.

Through reviewing several geospatial catalogues, I have identified the advantages of geospatial catalogues in facilitating discovery, search and access of EO data. In this section, I first describe the development of EO data management and then discuss several metadata catalogue services related to EO data. At the end of this section, interfaces, information models, and granularity of search for several online operational geospatial catalogues are examined.

2.1.1 Discovery, Search, Access and Usage of EO Data

The methods of discovery, search, access and usage of EO data have changed significantly since the last century. There have been six distinct phases in data management (Gray, 1996). The first phase, when data was manually processed, can be dated back to 4000BC to the 20th century. In the next step, punched-card equipment and electro-mechanical machines sorted and tabulated millions of records. In the third phase, data was stored on magnetic tape and computers used stored programs to batch process sequential files (Gray, 1996). In the 1960s, the introduction of online data network databases demonstrated the start of the fourth phase for data management. The fifth phase started with the emergence of relational database and client-server computing technology in the early 1980s. We are now in the early stages of sixth generation systems that store richer data types, notably documents, images, voice, and video data (Gray, 1996). Data management can be applied to a broad range of fields besides EO in GIS. It can also help to specify, develop, integrate, and test tools and middleware infrastructure to coherently manage and share petabyte-range information volumes in high-throughput production-quality Grid environments (Hoschek et al., 2000).

As a new and emerging technology in the early 1970s, GIS had a profound influence on the capabilities of geographic analysis, and in retrospect marked a turning point in the reinforcement of geography as an explicitly spatial discipline (Dragicevic, 2004). With the development of the Internet, GIS is now able to make its concepts more open, accessible, and mobile to everyone, thereby facilitating notions such as democratization of spatial data, open accessibility, and effective dissemination (Dragicevic, 2004). Dragicevic also points out that Web-based GIS has enhanced the open use of GIS in three main directions: 1) spatial data access and dissemination, 2) spatial data exploration and geovisualization, and 3) spatial data processing, analysis and modeling. The advantages of GIS in these three directions can be embodied simultaneously in one system. Landrau (2002) designed a mechanism for accessing geospatial data to support research, monitoring and environmental management activities in the island municipality of Vieques and implemented a prototype of this mechanism. The functions of the prototype he developed illustrated the benefits of the integration of GIS and the internet in these three directions: 1) search, view, downloading, and uploading datasets; 2) query and zoom of datasets; and 3) printing interactive maps. In Landrau’s research, only one simple database was involved. However, the discovery, search, access and usage of EO data usually involve interactions with multiple distributed databases. Rather than centralizing geographic information into a unique database, an interesting solution is to federate all information stored into different databases or sites (Laurini, 1998). As a result, the integration of distributed catalogues of EO data is a solution for enhancing the interoperability of EO data discovery, search, and access.

2.1.2 Metadata Catalogue Service

A large number of catalogues have been developed to support the discovery, search, access and usage of Earth observation data. Although spatial agencies have developed their own diverse catalogues in the past, it has become increasingly important to build catalogues using commonly recognized standards. Some early efforts on metadata catalog services were made by Singh et al. (2003). The authors realized that metadata or descriptive information about data needed to be developed to manage large data sets efficiently. As such they presented a design of a metadata catalog service (MCS) that provided a mechanism for storing and accessing descriptive metadata and allowed users to query for data items based on desired attributes (Singh et al., 2003). In their research, the definition, role, requirements, and components of a metadata service are described. Singh also presented the design and implementation of MCS, described the experiences in using MCS with two different applications and discussed the scalability of MCS. This can be considered as an early design and implementation of metadata catalogue service; however, it is not as popular as the OGC metadata catalogue services presently available.

The Open GIS Consortium has made great efforts in the field of spatial data sharing, service categorization and standardization of service interfaces. It offers a series of international standards to support interoperable solutions for geospatial interoperability. An increasing number of organizations are developing online services for sharing geospatial information based on OGC standards. For this dissertation, the specific standards of the OpenGIS catalogue service implementation specification (Nebert et al., 2007) and the ISO metadata application profile (Voges and Senkler, 2007) are adopted in the implementation of the integrated framework. With the goal of offering a standard search interface and metadata information model, a lot of effort has been made in developing OGC CSW standards since 1999. Nowadays, OGC CSW is widely adopted in the field of GIS, especially in geospatial catalogue web services. In 2005, by combining OGC CSW technologies and grid technology which enables large-scale data sharing, a grid-enabled catalogue service was designed and implemented by Chen et al (2005). The OGC Catalogue Service is grid-enabled by introducing Grid Services into the catalogue service and enabling its interoperability with other useful grid services to facilitate the sharing of geographic information (Chen et al., 2005; Wei et al., 2005). While CSW significantly facilitates the discovery of data and services, current discovery processes are based on the static keyword match without the full exploration of the underlying semantics, such as hierarchical relationships among metadata entities. Semantic augmentations to CSW can improve the discovery ability of data and services (Yue et al., 2006). Extending the elements of the Electronic Business Registry Information Model (ebRIM) profile, as recommended by CSW, provides the semantic information of geospatial data registration and services. Based on CSW and data typed indices, an efficient search and discovery system for heterogeneous EO metadata can be implemented (Kojima et al., 2010). An R-tree-based polygon index is added as a plug-in in this system since the indices of metadata have their own data types and ranking/scoring mechanisms. With this method, CSW is utilized by Kojima et al. (2010) for an efficient search and discovery system for EO data. Shen et al. (2012) developed a catalogue service for internet GIS services supporting active service evaluation and real-time quality monitoring which solved three main problems of catalogue services (Shen et al., 2012):

1) lack of the capability to discover services actively,
2) lack of the capability to monitor service status,
3) lack of accurate quality description for published GIS services.

This catalogue service is quite useful for active service evaluation and real-time quality monitoring. However, with interoperability and data sharing, this system is inadequate since it focuses on service registration, status, and service qualities.

Govedarica et al. (2010) give a short review of the OGC metadata catalogue services and its key role in geospatial resource discovery in Spatial Data Infrastructures (SDI). They point out that due to the lack of appropriate documentation of data and lack of metadata semantics, the full potential of metadata catalogues have not yet been achieved. Thus it is necessary to get a set of general metadata properties that can be used to characterize any resource. Four types of metadata models are described by Govedarica et al. to extend the catalogue abstract information model: 1) Dublin Core, 2) ISO 19115 Geographic Information –Metadata, 3) Advancing Open Standards for the Information Society (OASIS) Electronic Business using eXtensible Markup Language (ebXML) Registry Information Model (ebRIM) (Fuger et al., 2005) and 4) Web ontology language. The problem of the semantics of data is also discussed by the authors. Govedarica et al. conclude that the OGC Catalogue specification, with which various vendors must comply to in order to achieve interoperability, enables access of geospatial metadata independent of the nature of search client applications.

Metadata catalogues are developed with different query interfaces and metadata information models, and have different granularities of search. Some current operational online catalogues are compared in table 1: NOAA CLASS, NASA ECHO, China Academy of Optic-Electronic (AOE), the Brazil National Institute for Space Research (INPE), the National Oceanographic Data Center (NODC) Group for High Resolution Sea Surface Temperature (GHRSST), the Japan Aerospace Exploration Agency (JAXA), the Canada Centre for Remote Sensing (CCRS), the U.S. Geological Survey (USGS) Landsat, and GCMD.

Table 1 Comparison of catalogues for their interface, metadata information model and searching level

Catalogue data provider	Query interface	Metadata information model	Granularity of search	
			Keyword ->dataset	Dataset ->granule
NOAA CLASS	Web GUI	NOAA-defined model	No	Yes
NASA ECHO	ECHO API	EOSDIS Core System (ECS) science data model	Yes	Yes
AOE	CSW	ISO 19115 XML model	No	Yes
INPE	Web API	INPE-defined XML model	No	Yes
NOAA GHRSST	CSW	ISO 19115 XML model	No	Yes
JAXA	CSW	ISO 19115 XML model	No	Yes
CCRS	CSW	ISO 19115 XML model	No	Yes
USGS	Web API	USGS-defined XML model	No	Yes
GCMD	CSW	ISO 19115 XML model	Yes	No

As demonstrated in table 1, some of the catalogues offer standard OGC CSW query interfaces and adopt an ISO standard metadata information model while others offer self-defined APIs and metadata information models. When users want to search EO data from multiple catalogues, they need to do extra research in understanding the query interfaces and metadata information model. Furthermore, they need to know to which catalogue their required data belong. Most of the catalogues discussed above do not support two levels of searching granularities. Therefore, to hide users from the disparity of heterogeneous catalogues for EO data search, a framework that integrates distributed catalogues and supports searching abilities at both dataset level and granule level is developed.

2.2 Integrated Catalogue for Distributed EO Data Search

The large volume of geospatial data resources, the availability of on-line open data servers, and the existence of interoperability standards and technology form a common foundation for the sharing and interoperability of geospatial data. Based on this, many value-added services and applications of national and international importance can be built (Di and Ramapriyan, 2010). Different geospatial data are collected by different space agencies and they use different catalogues to manage data. The requests from users and clients to access data may require information from more than one source requiring users to achieve a good understanding of metadata information models and catalog interface protocols for each catalogue. It is important to find a mechanism to integrate the catalogues to facilitate data searching, especially for EO data. Efforts have been made towards the integration in SDIs (Manso et al., 2009; Vaccari et al., 2009). From an information technology point of view, the challenge is to implement interoperable discovery services for data and processing resources that are collected and managed using multidisciplinary standards and tools (Nativi and Bigagli, 2009). Various models and frameworks have been developed to solve the challenges in the concept of integration. By reviewing these models and frameworks, I have identified current strategies and deficiencies in the integration of catalogues for distributed EO data search. In this section, I first review several models, frameworks, and systems for integration among discovery and access services for EO data then introduce the mediator-wrapper structure in developing the framework for integrating catalogues. At the end of this section, current keywords search technologies will be introduced.

2.2.1 Federation Service and Interoperability

Heterogeneity problems need to be solved. Visser et al. (2002) addressed these problems on three levels: the syntactic, structure, and semantic level. They point out that it is crucial to note that the problems of interoperable GIS can be solved only if solutions (modules) on all three levels of integration are working together. It is not possible to solve the heterogeneity problems separately (Visser et al., 2002). In Shvaiko et al. (2010), the authors describe work on the implementation of a semantic geo-catalogue for a regional SDI which focuses on a discovery service implemented by means of CSW. The overall system architecture of the geo-catalogue implementation follows the standard three-tier paradigm with front-end, business logic and back-end layers, after which the semantic query processing methods are added to extend the GeoNetwork catalogue search function. This search illustrates how one CSW-based catalogue can be integrated into SDI with metadata management, user/group management and system configuration. This method is similar to the concept of standardizing catalogues before integrating them into the framework, however, the issues of interoperability for distributed EO data search are not solved.

Two types of SDI interoperability issues are addressed by Vaccari et al. (2009): Geo-data interoperability issues and the Geo-service interoperability issue. Since each geo-data producer adopts internal rules in order to manage its geographical datasets, and moreover, as geographical datasets have specific properties different from other types of data, heterogeneity at the data level arises for the following six reasons: different syntax, different structure, different semantics, implicit linking, massive datasets, and multiple versions. The general issues for service integration are (Vaccari et al., 2009): geo-service discovery, geo-service integration, maps as implicit interfaces, geometry based information, and specific topological operations. After addressing each of these issues, the authors discuss possible solutions to achieve Geo-data interoperability and Geo-service interoperability. For Geo-data interoperability, they examined several related works that use ontology to reduce heterogeneity. For the Geo-service interoperability, they state that OGC specifications and SOA technological solutions provide syntactic interoperability and cataloguing of geographic information (Vaccari et al., 2009) based on OGC CSW.

After the description of all the issues and possible solutions in Geo-data interoperability and Geo-service interoperability, the authors introduce an application scenario to be used as a motivating example for the description of their approach on a geo-service semantic integration. In Vaccari's research, the structure preserving semantic matching (SPSM) approach is used to support ontology matching between different service providers. In conclusion, the author states his research is a focused investigation on a semantic interoperability approach to integrate geo-services. A major drawback in this research is that the peer-to-peer (P2P) approach requires a peer to know which interaction model it wants to execute and with which peers it will be interacting. Thus extra work is still needed to figure out the relationship between peers and models.

Bai et al. (2007) proposed to build a federation service to fulfill distributed and integrated metadata discovery and point out that there are four main challenges in building the catalog federation: protocol adaptation, query dispatch, query criteria translation and query result integration. Three distinct geospatial catalogue services: the NASA ECHO, the George Mason University (GMU) CSW and the U.S. Department of Energy (DOE) Earth System Grid (ESG) Simulation Data Catalogue are investigated. Bai also analyzes the metadata conceptual model, query language and communication protocol for each of the three catalogues and proposes strategies for protocol adaptation, query dispatch, query translation and query result integration. This federated service follows mediation-wrapper architecture. As for protocol adaptation, a wrapper is developed for NASA ECHO since the other two catalogue services support OGC CSW. Three patterns for dispatching queries are defined in Bai et al.'s (2007) research: opaque, translucent, and transparent. Catalogue Federation Service (CFS) adopts the opaque pattern and dispatches queries depending on criteria cited in the user's query. The strategy for query translation involves transformation of four layers: the metadata term, query criterion, query criteria and query payload (Bai et al., 2007). There are three patterns for integrating query results: opaque, translucent, and transparent. CFS chooses the opaque pattern to offer a unique information model for the results. Given those strategies, the author describes the GMU CFS and uses ECHO as an example to discuss CFS. Bai et al. (2007) provide a federation service for geospatial catalogues through a case study of building integration over three legacy catalogue services. CFS always queries first against the GMU CSW, then against the OGC CSW for ECHO, and finally the ESG catalogue. Essentially, the query is in a sequential order process. This is only practical when there are only a few federated catalogues. However, when the integrated catalogue holds a large number of distributed catalogues, it is not efficient to search against every record of the embedded catalogues. How to efficiently dispatch queries to corresponding catalogues has been a major issue in developing an integrated framework for distributed catalogue search of EO data. In my research, a directory for datasets from embedded catalogues is involved in resolving this issue. This will be discussed in detail in chapter 4.

An advanced catalogue service featuring additional functionalities and a federated, extensible data model is developed by Nativi and Bigagli (2009). They point out there are several shortcomings in current catalogue specifications: data model heterogeneity results in interoperability mismatch between the different “sub-types” of catalogs, in spite of their claimed conformance to the same abstract specifications; a distributed search is not addressed by the specification, leaving space for arbitrary behavior that undermines interoperability; the specification’s synchronous mess exchange pattern hinders usability, since the user must wait until his/her request has been fully processed. In order to solve these shortcomings, a catalogue is built by extending OGC CSW with three additional functionalities: messaging, distribution, and mediation. Messaging is used to provide asynchronous searching by incremental query, query feedback, and query interruption. Distribution is for request routing and response aggregation. The main task of an ideal "mediation component" is to integrate a heterogeneous server by adapting its technological (protocol), logical (data model) and semantic (concepts and behavior) model (Nativi and Bigagli, 2009). In this research, mediation functionality is performed by specific "Accessor" components and the distribution functionality is performed by a distributor. A distributor and several "Accessors" may be chained to obtain a catalogue solution providing discovery services for heterogeneous geospatial resources (Nativi and Bigagli, 2009). Several strategies are described in which a CSW ebRIM data model extension is used to unify local models for data providers. These strategies federate a new data model and support EO data access. In summary, Nativi and Bigagli's research extends the SOA approach and present catalog standard specification to support discovery, mediation and access for EO data. However, it does not solve the dispatching issues for the mediator when a large number of EO catalogues are integrated and serve as federation members.

Although there is vast literature available on interoperability models and their respective interoperability levels, limited research has been carried out on the development of interoperability models for the implementation of Spatial Data Infrastructures (Manso et al., 2009). An integrated interoperability model is developed and consists of seven interoperability levels: technical, syntactic, semantic, pragmatic, dynamic, conceptual and organizational by Manso et al. (2009).With this model, elements of spatial metadata are classified into one or more interoperability levels. Thus the important role of metadata elements in the formalization of interoperability models for the implementation of Spatial Data Infrastructures is demonstrated.

Andrade et al. (2011) develop a distributed architecture, based on a federation of SDIs which interact among themselves, using query propagation to facilitate data discovery and sharing. Different SDIs are organized in a hierarchical architecture. An advantage of this architecture is that when the integration SDI wants to propagate a query, it simply sends the query to all local SDIs of the federation. On the other hand, when a local infrastructure wants to propagate a query, this query is sent to the integration SDI, which is responsible for routing it to the other infrastructures in the federation. This improves the capability of discovering resources. By applying OGC and ISO standards, the heterogeneity problems among the interconnected infrastructures are solved. A query processing service compliant with the OGC CSW getRecords method is developed for query matchmaking, query mapping and forwarding queries to other SDIs in the federation. Two major deficiencies of this architecture are: 1) the semantic conflicts in resource discovery and heterogeneity problems in accessing resources still exist; 2) a standard method to describe SDIs that constitute the architecture is lacking, which means a unified metadata information model is needed to enable an efficient search and better routing algorithms.

Wang (2012) does research on a web data integration framework based on cloud computing which presents several means of data integration. The researcher introduces the technology of cloud computing to provide all kinds of clients with different services such as software, hardware, and data by web server clusters. At present, the relatively mature data integration methods are federated database-based middleware models and data warehouse. With the increase of integrated systems, the cost will be doubled (Wang, 2012). Data warehousing is mentioned as another solution, in which data sources convert to a unified model to store the integrated data. However, this solution is difficult to realize, since there is no fixed data model for heterogeneous Web data. Using teaching resources from different universities as an example, the research developed an integrated framework. Three types of methods for data integration are proposed: the first is to link all university web sites in a cloud system; the second is to integrate data inside a cloud system which links several representative universities; the third is to provide integrated information referring to external network resources. With the first method, the cloud system does not play a role since it is only providing a user with all the information. With the second method, since only several representative universities are chosen, the searching results are incomplete and may not be the best. The third method is relatively good but integration efficiency is lower than the second option. The researcher developed the integrated framework based on two integration mechanisms: virtual view and data warehousing. In the virtual view method, the data is not stored locally, still remaining in their originating system. In the data warehouse method, the shared information extracted from different data sources is stored in a central database before a user puts forward an inquiry request (Wang 2012). The first method provides the user with the latest datum for data which are frequently updated, while the latter method offers a fast inquiry interface for relatively stable data. Since the purpose of Wang's research is for course teaching, the strategy for data source selection is simply based on the rank of the courses and disciplines as determined by the colleges or universities. This framework is efficient and flexible for integrating teaching resources. However, at least two aspects are not applicable for the integration of EO data: one is that due to the large volume of EO data collected every day, it is not possible or too costly to build a centralized data warehouse for storing all the relatively stable data; the other is the data selection strategy since it is difficult to set a rank to spatial information.

As a summary of the efforts made in the research in this section, Manso et al. (2009) developed an interoperability model for the implementation of SDI which consisted of seven interoperability levels to demonstrate the importance of metadata elements in the field of interoperability. All other research mentioned above contributes to the development of an integrated framework or model for distributed data discovery and access to some extent. After carefully reviewing the strategies, architectures, and technologies in the above research, I learned the current approaches for integrating service or data to enhance interoperability of data discovery and access. The pros and cons of the above models and frameworks are summarized in table 2.

Table 2 Summary of pros and cons of current models/frameworks to integrate data discovery and access

Model/Framework	Pros	Cons
Shvaiko et al. (2010)	One CSW-based catalogue being integrated into SDI with the realization of metadata management, user/group management and system configuration.	Interoperability issues for distributed EO data search are not solved.
Vaccari et al. (2009)	A semantic interoperability approach in order to integrate geo-services.	The adopted P2P approach requires a peer should know which interaction model it wants to execute and with which peers it will be interacting. Thus extra work is needed to figure out the relationship between peers and models.
Bai et al. (2007)	Built a federation service to fulfill distributed and integrated metadata discovery; standardize catalogue with OGC CSW.	Queries of catalogues are in a sequential order against every embedded catalogue. Not efficient to integrate large number of catalogues.
Nativi and Bigagli (2009)	Extended the SOA approach and the present catalog standard specification to support discovery, mediation and access for EO data.	The dispatching issues for the mediator when integrating a large number of EO catalogues are not solved.
Andrade et al. (2011)	Developed a distributed architecture based on a federation of SDIs which interact among themselves, using query propagation to facilitate data discovery and sharing; classification of SDIs.	Heterogeneity problems in accessing resources are not solved; lack a standard method to describe SDIs
Wang (2012)	Efficient and flexible for integrating teaching resources based on cloud computing architecture.	Not applicable for integration of EO data as for building data warehouse or ranking data sources.

Therefore, the current framework needs to be improved to support a distributed catalogue search for EO data. As mentioned in 2.1.2, metadata catalogues are developed with different query interfaces and metadata information. Several architectures are adopted in the above efforts, from which the mediator-wrapper architecture will be used in the ICF to standardize heterogeneous catalogues before they are integrated into the target framework.

2.2.2 Mediator-wrapper Architecture

The mediator-wrapper architecture was proposed by Wiederhold (1992). Interoperation with the diversity of available sources requires a variety of functions. In a 2-layer client-server architecture, all functions have to be assigned either to the server or to the client modules; with a third, intermediate layer, which mediates between the users and the sources. Many functions, in particularly those that add value and require maintenance to retain value, can be assigned there (Wiederhold, 1999). Mediator-wrapper architecture allows multiple types of models, as required, to be integrated. It is widely used in integrated access to multiple data and information sources. Nowadays, well-established architectures and standard technologies are available to address and implement data interoperability. In particular, mediation provides a valuable and flexible approach for harmonizing data (Bigagli et al., 2005). Semantic mediation can play an important role in this context in that information may not be processed from only one data source, but from combinations of multiple heterogeneous data sources with different representations of a common domain (Suwanmanee et al., 2005). Various data integration systems are built based on mediator-wrapper architecture with different technologies and purposes (Langegger et al., 2008; Beneventano et al., 2011 and Bakhtouchi et al., 2012). On a broader basis, interoperability may be seen as the capacity to move information across the boundaries between the source and the destination of such information (Bigagli et al., 2005). Data integration problems are often the result of inputs from a set of distributed, heterogeneous, autonomous, and evolving data sources. Further, each data source has its own scheme and population. The goal is to provide a unified description of source schemes using an integrated schema and mapping rules allowing access to data sources (Bakhtouchi et al., 2012). In my research, to integrate distributed catalogues for EO data search, a set of mapping rules will be followed to accommodate heterogeneous catalogues with OGC CSW. These rules will be introduced in Chapter 3.

Mediator-wrapper architecture has several advantages: first, the specialized components of the architecture allow the concerns of different kinds of users to be handled separately; second, mediators typically specialize in a related set of component databases with "similar" data, and thus export schemas and semantics related to a particular domain (Özsu and Valduriez, 2011). In this research, the mediator-wrapper architecture will be applied to develop integrated catalogue to support two levels of searching granularities, especially in the granule level search. Wrappers are developed for each catalogue to offer standard query interfaces and searching results. Based on queries from clients/users and the analysis from a dataset level search, the mediator dispatches queries to corresponding catalogues. Keyword search functionality will be exploited in a dataset level search to offer users an opaque method to search EO data.

2.2.3 Keyword Search

Keyword search is the most popular information discovery method because the user does not need to know either a query language or the underlying structure of the data (Hristidis and Papakonstantinou, 2002). While keyword searching is widely used to search documents on the Web, querying of databases currently relies on complex query languages that are inappropriate for casual end-users since they are complex and hard to learn (Hulgeri et al., 2001). Keyword queries offer a convenient alternative to traditional SQL in querying relational databases with large, often unknown, schemas and instances (Bergamaschi et al., 2011). A lot of research has occurred in keyword search of relational databases (e.g. Aditya et al., 2002; Agrawal and Das, 2002, Luo et al., 2007, Tata and Lohman, 2008). Generally, these works consider a database as a network of interconnected tuples. Each database is detected as containing the keywords in the query, and generates connected components based on how these tuples are associated. Connected tuples are returned as an answer to the query (Bergamaschi et al., 2011). While in the context of web services matching and ranking, semantic understanding of Web services may provide added value by identifying new possibilities for composition of services (Segev and Toch, 2009). Web Ontology Language (OWL) provides a mechanism to enable the use of semantics (Yue et al., 2006). By extending the ebRIM elements, the semantics defined in OWL can be organized in CSW. Although the concept of keyword search has been adopted in geospatial catalogues in recent years, little has been done in enabling keyword search functions at the dataset level, as Table 1 shows.

Having reviewed all the efforts made in the development of integrated catalogue for distributed EO data search, this dissertation aims to address several challenges in distributed catalogue search and enhance the interoperability of EO data discovery, search and access. Three main research solutions are provided: 1) designing an extensible framework which can be used in extending the scope and granularity of EO data search for integrating the distributed catalogues; 2) offering methods to enable keyword search for dataset metadata; 3) developing efficient algorithms for standardizing heterogeneous catalogues and dispatching queries to them; Finally, a unified interface is developed that users either with or without background knowledge of EO data catalogues can use to retrieve the EO data they demand.

CHAPTER THREE INTEROPERABLE SEARCH INTERFACES AND INFORMATION MODELS

Several standards will be applied and implemented towards the objective of providing users with a standards-compliant and universal interface to discover data from multiple heterogeneous catalogues. The OGC specifications and the ISO standards form the basis of most existing catalogue interface implementations (Senkler et al., 2004). The OGC CSW specification provides an implementation-specific model to define a set of interfaces between clients and catalogues services and is applied in the integration of distributed catalogues to develop an interoperable search interface. To solve the problem that heterogeneous information models are utilized by different data providers, the ISO 19115 and Dublin Core specifications were adopted. In this chapter, I first make an overall introduction to the OGC CSW and present several types of current search interfaces. At the end of this chapter, several metadata information models supported by OGC CSW are described.

3.1 Search interfaces for EO Data Discovery

Many search interfaces have been developed in the field of EO data discovery to facilitate the interaction between users and the underlying information. To improve the usability and functionality of these interfaces, people utilize different methods to develop user interfaces, depending on their distinct goals (e.g. Marlin et al., 1998; Myers and Rausch, 2000; Carvalho et al., 2012). Distributed catalogues with diverse search interfaces for discovery and access of EO data have been developed by various spatial agencies. In this research, those interfaces are categorized into three types: GUI, API, and CSW interfaces, according to the extent of the underlying information exposed to users, as illustrated in Figure 2.

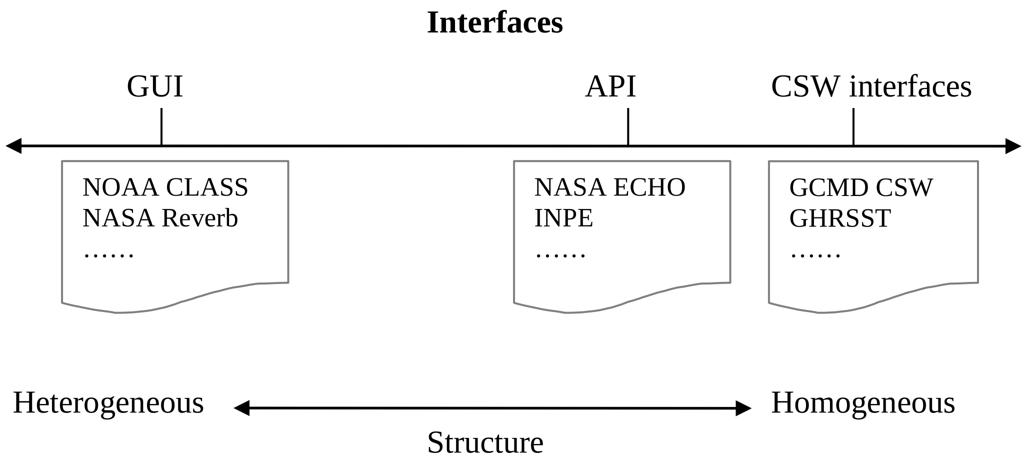


Figure 2 Types of catalogue interfaces for EO data discovery

GUIs require minimum EO domain knowledge from users to search over catalogues. However, they are the most heterogeneous thus the most difficult to be integrated to support distributed search. Examples of GUIs include the NOAA CLASS and NASA Reverb system. APIs, such as NASA ECHO API and INPE-Web API, suit for developer to interact with underlying catalogues. They are usually developed and exposed by data providers. Base on APIs, various GUIs can be developed to facilitate users with less domain knowledge since diverse structures and different programming language are adopted in APIs. CSW interface is standard-compliant approach to offer a unified interface for distributed catalogues. It supports the Common Query Language (CQL) developed by OGC to achieve high degree of flexibility of query. Increasing number of catalogues such as GCMD CSW and GHRSST CSW, have been developed with CSW interfaces to enhance interoperability of EO data sharing. Catalogues developed based on OGC CSW specification can be easily integrated to extend the scope of EO data search. In the following section, the overall concepts of OGC CSW specification are introduced.

3.2 OGC Catalogue Service

Catalogue services support the ability to publish and search collections of descriptive information, or metadata, for data, services, and other information. They also support various query languages to find and return results using well-known content models, or metadata schemas, and encodings (Nebert et al., 2007). The OGC catalogue services implementation specification provides an implementation-specific model to define a set of interfaces between clients and catalogues services. In general, it specifies the interfaces, bindings, and a framework for defining application profiles required to publish and access digital catalogues of metadata for geospatial data, services, and related resource information (Nebert et al., 2007). In the following part, the query language support and core catalogue schema for the abstract information will be discussed, followed by the general catalogue interface model of OGC CSW. I then describe the three protocol bindings: Z39.50 protocol binding, Common Object Request Broker

Architecture (CORBA)/Internet Inter-ORB Protocol (IIOP) protocol binding and Hypertext Transfer Protocol (HTTP) protocol binding, supported by this specification. The HTTP protocol binding, which is adopted in this research, will be described in detail.

3.2.1 Query Language Support

In order to achieve high degree of query flexibility, the OGC general catalogue model provides a minimum set of data types and query operations. To achieve the interoperability goal, this minimal abstract query language shall be supported by all OpenGIS-compliant catalogue services. During the development of OGC_Common Query Language, four assumptions were made (Nebert et al., 2007):

1) The query will have syntax similar to the SQL "Where Clause."
2) The expressiveness of the query will not require extensions to various current query systems used in geospatial catalogue queries other than the implementation of some geo operators.
3) The query language is extensible.
4) OGC_Common supports both tight and loose queries.

There are more than one query languages or messaging mechanisms identified within OGC specifications. To avoid potential confusion it may cause, it is important to identify the name and version of required query language in the application profile. Application profiles are defined as schemas which consist of data elements drawn from one or more namespaces, combined together by implementers, and optimized for a particular local application (Heery and Patel, 2000). Besides the Core, the OGC CSW implementation specification also defined the ISO metadata application profile, the OWL application profile, the EO application profile, and the FGDC Content Standard for Digital Geospatial Metadata (CSDGM) application profile. Each application profile defines a set of queryable properties and returnable properties.

3.2.2 Core Catalogue Schema

For the purposes of information interchange within an information community, a metadata schema may be defined to provide a common vocabulary which supports search, retrieval, display, and association between the description and the object being described (Nebert et al., 2007). The adoption of metadata schema is not required by the specification. However it will certainly enhance query interoperability among catalogues, especially those using the same protocol binding. OGC CSW defines a set of core queryable properties and core returnable properties based on the nomenclature and syntax of Dublin Core Metadata. Specific information content, syntax and semantics need to be addressed in each application profile. Detail information about OGC CSW Core profile is discussed in section 3.3.2.

3.2.3 General Catalogue Interface Model

The General Catalogue Interface Model provides a set of abstract service interfaces that support the discovery, access, maintenance and organization of catalogues for geospatial information (Nebert et al., 2007). In the implementation of the general catalogue interface model, protocol binding shall be included. According to the catalogue service specification, three protocol bindings are supported, including CORBA, Z39.50, and HTTP protocol bindings. In this research, the HTTP protocol binding was adopted for the communication of the integrated framework with underlying distributed catalogues services.

The Catalogue Service may leverage one of three sources: a metadata repository local to the catalogue service, a resource service, or another Catalogue Service, to respond to a Catalogue Service request (Nebert et al., 2007). In this research, to support distributed catalogues search and two levels of searching granularities, the aimed framework leveraged all of the three sources. As for the NOAA CLASS catalogue, it used a resource service to respond to the catalogue service request. While for the NASA ECHO catalogue it used a local metadata repository to respond to clients' requests. Figure 3 illustrates the interfaces with catalogue services compliant to OGC CSW standards:

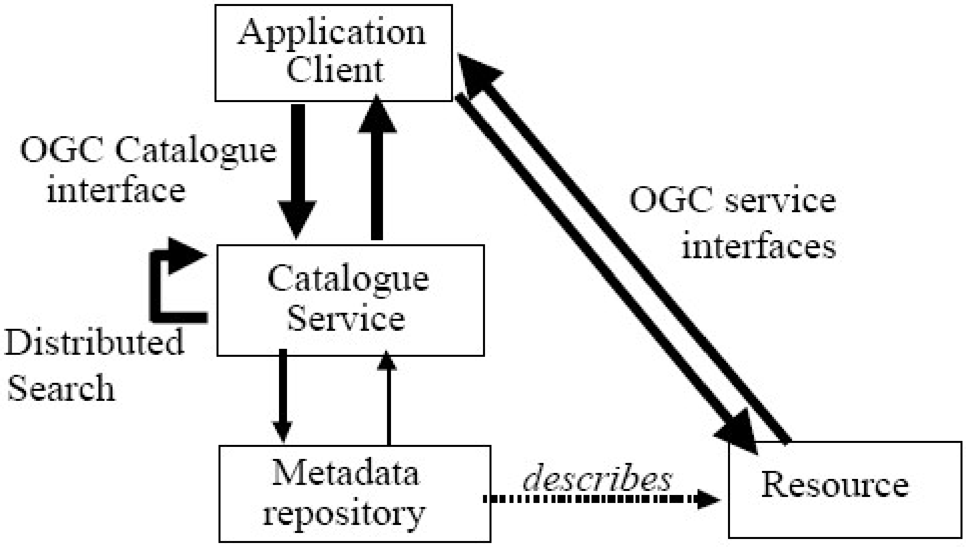


Figure 3 Interfaces with OGC catalogue services (Nebert et al., 2007)

In the concept of catalogue interface model, we defined the catalogue service interfaces as a class named Catalogue Service Class. It is associated with five other classes: the OGC_Service class, Discovery class, Session Class, Manager Class, and the Brokered Access class. Each class has different abilities by providing a set of corresponding operation requests. A tree view of the relationship between catalogue service classes and operations is given as Figure 4.

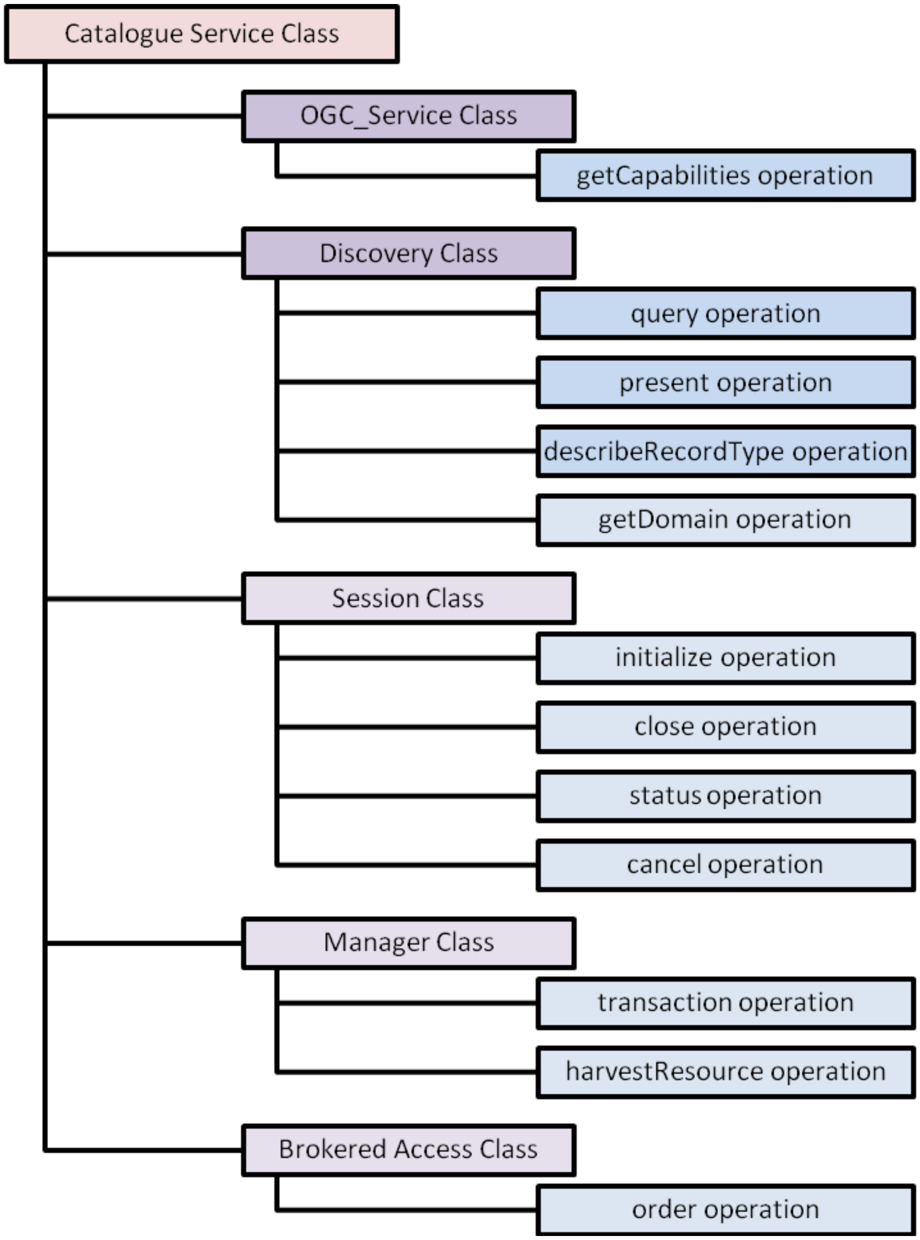


Figure 4 Relationship between catalogue service classes and operations.

As shown in Figure 4, the catalogue service class functions as the foundation of all the other classes. These interfaces and operation can be further extended and specialized by specific application profiles. The names of the classes and operations may be changed according to different protocol binds as well. However, the semantics and granularity of interaction of these interfaces and operations should remain the same as in the concept of catalogue interface model.

The OGC_Service class provides the GetCapabilities operation for users or clients to retrieve service metadata for the whole catalogues services. This operation is mandatory for the catalogue service class.

The Discover class provides four operations: query, present, DescribeRecordType and GetDomain for users or clients to discover geospatial information provide by catalogues. This class is required for the catalogue service class. However, not all of its operations are mandatory. If the supported protocol binding of the catalogue service class is HTTP binding, the query, present, and the DescribeRecordType operations are mandatory operations while the GetDomain operation is optional.

The optional Session class allows the use of interactive sessions between a client and a server, by providing four stateful operations: “initiate”, “close”, “status”, and “cancel” (Nebert et al., 2007).

The Manager class and Brokered Access class are both optional. The Manager class allows clients to insert, update and/or delete catalogue content by providing a transaction operation and a harvestResource operation. The Broker Access class provides an order operation, which allows clients to place an order for an identified registered resource when that resource is a data product that is not directly accessible to clients (Nebert et al., 2007).

In summary, the OGC_Service class and the Discovery class are required for the catalogue service class and the Session class, Manager class, and Brokered Access class are optional. Among the 12 operations, the GetCapabilities, query, present, and DescribeRecordType operations are required for the catalogue service class to support the HTTP protocol binding. In the following part, three recognized protocol bindings: Z39.5, CORBA/IIOP and HTTP protocol, supported by OpenGIS catalogue service specification will be described separately with the main focus on the HTTP protocol binding.

3.2.4 Protocol Binding

The Z39.50 protocol binding uses a message-based client server architecture implemented using the Z39.50 Application Service Definition and Protocol Specification [ISO 23950]. At a minimum, Catalogue Services implemented using the Z39.50 protocol binding shall support the Discovery and Session operation groupings (Nebert et al., 2007). The intention of the CORBA protocol binding is to follow the General Model closely. The CORBA protocol binding is described in IDL (interface definition language) of the object management group (Nebert et al., 2007). Interfaces of CORBA protocol binding follow the General Model as closely as possible. The discovery and session interfaces shall always be supported by catalogues services implemented based on CORBA. Separate services of the general model are all inherited by the central interface CatalogServices of CORBA protocol binding.

HTTP is an application-level protocol for distributed, collaborative, hypermedia information systems. It is a generic, stateless, object-oriented protocol which can be used for many tasks (Fielding et al., 1997). The request and response messages of HTTP can be either encoded in keyword-value pairs (KVP) format within a request uniform resource identifier (URI) or in Extensible Markup Language (XML) format. Requests can also be embedded in messaging frameworks. The GET and POST methods for HTTP requests were implemented in this research according to the corresponding operations.

All implementations of the HTTP protocol, and application profiles derived from the CSW protocol binding, shall support the response schema which is an XML realization of the core metadata properties. In all cases, elements of the underlying information model shall be mapped to the core metadata properties (Nebert et al., 2007). The full set of core properties are concretely materialized by the csw:Record element. Two additional elements, csw:BriefRecord and csw:SummaryRecord materialize the brief and summary views of the full set of core properties (Nebert et al., 2007). In this research, for the development of the integrated model, all of the three types of elements were used to describe the metadata information model. Full record, summary record and brief record represents the full view, summary view and brief view of the returned results. As mentioned above, the GET and POST methods of HTTP request methods were implemented in this research according to the corresponding operations. Table 3 shows the HTTP method binding and data encoding for mandatory operations implemented in this research. All the requests listed in the table, the specification supports both HTTP GET and HTTP POST method. For convenience, the GET method was adopted for the GetCapabilities and DescribeRecord requests while the POST method was used for the other two mandatory requests: GetRecords, and GetRecordById.

Table 3 HTTP method bindings and data encoding for mandatory operations of CSW

Request	HTTP Method binding(s)	Data encoding(s)
GetCapabilities	GET	KVP
DescribeRecord	GET	KVP
GetRecords	POST	XML
GetRecordById	POST	XML

These four types of requests in HTTP binding belong to the OGC_Service class and the Discovery class. Table 4 maps the general model operation to the catalogue service for the web operations. Here we only describe the four mandatory catalogue operations for the HTTP protocol binding. Except GetCapabilities, for all the other three operations, there are three common operation request parameters: request, service, and version, which need to be encoded either in KVP or XML format separately. The "request" parameter is to specify the type of the request sent to CSW. Its value is the name of the corresponding CSW operation. The "service" parameter is a fixed string with the value of "CSW". The "version" parameter indicates the associated CSW request version. In this research, since the developed CSW is based on OpenGIS Catalog services implementation specification version 2.0.2, this parameter is also a fixed string with value of "2.0.2".

Table 4 Mapping of mandatory operations: general model to CSW

General Model Operation	**CSW Operation**
OGC_Service.getCapabilities	OGC_Service.GetCapabilities
Discovery.query	CSW-Discovery.GetRecords
Discovery.present	CSW-Discovery.GetRecordById
Discovery.describeRecordType	CSW-Discovery.DescribeRecord

The four mandatory operations are: GetCapabilities, DescribeRecord, GetRecords, and GetRecordById. The GetCapabilities operation allows CSW clients to retrieve service metadata regarding the whole catalogue service from the server. The response to this operation is an XML encoded capabilities document with four sections to describe the catalogue service. The DescribeRecord operation allows CSW clients to retrieve information about the elements of the information model supported by the catalogue service. The response to this operation usually contains XML schema to define the record (Nebert et al., 2007). The GetRecords operation allows CSW clients to discover resources supported by the target catalogue service. It does the search for and presentation of resources. Except for the three common mandatory parameters of an operation request, the GetRecords request has another mandatory parameter named typeNames which is a list of one or more names of entities that are queryable in the catalogue's information model. In an XML encoded request, this parameter is not required. The response to this operation is an XML encoded document with the search results. The GetRecordById operation allows CSW clients to retrieve a catalogue record with its identifier. As indicated by the operation name, Id is one of the mandatory request parameter that the GetRecordById request must offer. Response to this operation is an XML encode document with the list of requested records.

For the exception reporting, according to the OpenGIS web service common implementation specification, an XML document indicating that an error has occurred will be generated and sent back to the user/client.

3.3 Metadata Models for OGC Catalogue Service

There are many different metadata formats, which use different ways to represent information about data. Many information models are developed as constraints for controlling metadata. In many circumstances, even if controlled metadata are used, each archive could employ its own semantics for these fields (Liu et al., 2006). Thus the lack of interoperability is one of the major issues in metadata sharing. One straightforward solution is to apply commonly recognized standards to metadata models. OGC CSW supports Core profile metadata, which is extended from Dublin Core and several other metadata models, such as OWL application profile, EO application profile, FGDC CSDGM application profile, defined in ISO metadata application profile. This section will focus on introduction for two types of metadata models, Dublin Core and ISO 19115 metadata, supported by OGC CSW in detail by identifying its queryable properties and returnable properties accordingly.

3.3.1 Dublin Core

The common CSW record syntax is an XML-based encoding of Dublin Core metadata terms. It represents a concrete realization of the core metadata properties abstractly specified in catalogue abstract information model (Nebert et al., 2007). With a set of core queryable properties, queries can be executed against any OGC catalogue service regardless of the information model it applied. The core queryable properties should be realized in core queryable schemas according to different protocol bindings. As for HTTP protocol binding, a set of core queryable and returnable properties is shown in table 5.

Table 5 Dublin Core queryable and returnable properties mapped to XML elements

Name	XML element name	Common queryable	Common returnable
title	dc:title	Yes	Yes
creator	dc:creator		Yes
subject	dc:subject	Yes	Yes
description	dc:abstract	Yes	Yes
publisher	dc:publisher		Yes
contributor	dc:contributor		Yes
date	dc:modified	Yes	Yes
type	dc:type	Yes	Yes
format	dc:format	Yes	Yes
identifier	dc:identifier	Yes	Yes
source	dc:source	Yes	Yes
language	dc:language		Yes
relation	dc:relation	Yes	Yes
coverage	ows:BoundingBox	Yes	Yes
rights	dc:rights		Yes

Besides the queryable properties shown in the table, there is another queryable property “AnyText” which targets for full-text search of character data types included in the common queryable elements. They should all be included in a binding protocol regardless of the underlying information model. A NULL value should be assigned to a core property if the underlying catalogue information model does not support this property. However, to compose a searching result, not all the commonly returnable properties in table 5 need to be populated. It varies depending on different protocol bindings and the record type returned to users. With the HTTP protocol binding, three types of records are supported, which are: csw:Record element, csw:BriefRecord element, and csw:SummaryRecord element. The sets of Dublin core elements materialized by these record elements are different. The Dublin core elements dc:identifier and dc:title are mandatory for all the three types of records defined in OGC CSW though they are optional in the Dublin core schema.

3.3.2 ISO 19115

The ISO metadata standards, specifically those in the ISO 19000 series, are currently emerging as the primary means to represent metadata associated with geographic information in Earth science data products (Hua and Weiss, 2011). The OGC ISO metadata application profile specifies the interfaces, bindings, and encodings required to publish and access digital catalogues of metadata for geospatial data, services, and applications that comply with the given profile. The intention was to implement a generally understood information model based on standard metadata with only a few relationships among the catalogue items (Voges and Senkle, 2007).

The capabilities classes defined in the ISO profile for the catalogue services are the OGC_Service class and the Discovery class. The OGC_Service class should provide operation for requesting service metadata. The service metadata descriptions should consist of identification information inherited from ISO19115:MD_Identification, metadata describing the service instance and optional metadata or reference. The ISO catalogue information model for CSW provides a standard method to describe and encode information resources. Dataset, dataset collection, and application types of information resources are described with ISO19115 while information resources of the service type are described with ISO19115 or ISO19119. Extensions to ISO specifications can also be made to support specific description of resources. Table 6 lists core queryable properties and returnable properties to a common XML Record format which means the properties are mapped to the information model based on the ISO profile (Voges and Senkle, 2007).

Table 6 Core returnable properties mapping to ISO information model (*:queryable and returnable)

Dublin Core Metadata name	**Returnable property mapping to ISO information model**	**Name used in OGC queryables**
title*	MD_Metadata.identificationInfo.AbstractMD_Identification.citation.CI_Citation.title	Title
creator	MD_Metadata.identificationInfo.AbstractMD_Identification.pointOfContact.CI_ResponsibleParty.organisationName[role.CI_RoleCode@codeListValue=’originator’]	
subject*	MD_Metadata.identificationInfo.AbstractMD_Identification.descriptiveKeywords.MD_Keywords.keyword plus MD_Metadata.identificationInfo.MD_DataIdentification.topicCategory	Subject
description*	MD_Metadata.identificationInfo.AbstractMD_Identification.abstract	Abstract
publisher	MD_Metadata.identificationInfo.AbstractMD_Identification.pointOfContact.CI_ResponsibleParty.organisationName[role.CI_RoleCode@codeListValue=’publisher’]	
contributor	MD_Metadata.identificationInfo.AbstractMD_Identification.pointOfContact.CI_ResponsibleParty.organisationName[role.CI_RoleCode@codeListValue=’author’]	
date*	MD_Metadata.dateStamp.Date	Modified
type*	MD_Metadata.hierarchyLevel.MD_ScopeCode@codeListValue	Type
format*	MD_Metadata.distributionInfo.MD_Distribution.distributionFormat.MD_Format.name	Format
identifier*	MD_Metadata.fileIdentifier	Identifier
source*	not supported	Source
language	MD_Metadata.language	
relation*	MD_Metadata.identificationInfo.AbstractMD_Identification.aggregationInfo	Association
coverage*	MD_Metadata.identificationInfo.MD_DataIdentification.extent.EX_Extent.geographicElement.EX_GeographicBoundingBox.(westBoundLongitude, southBoundLatitude, eastBoundLongitude, northBoundLatitude)	BoundingBox
rights	MD_Metadata.identificationInfo.AbstractMD_Identification.resourceConstraints.MD_LegalConstraints.accessConstraints@codeListValue	

Besides the queryable properties shown in Table 6, the common queryable property of “AnyText” is also supported in ISO profile as the whole resource text. ISO profile supports a set of additional search properties and additional returnable properties such as alternateTitle, Language, SpatialResolution, TemporalExtent, DistanceValue, DistanceUOM, and so on. As for the data bindings, currently XML is the only data binding supported. The entire information object that is to be managed by a catalogue service complying with this profile must apply this presentation form (Voges and Senkle, 2007). A set of rules apply to XML encoding are provided in ISO profile for defining dataset, dataset collection, service and application.

The ISO application profile also supports three types of result sets: brief, summary and full. In order to get a valid form of those resultsets, the corresponding sets of metadata elements must have valid returnable properties defined. For example, BoundingBox, Identifier, GraphicOverview, ServiceType, ServiceTypeVersion, Title, and Type are defined as valid returnable properties for a brief record. The collaboration for the ISO based catalogue with catalogs based on other CSW profiles such as OWL or EO profile is achieved by using the CSW common profile including the core queryable properties and common record schema applied in it. The ISO based profile imports the HTTP protocol from the OpenGIS Catalogue Services Specification. Table 7 shows the mapping of CSW ISO operations to the CSW mandatory operations.

Table 7 Mapping of CSW ISO operations and CSW operations to general model mandatory operations

General Model Operation	**CSW Operation**	**CSW ISO Operation**
OGC_Service.getCapabilities	OGC_Service.GetCapabilities	OGC_Service.GetCapabilities
Discovery.query	CSW-Discovery.GetRecords	CSW Discovery.GetRecords
Discovery.present	CSW-Discovery.DescribeRecord	CSW Discovery.DescribeRecord
Discovery.describeRecordType	CSW-Discovery.GetRecordById	CSW Discovery.GetRecordById

Both the ISO Metadata profile and the Core metadata profile were adopted in this research. Based on the syntax and/or semantics restrictions or variations of some parameters of the request and response for each operation defined by the OGC ISO metadata application profile, the information models adopted by underlying catalogues embedded in the integrated catalogue framework can be mapped to the standard ISO profile. This leads to the enhancement of the interoperability of information resources sharing. With the same method, query results can be mapped based on Core profile. The collaboration of ISO based catalogue and Core profile based catalogue can be achieved through the set of common queryable and returnable properties. To facilitate users without much domain knowledge to search with the integrated catalogue, a GUI was also developed based on the underlying catalogues.

CHAPTER FOUR INTEGRATED CATALOGUE FRAMEWORK

This chapter presents the design and development of the ICF, which discovers and integrates the distributed catalogues for satellite data and information products. To achieve interoperability of EO data discovery and sharing, this framework is built based on the standards and metadata information models described in Chapter 3. The integrated catalogue framework offers a harmonized interface for users to discover and access data resources from distributed catalogues on two levels of granularities: dataset level and granule level. This chapter first gives an overview of the ICF, then examines challenges in its design and development, and finally discusses methods of applying the specifications in the overall architecture. Detailed design of the two search granularity levels is introduced, including architectures and functions for core modules on each granularity level.

4.1 Overview

In the geospatial domain, a geospatial catalogue service provides a network-based meta-information repository and an interface for advertising and discovering shared geospatial data and services; the most widely used interface specification for geospatial catalogue services is the OGC CSW (Yue et al., 2011). However, due to the fact that catalogues are isolated from each other, heterogeneities exist in both interfaces and metadata information models, thus, lack of interoperability of EO data discovery and sharing remains a major issue. The ICF is designed to help users with limited domain knowledge to search geospatial resources over multiple distributed through a unified interface. As mentioned earlier, distributed EO data search is a 2-dimensional search. The designed framework should be able to fulfill the objective of extending both the scope and the granularity of EO data search. Figure 5 illustrates the context view of the ICF.

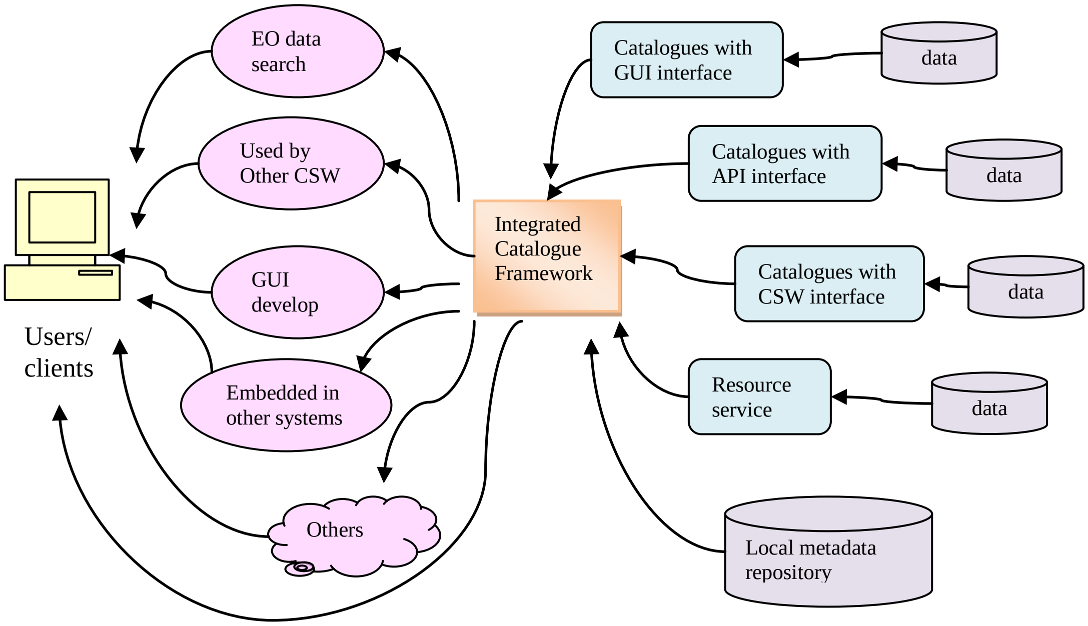

EO data search
Used by Other CSW
GUI develop
Embedded in other systems
Others
Users/ clients
Integrated Catalogue Framework
Catalogues with GUI interface
Catalogues with API interface
Catalogues with CSW interface
Resource service
Local metadata repository
data
data
data
data

Figure 5 Context view of the Integrated Catalogue Framework (ICF)

Various types of catalogues, such as catalogues with GUI interfaces, catalogues with API interfaces, catalogues with OGC CSW interfaces, resource services, and local metadata repositories can be integrated into ICF. The ICF significantly extends the scope of EO data search by linking to resources underlying multiple catalogues. On the other hand, the ICF can be used for direct EO data search, embedded in other systems, used as resources for GUI development, reused by other CSW, and other usages. A computational view of the framework is shown in Figure 6.

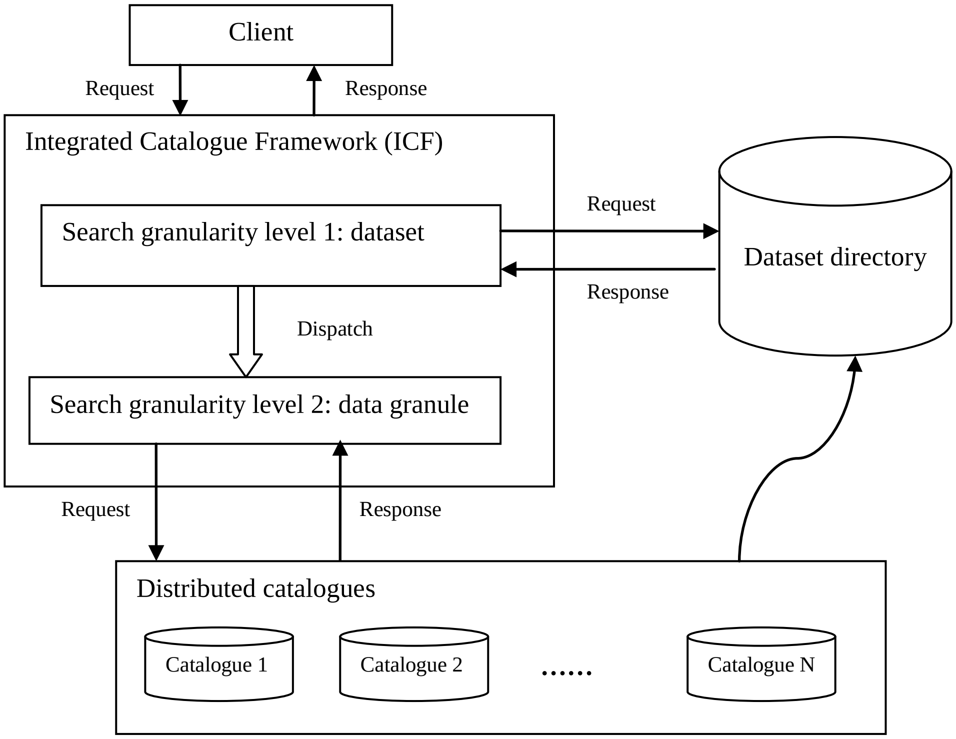


Figure 6 Computational view of the Integrated Catalogue Framework (ICF)

From Figure 6 we see how the integrated catalogue framework supports distributed catalogue search. When a client sends a query to the ICF, two levels of searches are processed: dataset level search and data granule level search. Data granule is defined as 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 attributed defined by the owning dataset (Shao et al., 2012). Dataset is defined as a group of granules with certain commonalities, for example from the same phenomenon. Datasets have characteristics that are common across all the granules they "own" and templates for describing additional attributes not yet part of the metadata model (Shao et al., 2012). When ICF receives a query, it first processes the query on the dataset level and search a local dataset directory to retrieve datasets that are relevant to the query. Then it reorganizes the query on the granule level by combining searching criteria with the previously retrieved datasets information to compose several queries for data granules. Those queries are dispatched to corresponding catalogues to which the datasets belong. The query responses from those catalogues are merged and returned to clients. Since the underlying distributed catalogues have different query interfaces, query languages and metadata information models, there are challenges in developing the ICF. These challenges will be discussed later. Section 4.3 and 4.4 describes the detailed methods including architectures and functions in how to deal with these challenges in building the two levels of search granularity. Figure 7 presents an information view showing the data flow of the ICF.

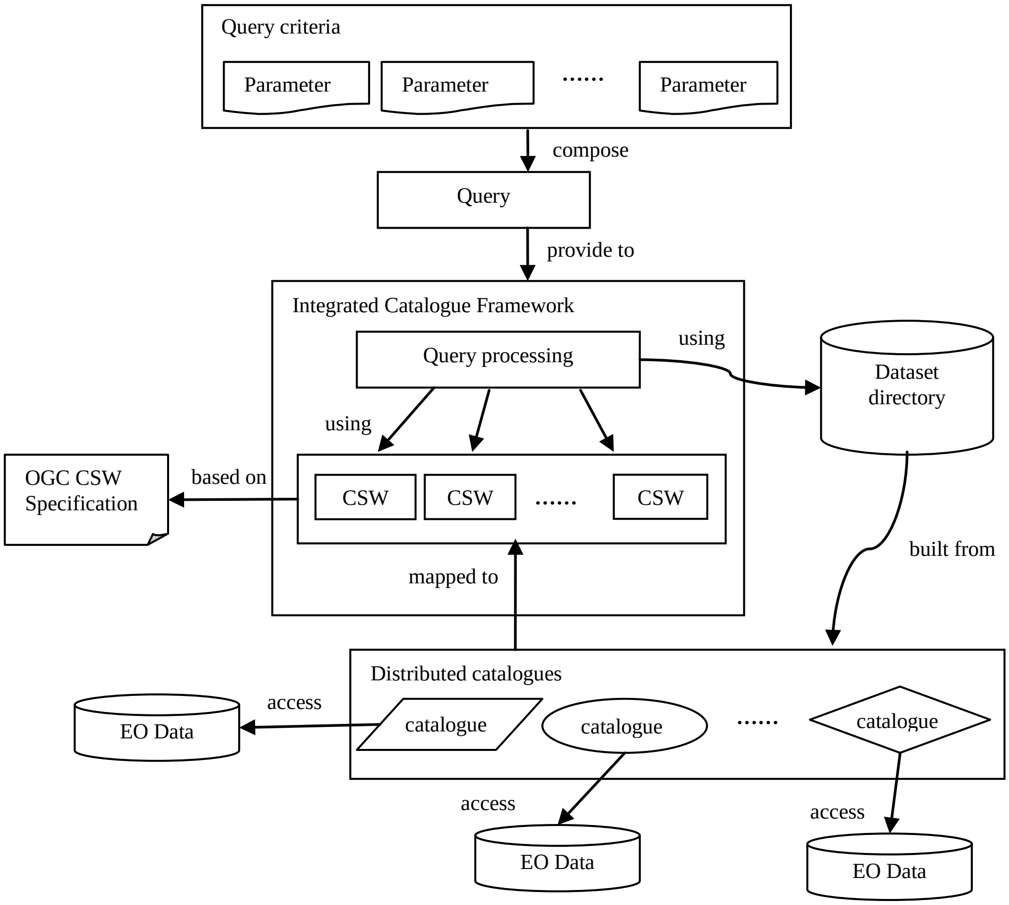


Figure 7 Information view of the Integrated Catalogue Framework (ICF)

Query criteria such as keywords, spatial and temporal constrains, result set types, requested profiles and other parameters are used to compose queries which are provided to the ICF. The query processing component utilizes the dataset directory to get relevant dataset information and determine target catalogues. The dataset directory is built with information extracted from distributed catalogues which index various EO data repositories. Those catalogues are mapped to standard CSW services based on OGC

CSW specification. Once target catalogues are determined, the CSWs are used by the query processing component to fulfill queries from users.

Integrated access to information that is spread over multiple, distributed and heterogeneous sources is an important problem in many scientific disciplines (Naumann et al., 1999). Due to the heterogeneity of query languages, communication protocols, metadata information model of catalogues, there lacks of interfaces for users without much domain knowledge to search EO data from distribute catalogues.

There are mainly five challenges in developing the ICF:

1) Keyword search: Most geospatial catalogues index EO data on the data granule level. They offer interfaces for users to search in specified datasets, which means users are required to have an advanced knowledge on datasets and catalogues to which their required data granules belong. To help users who are without much domain knowledge, dataset level keyword search function need to be provided. A dataset directory is built to address this issue. The granularity of EO data search is extended by developing dataset level keyword search for distributed catalogues serving granule level EO data.

2) Mapping of query languages: Mapping of query languages is a major challenge in developing the integrated framework. The query criteria, query format and query language are different for each catalogue, or even for different types of data archives in one catalogue. Thus before queries are sent to affiliated catalogue, they need to be converted to a specific format so that they can be recognized by these catalogues. To solve this problem, a wrapper component of the ICF needs to be developed to provide the

function of query mapping. After receiving queries, the wrapper transformed the format and content of the query to get it ready for corresponding system connectors.

3) Query dispatching: The challenge in query dispatching is to decide which specific catalogue should the query be dispatched when the ICF receives a query. To solve this problem, the mediator component in this framework is developed to function as the dispatching center. It takes the query and sends it to the affiliated catalogues services after analyzing the catalogue ID of the target dataset retrieved from dataset directory.

4) Protocol adaptation: Another challenge in developing the integrated catalogue framework is protocol adaptation. Spatial agencies may provide users/clients with different protocols to enable the communication between users/clients with their catalogues services. To solve this problem, the system connector component in the ICF is developed for the wrapper to communicate with their associated catalogue without worrying about protocols. The system connector component plays a role of connecting wrapper and online catalogue services, it interacts with catalogues directly. For each catalogue that is integrated into the framework, a corresponding system connector is developed so that the integrated model supports protocol adaptation.

5) Mapping of metadata information model: Mapping of metadata information model is another major challenge. The heterogeneity of spatial information that catalogues services provide makes the adoption of different metadata information models. Some of those metadata information models may follow international standards, unfortunately, there are still many catalogues utilize their own self-defined metadata whose contents and formats vary a lot. Thus in ICF, the mapping of metadata information

model is required. To solve the problem in mapping of metadata information model, the OGC CSW specification and the OGC CSW-ISO metadata application profile (Voges and Senkler, 2007) are adopted. Based on these specifications, two standard profiles are used in mapping of metadata information model: the ISO profile based on the ISO 19115 standards and Core profile based on the standards developed by Dublin Core Metadata Initiative (DCMI, 2010). When query results are returned from multiple catalogue services, they vary according to the characteristics of catalogues which makes it difficult to merge those query results. The solution for this issue also lies in the mapping of the respondent metadata.

After analyzing the context view, computational view and information view of the ICF and addressing five major challenges, the approach to design and develop the framework is described in the following section.

4.2 Approach

This section discusses specifications applied to the development of the framework. Specific operations defined in the specifications are introduced first. Then the architecture of the ICF is described. The mediator-wrapper architecture adopted in the framework architecture is discussed. The mediator and the wrapper components will be described in detail separately. Since it is necessary to build a system connector for each catalogue used in this research, at the end of this section, the functions of system connectors are discussed. The methods used in the interaction between system connectors and catalogue services will be explained too.

4.2.1 Adopt Standards in Integrated Catalogue Framework

The integrated catalogue framework is developed in compliance with OGC CSW specification. OGC CSW defines a set of interfaces for EO data discovery and access that can be supported by several protocol bindings. Among the three types of protocol bindings OGC CSW supported, the HTTP protocol binding is adopted in this research. To support the operations affiliated with interfaces defined by OGC CSW, the GET and POST methods are employed in the HTTP protocol binding. General concept in CSW with HTTP protocol binding is described in Chapter 3; we focus on the four mandatory operations which need to be implemented in HTTP protocol bindings in this section. These four mandatory operations are the GetCapabilities operation of OGC_Service class, the DescribeRecord operation, GetRecords operation and GetRecordById operation of the Discovery class as illustrated in Figure 8.

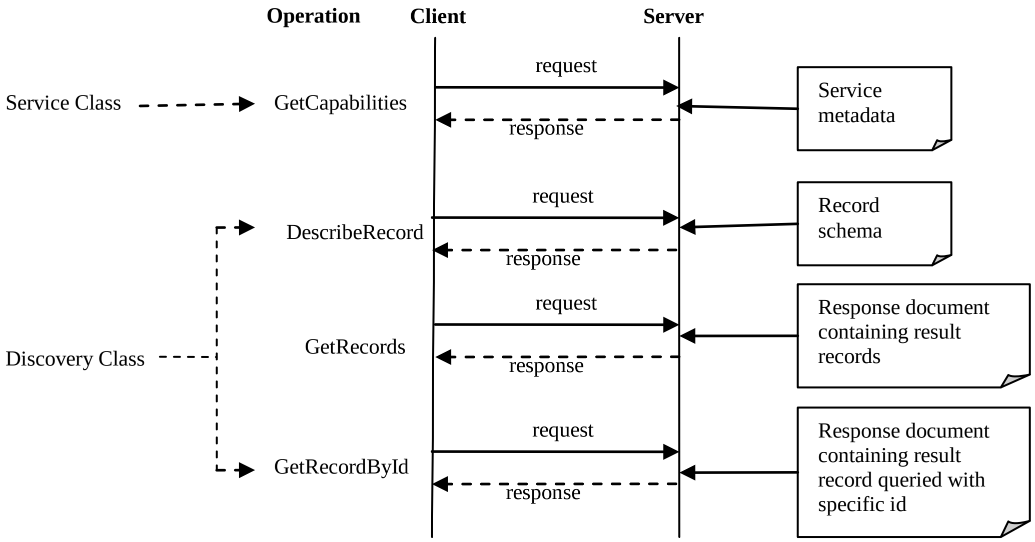


Figure 8 Overview of mandatory operations in HTTP protocol binding

The HTTP GET and POST methods are implemented for these four operations. For the GetCapabilities operation, the GET method and KVP encoding are used to request the service metadata. The value of the “request” parameter shall be “GetCapabilities” to inform the server to return the capabilities document as the response. The capabilities document should contain sections of: ServiceIdentification, ServiceProvider, OperationsMetadata, and Filter_Capabilities.

The ServiceIdentification and ServiceProvider sections provide metadata about a specified CSW implementation and the organization offering this CSW service. Information about title, abstract, keywords, service type, version, service provider name, site and contact information of service provider are included in this section.

In the OperationMetadata section, all the operations implemented by the ICF shall be listed: GetCapabilities, DescribeRecord, GetRecords, and GetRecordById including information indicating whether HTTP POST and GET method are supported by the corresponding operation. The names and possible values of parameters affiliated with each operation can also be used in the OperationsMetadata section.

The filter_Capabilities provide users with metadata about the filer capabilities of the server such as the spatial operator bounding box and comparison operators. An additional section FederationMetadata is added to provide users with more information about the catalogues embedded in the ICF. For each catalogue, the list of datasets it is capable to serve is included in this section. The static structure of the capabilities document is shown as Figure 9.

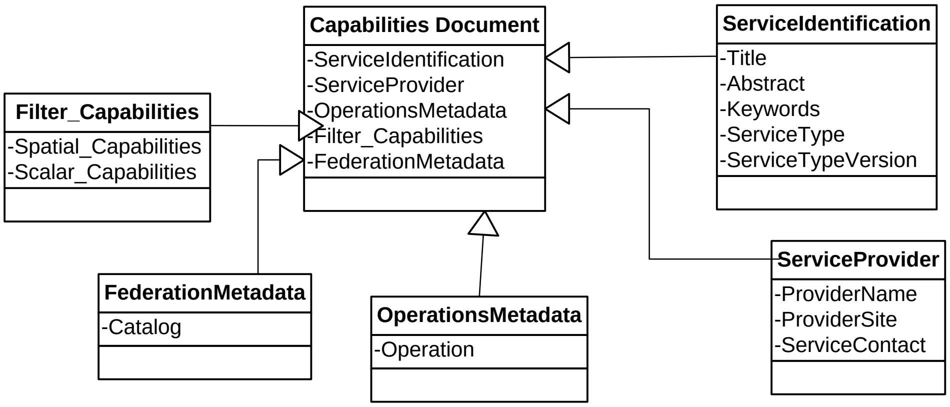


Figure 9 Static structure of capabilities document

For the DescribeRecord operation, the GET method and KVP encoding is implemented to discover elements of the information models supported by the ICF. Three mandatory parameters need to be specified in the KVP encoding for the DescribeRecord operation request: request, service and version. In this research, the fixed values for service and version are “CSW” and “2.0.2”. The value of the “request” parameter shall be “DescribeRecord” to inform the server to return the record schema as the response. The DescribeRecord response is an XML encoded document containing a DescribeRecordResponse element which is a container for zero or more SchemaComponent elements. Since both Core profile and ISO profile are supported in this research, the DescribeRecordResponse contains two SchemaComponent elements. Figure 10 shows the static structure of the DescribeRecord response.

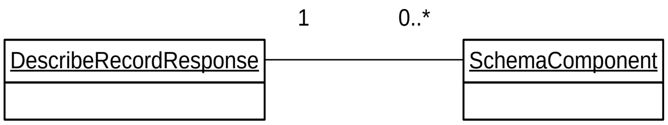


Figure 10 Static Structure of DescribeRecord Response

The HTTP POST method and XML encoding are applied in the GetRecords operation for resource discovery. A GetRecords request is encoded using the Query element which contains parameters typeName and Constraint. The typeName parameter specifies which set of elements of the specific information model should be queried. The Constraint parameter specifies the query criteria applied in the query, such as subject, spatial and temporal constrains. The query result is generated based on the required set of elements of a specific information model. The SearchResults element contains actual response to a GetRecords request with the set of records return by the GetRecords operation. The structure of the SearchResults element is shown in Figure 11.

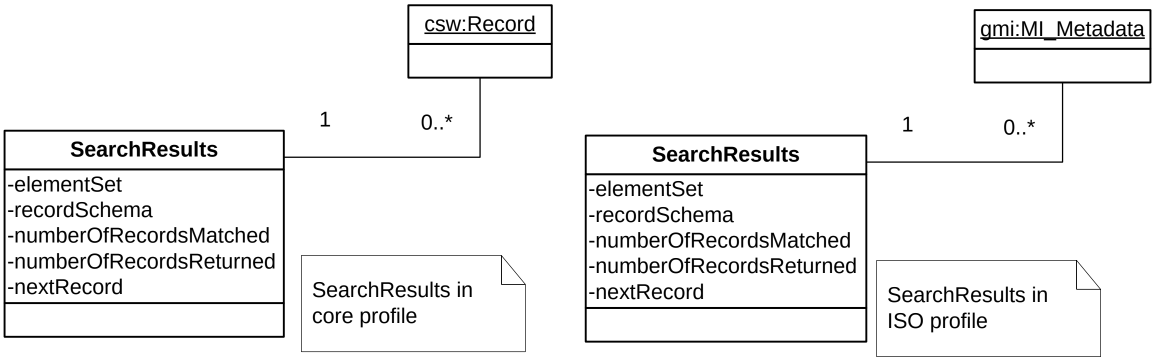


Figure 11 Structure of SearchResults element in Core profile and ISO profile

The GetRecordById operation retrieves catalogue record by specifying its Id. This operation provides a convenient short form to request record from a catalogue. The HTTP POST method and XML encoding are implemented for GetRecordById request.

Besides the three mandatory parameters need to be specified in the request, more than one <Id> elements can be used to retrieve multiple catalogues records. These records are generated based on the specific profile chosen by users. The GetRecordById operation can be considered as a subset of the GetRecords operation.

By implementing the GetCapabilities, DescribeRecord, GetRecords, and GetRecordById operations with HTTP protocol bindings in this research. The OGC CSW specification is applied in designing standard interfaces of the ICF. The ISO application profile is applied in the description of records retrieved from distributed catalogues.

4.2.2 Architecture

In this research, the mediator-wrapper architecture is used to design the ICF to achieve distributed catalogue search of earth observation data. The mediator architecture is one of those that have been proposed to address the problem of integration of heterogeneous information (Garcia-Molina, 1997). Lynden et al. (2008) used the mediator-wrapper architecture in his research of Service-based data integration to support compiling and executing queries over multiple web databases. Jafari et al. (2010) also adopted the mediator-wrapper concept in building a web-based renewable energy monitoring and management system. The overall architecture of the ICF is illustrated in Figure 12.

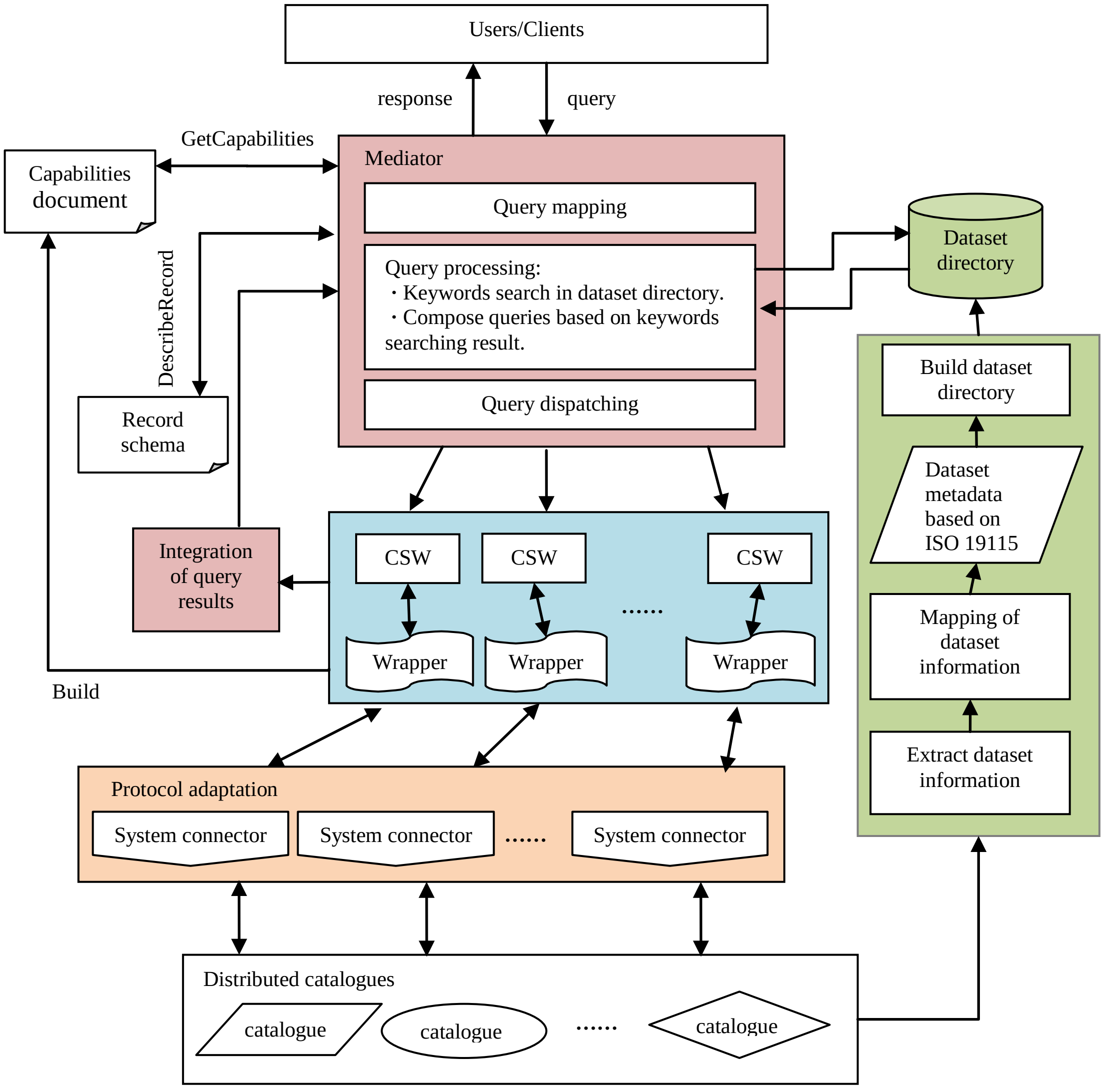


Figure 12 Architecture of Integrated Catalogue Framework (ICF)

With the ICF, both the scope and granularity of EO data search can be extended. By integrating multiple distributed catalogues and applying a well-recognized specification to wrap their interfaces and information models, the scope of EO data search is significantly expanded and the interoperability of these catalogues is enhanced.

The design and development of dataset-level keyword search for distributed catalogues serving granule level EO data not only extends the granularity of EO data search, but also facilitate users without domain knowledge to search distributed catalogues. The architecture can be divided into four major modules: the mediator module, wrapper module, system connector module and the dataset directory module. These four modules form two levels of search granularity: the mediator module and dataset directory module forms dataset level search; the wrapper module and system connector module form granule level search while the mediator module is partially involved in the granule level search.

For the GetCapabilites and DescribeRecord operations, two documents are generated as responses for users' requests correspondingly. Since the metadata information models applied in developing the ICF are based on the Core profile and ISO metadata application profile of OGC CSW, the record schema document will remains to be the same when a new catalogue is integrated into the framework while the FederationMetadata section of the capabilities document need to be updated with information of the new catalogue and datasets it supports.

As geospatial data becomes more widely available and used by people, their keyword based querying will become an important interface (Ganeshan et al., 2010). In the ICF, keyword search is applied on the dataset level to first identify which datasets are relevant to user's search criteria. Since datasets have information that is common across all the granules they contain plus a template for additional attributes, their metadata information is stored in a dataset directory to facilitate the search. For each catalogue embedded in the framework, there are three steps in establishing their dataset metadata information in the dataset directory:

1) Extract all dataset metadata, such as data description, list of keywords may be applied in searching, spatial extent, temporal range from individual catalogues.

2) Map extracted metadata to dataset metadata based on ISO 19115. This step can be skipped if the dataset information provided by catalogues is already based on ISO 19115.

3) With the standard dataset metadata from step 2, multiple methods can be used to build the dataset directory by applying different keyword searching strategies.

Once the dataset directory is built, it will mainly interact with the mediator which works as a dispatching center. The mediator takes search criteria from users/clients and extracts keywords from the queries to search the dataset directory to identify those datasets that potentially contain the data granule of interest to users. Next, in the data granule level search, the query is mapped and recomposed as several sub-queries since these datasets identified in the previous step may reside in multiple catalogues. Searching criteria and identified dataset information are used in the sub-queries composition procedure before they are sent to corresponding wrappers.

In the ICF, a wrapper is built for each backend catalogue. Each wrapper has a system connector that connects with its corresponding catalogue. The wrapper takes standard requests from user and delivers them to corresponding system connectors after adapting them to fit the connector. Wrappers not only convert requests from users or clients, but also wrap the searching results from system connectors. The search results sent back to the ICF system connectors vary according to the backend catalogues while the search results users receive from the ICF shall be consistent with the requested type. Thus the mapping and integration of search results are needed to provide users with search results consistent with standard information models. Wrappers parse the user-provided query to retrieve all necessary information to compose requests for each individual system connectors. Upon receiving those requests, system connector interacts with its corresponding catalogue to retrieve results of matched data granule information. These results are delivered back to the wrappers, which will wrap the results based on metadata information models for the mediator to return to users. Since system connectors directly interact with heterogeneous catalogues, they varies a lot depends on the speciality of catalogues.

4.3 Search Granularity Level 1: Dataset Search

As mentioned earlier, most geospatial catalogues only serve granule level EO data, which means users need to know which dataset their required data belong to in order to search data from multiple catalogues. To extend the granularity of EO data search and release users from the complexity of the relationship between datasets and data granules, a dataset level search is designed for those catalogues. Two major modules are involved in the dataset level search: the mediator module and the dataset directory module. There are three challenges in the dataset level search: enable keywords-based search, mapping of query languages, and query dispatching. These modules were designed to address these challenges. Functions of each module are described in Figure 13.

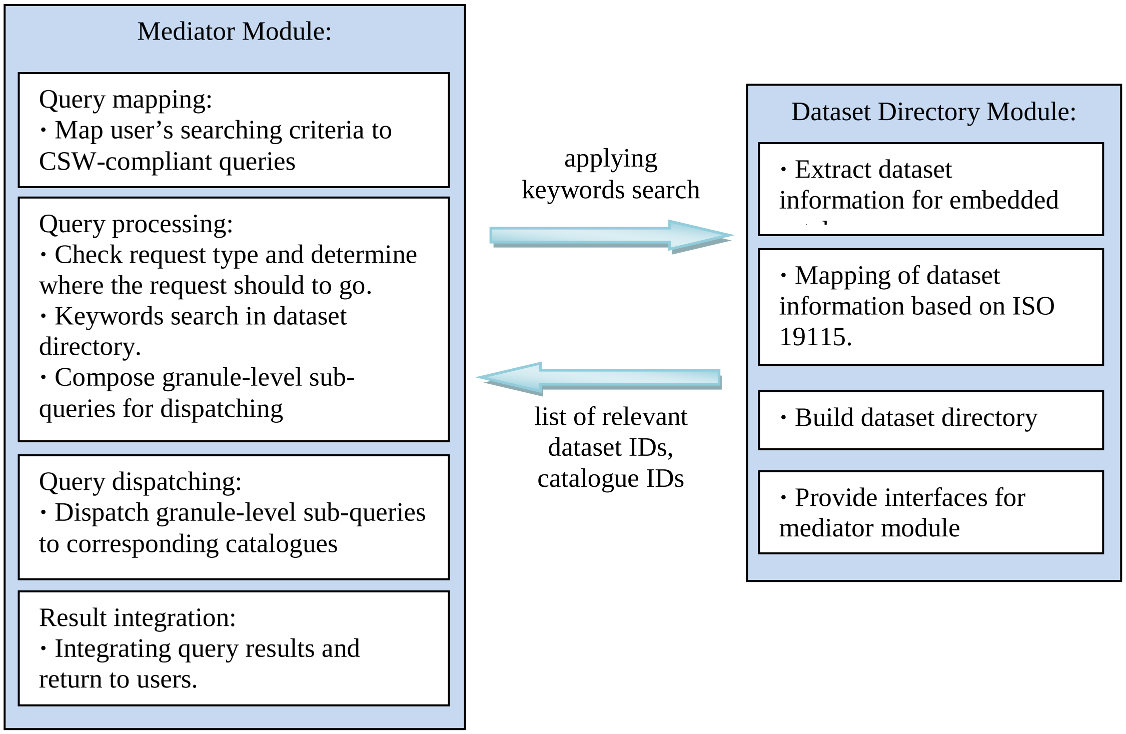

Mediator Module:
Query mapping:
• Map user's searching criteria to CSW-compliant queries
Query processing:
• Check request type and determine where the request should to go.
• Keywords search in dataset directory.
• Compose granule-level sub-queries for dispatching
Query dispatching:
• Dispatch granule-level sub-queries to corresponding catalogues
Result integration:
• Integrating query results and return to users.
applying keywords search
list of relevant dataset IDs, catalogue IDs
Dataset Directory Module:
• Extract dataset information for embedded
• Mapping of dataset information based on ISO 19115.
• Build dataset directory
• Provide interfaces for mediator module

Figure 13 Functions of mediator module and dataset directory module

The mediator module is responsible for query mapping, query processing, query dispatching, and integration of query results. It interacts with users, the dataset directory module, and the wrapper module directly. The dataset directory module initially collects dataset information from each catalogue linked to the ICF and builds a dataset metadata repository after mapping the information to ISO 19115. This module provides interfaces for the mediator module to retrieve dataset metadata to enable the keyword search function. Keywords-based searches continue to provide one of the most useful techniques for users to find information on a given topic, but they also form a starting point for a number of higher-level semantic queries that can be used in an information search (McCurley, 2001). Currently, most search engines inside spatial web portals are based on direct keyword matching, which can not effectively ‘understand’ the meaning of user’s queries, especially when a user has limited geospatial knowledge (Li et al., 2008). Two approaches for improving keyword search in dataset metadata are exploited in this research. One is to build a local metadata repository that saves the spatial, temporal, and other necessary information together with the dataset metadata files based on ISO19115 standard on top of which to apply the Term Frequency Inverse Document Frequency (TF-IDF) technique to rank the relevance of datasets and determine the results of keyword search. The other is to leverage GeoNetwork which is an open source catalog application for managing spatially referenced resources, and utilize the keyword search functions provided by GeoNetwork to manage dataset metadata.

4.3.1 Keyword Search in Local Metadata Repository

Before applying the dataset-level keyword search, a local dataset metadata repository needs to be built. There are two ways of collecting ISO 19115 metadata for datasets. One is to develop a registration service for data providers to register their products with the ICF. Through the registration process, dataset metadata in ISO 19115 format can be created. The other is to harvest existing dataset metadata from catalogues linked to the ICF and map them to ISO 19115 format. Figure 14 illustrates the process of building local dataset metadata repository.

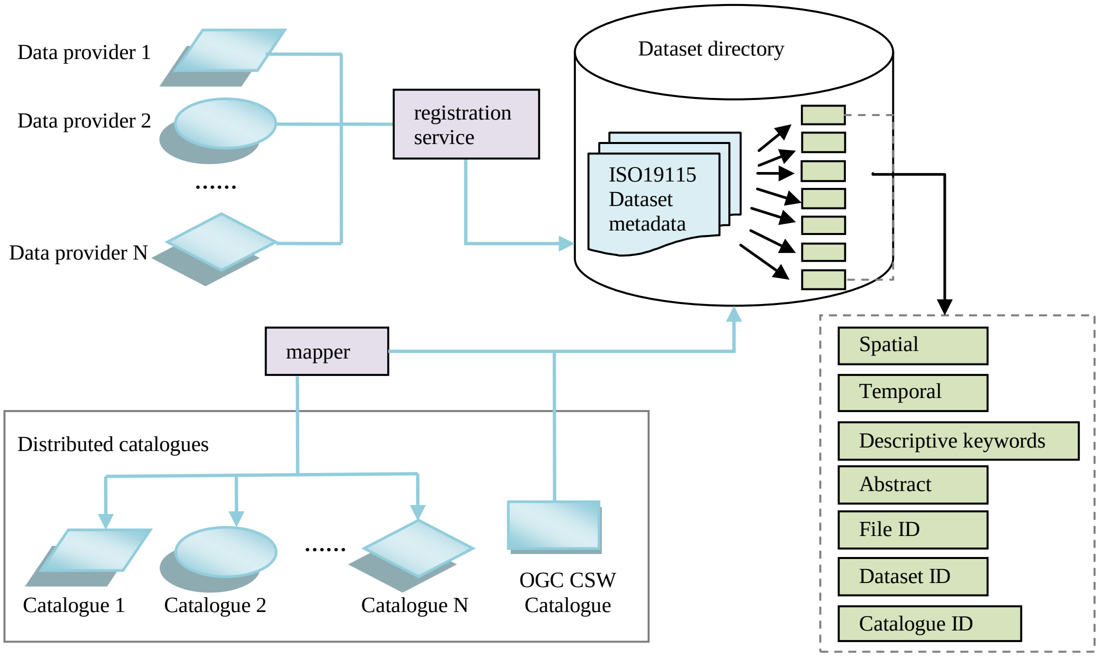


Figure 14 Build dataset directory with a local metadata repository

Once the ISO 19115 dataset metadata for all the datasets from distributed catalogues are collected, a table is generated to maintain the relationship between datasets and their corresponding catalogues. To facilitate the keyword search process from mediator module, relative information about the dataset should be stored in the table: spatial information including west bound longitude, south bound latitude, east bound longitude, and north bound latitude of the products in the dataset; temporal range including a starting and ending time point for the dataset; descriptive keywords which is a set of science keywords such as: agriculture, atmosphere, biological classification, biosphere, climate indicators, cryosphere, human dimensions, land surface, oceans, and so on; and abstract, which contains human-readable text describing the dataset. Both abstract and descriptive keywords belong to the identification information section of the

ISO 19115 metadata data. Besides, file ID records the identifier for ISO 19115 metadata of the dataset while dataset ID and Catalogue ID record corresponding identifier and relevant catalogue identifier of the dataset.

When a query is sent to the local dataset metadata repository, three steps will be performed to get the query results, as illustrated in Figure 15. The first step is to validate whether the spatial range of the dataset meets the searching criteria. Assuming the spatial range of a dataset is denoted as D and the bounding box specified by a query is denoted as Q, there are four types of relationships between D and Q: cover, covered by, overlap and disjoint. In this research, a loose method is used to validate the spatial range, which means a dataset is considered as unqualified for spatial validation only when D disjoint with Q (all points in D do not belong to Q and all points in Q do not belong to D). Otherwise, a dataset is considered as valid and continue to be used for further process.

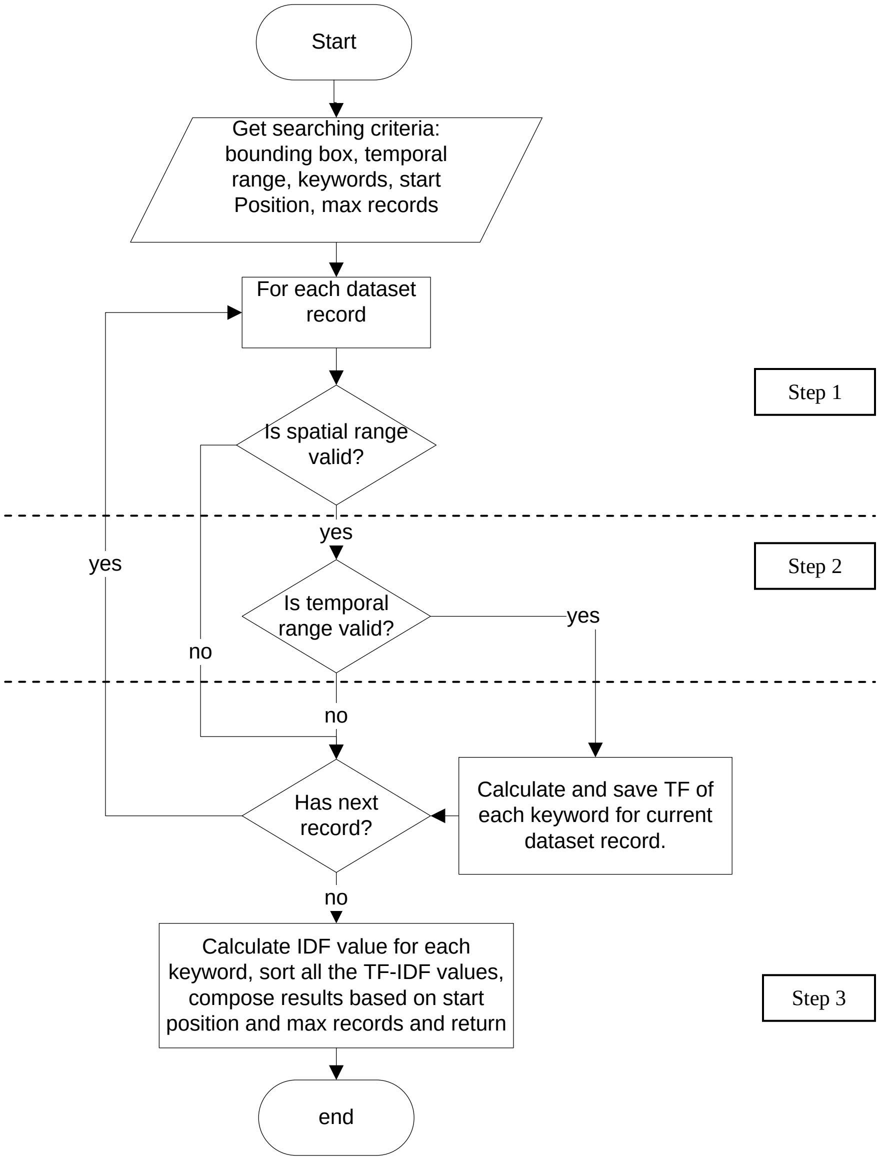

Start
Get searching criteria: bounding box, temporal range, keywords, start Position, max records
For each dataset record
Step 1
Is spatial range valid?
yes
yes
Step 2
Is temporal range valid?
yes
no
no
Has next record?
Calculate and save TF of each keyword for current dataset record.
no
Calculate IDF value for each keyword, sort all the TF-IDF values, compose results based on start position and max records and return
Step 3
end

Figure 15 Three steps for identifying searching results based on keywords

A simple way for implementing the spatial validation is to check whether any of the four vertices of Q locates inside D and then check whether any of the four vertices of D locates inside Q. If at least one of the validations return true, the dataset is qualified for spatial validation. The diagrams for checking whether a vertex is inside a spatial range D and for spatial validation of dataset are shown in Figure 16.

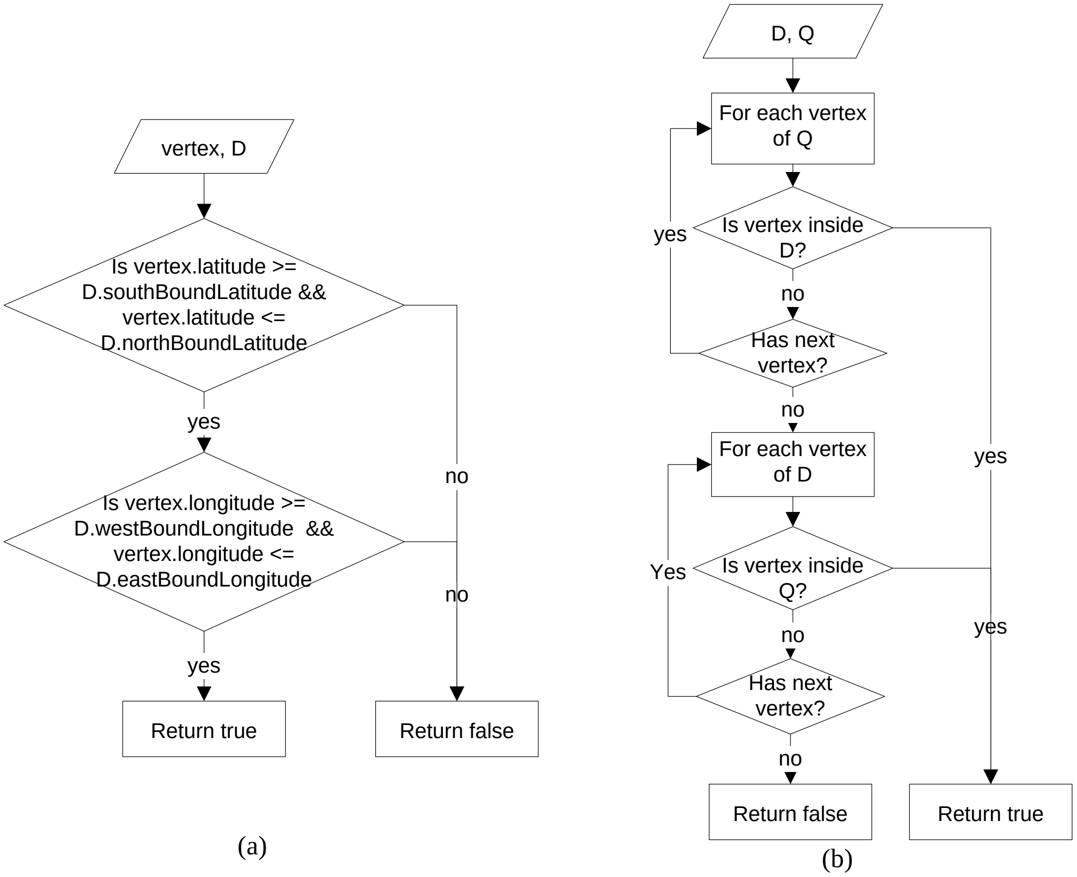


Figure 16 Inside validation (a) and spatial validation for dataset (D: spatial range of dataset; Q: spatial range of query) (b)

After the spatial validation, the second step is to check whether the temporal range of a dataset meets the searching criteria. Similarly, a loose method is used to validate the temporal range, which means only when the temporal range of dataset disjoints with the temporal range of the searching criteria, the dataset is considered as unqualified for temporal validation. Assuming the temporal range of a dataset is denoted as Td and temporal range specified by query is denoted as Tq, this step is implemented by checking whether the start or end time points of Tq located in Td , and then the start or end time points of Td located in Tq. If at least one of the validations return true, the dataset is qualified for temporal validation.

The reason for using loose method for the first two steps is to enlarge the search scope for the third step. Since in the granule level search, the spatial and temporal searching criteria will be applied again, the searching results are refined by specific bounding box and temporal range. On dataset level, after spatial validation and temporal validation, qualified records are extracted and ready for TF-IDF computing and ranking.

A term frequency factor is used as part of the term-weighting system to measure the frequency of occurrence of the terms in the document or query texts, however, term frequency factors alone cannot ensure acceptable retrieval performance hence a new collection-dependent factor must be introduced that favors terms concentrated in a few documents of a collection; the well-known inverse document frequency factor performs this function (Salton and Buckley, 1988). The concept of IDF can be dated back to 1972 when Karen Sparck Jones introduced logarithm in the term weighting system of her research in a statistical interpretation of term specificity and its application in retrieval (Jones, 1972). As the term implies, TF-IDF calculates values for each word in a document through an inverse proportion of the frequency of the word in a particular document to the percentage of documents the word appears in (Ramos, 2003). TF-IDF is the product of term frequency and inverse document frequency. In this research, descriptive keywords combining abstract information of a dataset is denoted as dataset description text *d*. Given a keyword *w*, there are various ways of defining the term frequency of *w*; a simplest choice is to use the raw frequency of *w* in *d*: f(*w*,*d*).

Definition 1 A simple term frequency tf(w,d) is defined as

$$\mathrm{tf}(w,d) = \mathrm{f}(w,d)$$

Where f(*w*,*d*) is the number of times that keyword *w* occurs in dataset description text *d*.

Definition 2 The inverse document frequency of *w* is given by

$$\mathrm{idf}(w, D) = \log(|D|/\mathrm{f}_{w,D})$$

Where *D* is a collection of text *d*, |*D*| is the number of texts in the corpus. $\mathrm{f}_{w,D}$ is the number of text *d*, where $w \in d$ and $d \in D$. IDF is used to measure whether a keyword is common or rare. When a keyword is common, its IDF value is low. For some extremely common keyword such as “the”, it may appear in every text in *D*, thus $\mathrm{f}_{w,D} = |D|$ and its IDF value is log(1) = 0.

Definition 3 Based on the above two definitions, the TF-IDF value of keyword *w* in *d*, $d \in D$ can be calculated as

$$\mathrm{tfidf}(w,d,D) =\mathrm{tf}(w,d)*\mathrm{idf}(w,D) = \mathrm{f}(w,d)* \log(|D|/\mathrm{f}_{w,D})$$

A high weight of TF-IDF is reached when the term frequency of a keyword in the give text is high and the document frequency of the keyword in the whole collection of texts is low, which means when a keyword appears in a small collection of texts and occurs many times in a given text, the TF-IDF weight for the keyword of the given text is high.

Definition 4 For a set of keywords in query Q, $Q = \{w_1, w_2, \ldots, w_N\}$, the TF-IDF value of Q in d, d $\in$ D can be calculated as

$$\mathrm{tfidf}(Q, d, D) = \sum_{i=1}^{N} [\mathrm{f}(w_i, d) * \log(|D|/\mathrm{f}_{wi,D})]$$

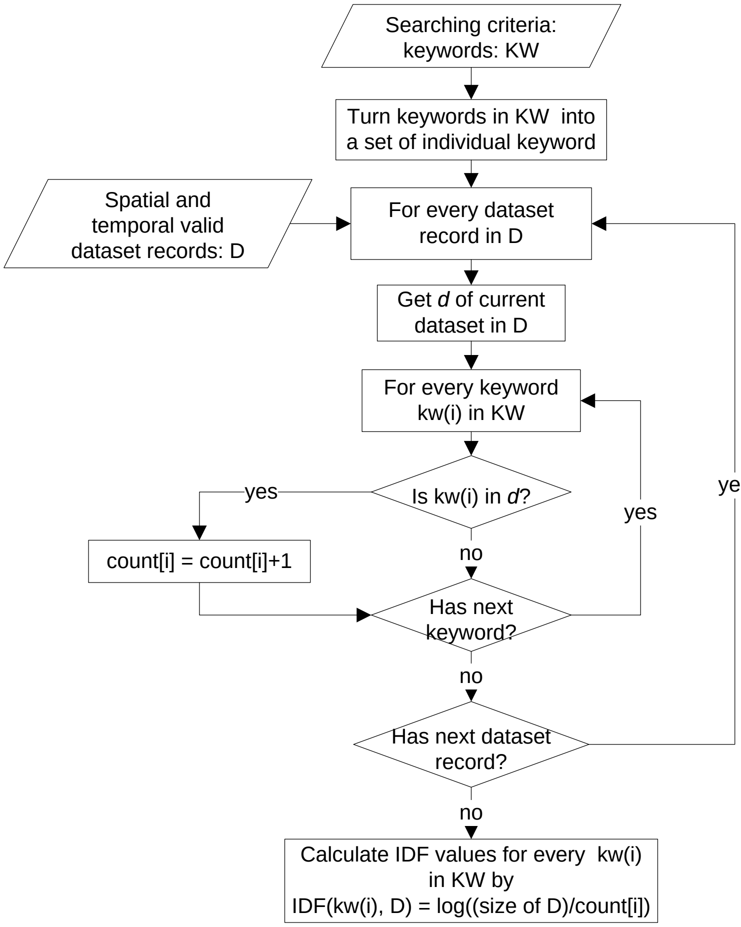
Example 1 Supposing the query “CO2 in atmosphere” is denoted as *q*, given a dataset description text *d* (combination of abstract and descriptive keywords sections in ISO 19115 dataset metadata), a dataset description text collection *D* which combines all the records qualified for spatial and temporal validation, the TF-IDF weight of the query in *d* (denoted as $W_{q,d}$) is given by

$$W_{q,d} = \mathrm{tfidf}(\text{“}CO2\text{”}, d, D) + \mathrm{tfidf}(\text{“}in\text{”}, d, D) + \mathrm{tfidf}(\text{“}atmosphere\text{”}, d, D).$$

$$= \mathrm{tf}(\text{“}CO2\text{”}, d) * \log(|D|/\mathrm{f}_{\text{”}CO2\text{”},D}) + \mathrm{tf}(\text{“}in\text{”}, d) * \log(|D|/\mathrm{f}_{\text{”}in\text{”},D})$$

$$+ \mathrm{tf}(\text{“}atmosphere\text{”}, d) * \log(|D|/\mathrm{f}_{\text{”}atmosphere\text{”},D}).$$

If we are having a corpus of documents which are all highly related with a specific domain then the TF-IDF weight of a word in a page gives the importance of that term for that document with respect to the whole corpus (Kumar and Vig, 2013). After the spatial and temporal validations of dataset records, the third step is to figure out relevant datasets with the keywords specified by a query. This objective is fulfilled by calculating and ranking the TF-IDF weights of the query for every valid dataset. In this research, the weight for a query in a dataset is determined by summing up the TF-IDF values for all keywords in the query. After that, all the TF-IDF weights are sorted and ranked; the result set is refined by two parameters of the query: start position and max records. By default, the IDs of datasets with the highest ten TF-IDF weights and their corresponding catalogue IDs are returned to the mediator module. Figure 17 describes the process of calculation TF-IDF values for a query. Before TF-IDF values are calculated, the IDF values of each keyword in the query are computed (Figure 17 (a)), and then TF-IDF weights of the query for every valid dataset are calculated (Figure 17 (b)).



(a)

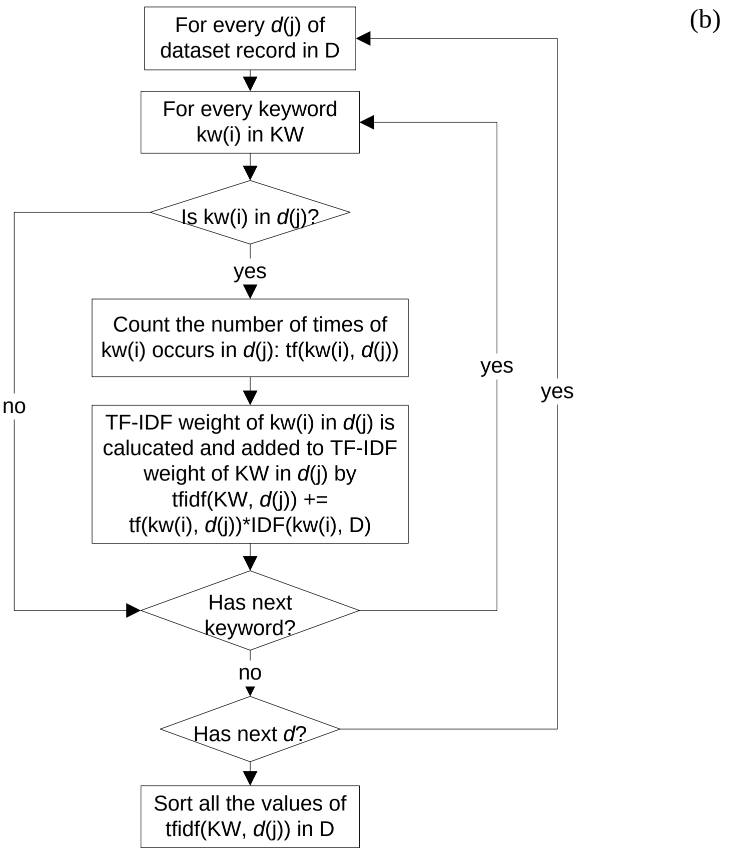

(b)
For every d(j) of dataset record in D
For every keyword kw(i) in KW
Is kw(i) in d(j)?
yes
no
Count the number of times of kw(i) occurs in d(j): tf(kw(i), d(j))
TF-IDF weight of kw(i) in d(j) is calucated and added to TF-IDF weight of KW in d(j) by tfidf(KW, d(j)) += tf(kw(i), d(j))*IDF(kw(i), D)
Has next keyword?
yes
no
Has next d?
yes
Sort all the values of tfidf(KW, d(j)) in D

Figure 17 Diagrams for computing IDF values (a) and TF-IDF weights of keywords (b) for datasets

As mentioned earlier, dataset description text *d* is a combination of the abstract and descriptive keywords sections in ISO 19115 metadata for a dataset. The variable *count[i]* in Figure 17(a) is used to count the number of *d* in the collection of dataset *D* where *keyword(i)* appears in *d*. Note that when *keyword(i)* does not appear in any dataset description text in *D*, *count[i]* equals to 0, the IDF value for *keyword(i)* is considered to be 0; when *keyword(i)* appears in every dataset description text, its IDF value is log(1), equals to 0 too. For other situations, 0 < *count[i]* < size of D, the IDF values will always be positive. The TF-IDF weights of a query in dataset description texts are used for ranking the relevance between keywords and dataset. Wu et al. (2008) justified mathematically and empirically that TF-IDF term weights could be the outcome of modeling relevance decision-making. In this research, the datasets with the highest TF-IDF weights of keywords are considered to be the most relevant query results.

4.3.2 Keyword Search in GeoNetwork

Besides applying the TF-IDF method to rank the relevance of dataset in a local metadata repository to enable keyword search function on the dataset level, the other method is to utilize the open source GeoNetwork application to manage dataset metadata. The search functions of GeoNetwork are extended by adding query processing methods. GeoNetwork is an open source software designed to improve accessibility of a wide variety of data together with the associated ancillary information (metadata), at different scale and from multidisciplinary sources; data are organized and documented in a standard and consistent way (Baldini et al., 2010). Before building the dataset directory with GeoNetwork, ISO 19115 dataset metadata need to be collected and inserted into GeoNetwork. Besides the methods for gathering ISO 19115 dataset metadata illustrated in Figure14, if the dataset information of a catalogue embedded in the ICF is registered in GCMD, we can simply harvest the ISO 19115 dataset metadata from the CSW provided by GCMD.

The prototype of GCMD was originally released by NASA in 1987, and it was renamed to GCMD in 1994; the GCMD is one of the largest public metadata inventories in the world; its primary responsibility is to maintain a complete catalog of all NASA’s Earth science data sets and services; however, the GCMD does not distribute datasets themselves (Tateishi et al., 2012). GCMD provides pointers to locations of data in various Earth science disciplines; the access here is at " directory level," which is represented by pointers indicating where the data collections of interest are held, To search and obtain specific data files covering a given region and/or a time interval, users would have to visit the respective data centers' sites (Yang et al., 2011). GCMD offers standard CSW interfaces for users to harvest dataset metadata. There are four steps for retrieving the ISO 19115 metadata for all datasets of a catalogue registered in GCMD.

1) Connect to GCMD CSW server, set request method as “POST” and set other request properties such as property “Accept” as “application/xml” and property “Content-Type” as “application/xml”;

2) Compose a CSW GetRecords request according to standards specified in section 3.2. Since we want to gather information of datasets at most, the “ElementSetName” is set as “full” and output schema is specified as ISO 19115. The datasets in target catalogues are identified by specifying corresponding properties of subject, identifier and comparison operators in the constraint and filter elements of the request. To get all the matched dataset records from target catalogues, the start position of the result is set as “1” and the “maxRecords” element is set as an integer larger than the maximum number of datasets in one catalog;

3) Post the request to GCMD CSW and retrieve the response;

4) Parse the response from GCMD CSW with XPath (path expressions to select nodes or node-sets in an XML document) of the set of nodes containing information for individual datasets. GCMD uses a unique identifier called “entryId” to identify datasets;

in this research, “entryId” is used as the file name for each individual ISO 19115 dataset metadata. Since the ISO 19115 dataset metadata itself does not store the relationship between datasets and catalogues, another file or table is needed to maintain the mapping relationship among the file identifier, dataset ID in its corresponding catalogue, and catalogue ID. Their relationships are illustrated in Figure 18.

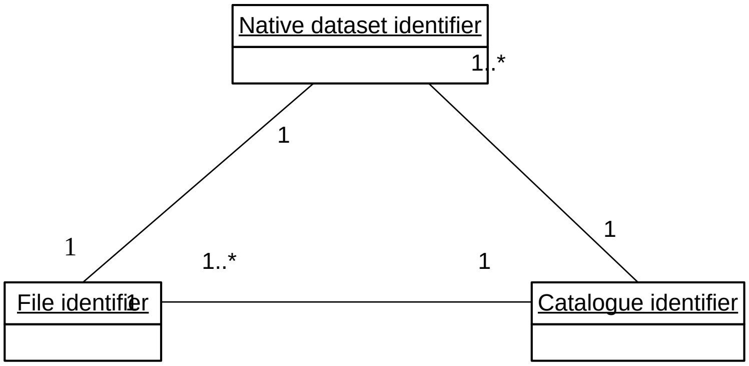
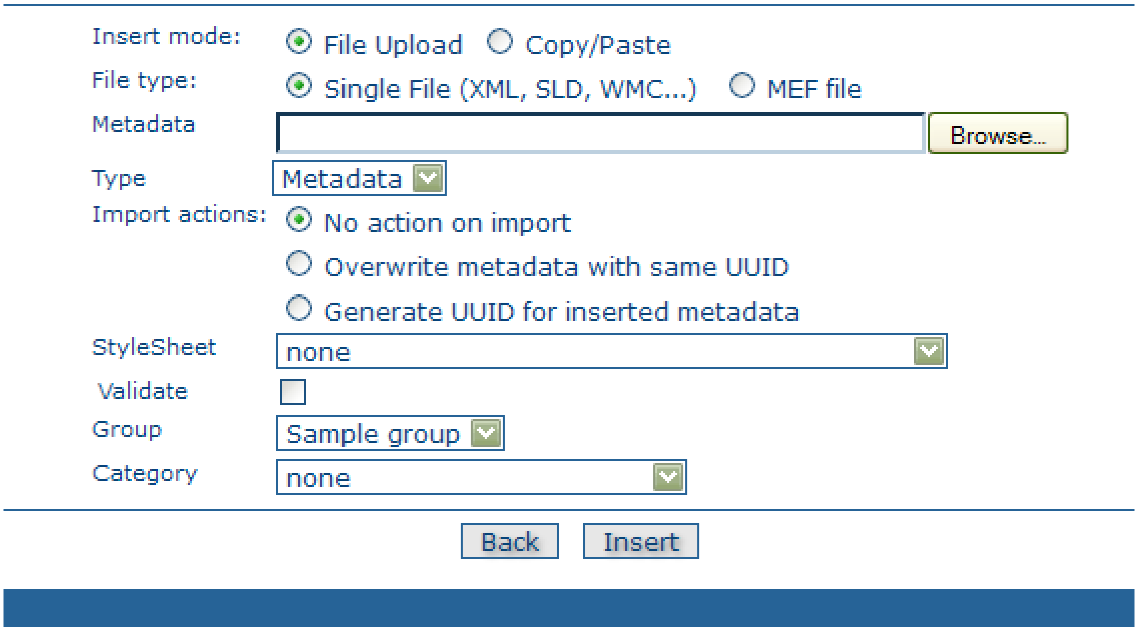


Figure 18 Relationship among file identifier, native dataset ID and catalogue ID

After ISO 19115 dataset metadata from all catalogues are collected, they are inserted in GeoNetwork. The GeoNetwork application has been developed with an intention to show connections between geodata, metadata and their standards; it has proved to be a very useful tool for direct support research and education activities at the Faculty of Science (Grill and Schneider, 2009). According to GeoNetwork user manual (GeoNetwork, 2013), support for ISO19115/ISO19119/ISO19139/ISO19110 and ISO Profiles, FGDC and Dublin Core formatted metadata is provided. Up-loading and downloading of data, graphics, documents, pdf files and many other content types are also supported; however, if we want to upload metadata files through the XML metadata insert tool, the metadata must be in one of the standards supported by GeoNetwork. In this research, dataset metadata in XML format is based on ISO 19115 standard thus it is applicable to insert the tool. GeoNetwork offers a convenient interface for importing metadata records (Figure 19), through this interface, all the dataset metadata for catalogues can be uploaded and then managed by GeoNetwork. To access the XML metadata insert tool, users should log in and select the "Metadata Insert" function in the administration page. A batch import interface is also provided to import all XML formatted metadata from a local directory.



Insert mode: ● File Upload ○ Copy/Paste

File type: ● Single File (XML, SLD, WMC...) ○ MEF file

Metadata [ ] Browse...

Type: Metadata

Import actions: ● No action on import
○ Overwrite metadata with same UUID
○ Generate UUID for inserted metadata

StyleSheet: none

Validate □

Group: Sample group

Category: none

Figure 19 XML metadata insert tool in GeoNetwork

As GeoNetwork already includes a CSW Server since it installs (Olfat et al., 2012), metadata can be retrieved by composing standard CSW GetRecords or GetRecordById request. In order to fulfill the keyword search function, the “any text” search function of the search and retrieve metadata service in GeoNetwork is adopted. Several methods for sorting the query results are supported: relevance, rating, popularity, change date, and title among which relevance is set as the default sorting criteria. GeoNetwork uses Lucene accuracy to determine the similarity between the query and search results. Apache Lucene is a high-performance, full-featured text search engine library written entirely in Java that is suitable for nearly any application requiring full-text search capabilities; it provides complex query processing by combining similarity search with the full-text search (Amato et al., 2013). In recent years, Lucene has become one of the most highly praise and most popular information retrieval library (Gao et al., 2012). Thus with the XML search service provided by GeoNetwork, the relevance sorting function for keyword search results is realized. Other searching criteria such as bounding box can also be applied in the search by adding it as filter of the request in constraint section. The results from GeoNetwork are parsed to extract the list of file identifiers, which are related with user’s query. Based on file identifiers, corresponding native dataset ID and catalogue ID of the dataset records need to be figure out. As mentioned above, a table is used to maintain the relationships among file ID, dataset ID and catalogue ID against which we can search necessary information and return results to the mediator module. The process of keyword search with GeoNetwork is illustrated as Figure 20.

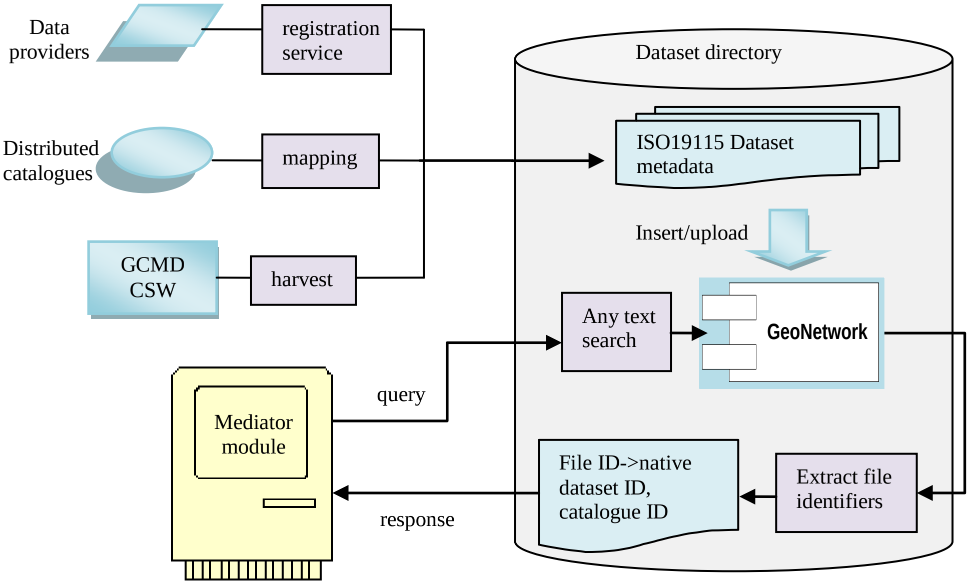


Figure 20 Process of keyword search in GeoNetwork

There are seven steps in the process of keyword search in GeoNetwork: 1) Gather ISO 19115 dataset metadata for catalogues embedded in the ICF; 2) upload the metadata with the aid of XML metadata insert tool in GeoNetwork; 3) connect to GeoNetwork; 4) set search criteria and compose any text search request; 5) send the request to GeoNetwork server and retrieve the result; 6) extract file identifiers for dataset records which are relevant to the query; 7) retrieve native dataset identifiers and catalogue identifiers with the file IDs from step 6 for future processing in the mediator module.

4.3.3 Query Processing, Mapping and Dispatching

The mediator module mainly provides four functions: query mapping, query processing, query dispatching and integration of query results, as described in Figure 13.

The first three functions are applied on the dataset level search. When a query is sent to the ICF, the mediator first analyzes the request type and determines proper actions need to be taken. If the request type is GetCapabilities or DescribeRecord, corresponding document is retrieved and returned to user; otherwise, searching criteria are extracted from the query and applied to keyword search in dataset directory and future processing (Figure 21).

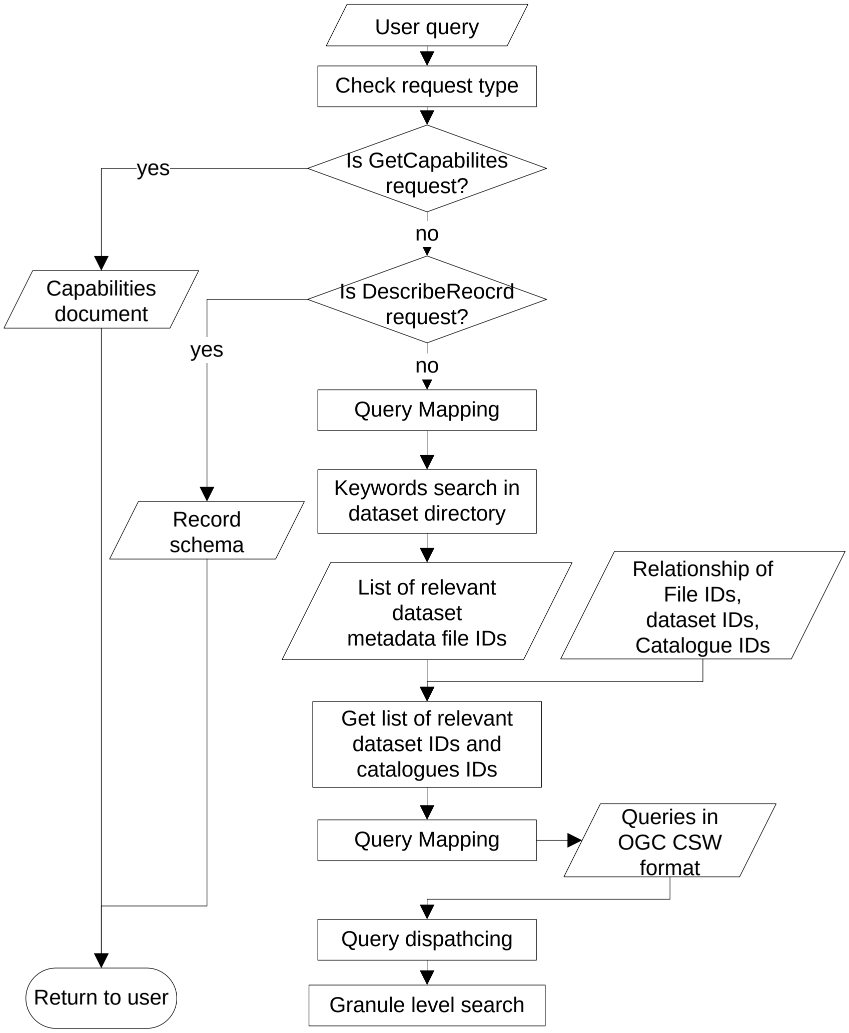

User query
Check request type
Is GetCapabilites request?
yes
no
Capabilities document
Is DescribeReocrd request?
yes
no
Query Mapping
Keywords search in dataset directory
Record schema
List of relevant dataset metadata file IDs
Relationship of File IDs, dataset IDs, Catalogue IDs
Get list of relevant dataset IDs and catalogues IDs
Query Mapping
Queries in OGC CSW format
Query dispathcing
Return to user
Granule level search

Figure 21 Diagram for query processing in mediator module

After relevant dataset metadata file identifiers are returned by the keyword search, corresponding native dataset IDs and catalogues IDs can be identified. Together with spatial, temporal and other search criteria, user's query request is interpreted and converted to several sub-queries. The objective of the query mapping process is to take search criteria to compose sub-queries based on OGC CSW standards. The mapping table between search criteria and OGC CSW requests is shown in table 8.

Table 8 Mapping of search criteria to OGC CSW request

Search Criteria	**OGC CSW attribute or property**
Native dataset ID	dc:subject
Spatial	ogc:BBOX
Temporal (Start, End)	dct:coverage.dataStart, dct:coverage.dataEnd
Start Record	startPostion
Maximum Record	maxRecord
Request element set type	ElementSetName
Profile	outputSchema

The native dataset ID is the identifier of a dataset used in the query process in its corresponding catalogue. The native dataset ID, spatial and temporal search criteria are mapped onto the constraint element of an OGC CSW request in which a set of comparison operators are allowed. The allowed comparison operators are defined in the capabilities document which can be accessed through GetCapabilities request. The start record and maximum record define the start position of the result set and the maximum number of records the user wants to retrieve respectively. Three types of element set are supported: full, summary and brief; the returned element sets varies according to different metadata information models chosen by the user: the profile parameter. Two types of profiles are supported in this research: OGC CSW Core profile (or Core profile for short) and ISO metadata application profile (or ISO profile for short). The advantage of query mapping is to release users from the burden to understand heterogeneous query interfaces of distributed catalogues. Since the catalogues embedded into the ICF are all wrapped based on OGC CSW specification, they offer standard CSW interfaces for EO data search. After applying keyword search of user's query to get relevant dataset records, the query is converted to individual sub-queries for each dataset, thus it is necessary to map each of the sub-queries to standard CSW requests.

Another major function of the mediator module is query dispatching. As mentioned earlier, there are three patterns of query dispatching: opaque, translucent, and transparent. In this research, to help users without much domain knowledge, both opaque and transparent patterns are supported by the ICF. Transparent query dispatching is easy since users already know where they can find data they need. By analyzing the catalogue ID and dataset ID specified by users, the mediator module simply composes CSW standard queries and dispatches them to corresponding catalogues. This can be achieved by exposing affiliated catalogue services to users such as lists of available catalogues and datasets in the ICF. In opaque pattern, users have no awareness of underlying catalogues. After user's query is processed by keyword search and query mapping, a list of granule-level queries compatible with OGC CSW format and their relevant catalogue IDs are created. Then the mediator module dispatches those queries to their corresponding catalogue wrappers based on the catalogue identifiers respectively.

4.4 Search Granularity Level 2: Granule Search

As proposed earlier, to integrate different catalogues into one harmonized model for users to search EO data via one common recognized metadata model, the mediator-wrapper architecture is adopted. In this research, the OGC CSW specification is utilized as the constraint for developing the “mediator” and “wrapper” modules for distributed catalogues. For dataset level search, the mediator module does query processing and mapping. It extends the search granularity beyond the granule level by enabling keyword search function of the dataset directory module. For granule level search, distributed catalogues are standardized and integrated to extend the scope of EO data search. Two modules form the granule level search in the ICF: the wrapper module and system connector module. The mediator module is partially involved with granule level search to integrate query results from wrappers and response to users. The design and development of granule level search in the ICF are described in detail in the following sections.

4.4.1 Overview of Granule Level Search

Distributed geospatial catalogues serving granule level EO data use diverse interfaces, query languages, and metadata information models, thus if users need to discover and access data from multiple catalogues, they have to understand how to interact with each catalogue. To extend the scope of EO data search and hide users from the complexity of catalogues embedded in the ICF, the integration of distributed catalogues is needed. The granule level search is designed for this goal. The concept is to develop a system connector for each of the catalogues and wrap those connectors into standard interfaces based on OGC CSW specification to offer harmonized service interfaces. Then those interfaces serve as plug and play functions for the mediator module.

Mediator analyzes query from user to convert it into a set of relevant sub-queries. It then decides which wrapper each sub-query should go to. As soon as the search results of sub-queries are returned from their corresponding wrappers, the mediator module integrates the results and returns them to user.

There are four challenges in the dataset level search: mapping query mapping, metadata information module mapping, protocol adaptation and query results integration. To deal with these challenges, three modules are created for the granule level search: the wrapper module, the system connector module and the mediator module. The wrapper module provides two levels of mapping. One is query mapping, which wraps the input of system connectors to hide users from heterogeneous query interfaces. The other is metadata information model mapping, which wraps the output of system connections based on metadata models supported by OGC CSW, discussed in Chapter 3. For protocol adaptation, the system connector model is designed to connect wrappers with their corresponding catalogues. Due to the heterogeneity of distributed catalogues, in this research, for each catalogues embedded in the ICF, an individual system connector is designed and developed to communicate with catalogue respectively. Before search results are returned to users, the mediator collects the responses from relevant wrappers and integrates them to compose the final query response for users. The functions of each module for granule level search are described in Figure 22.

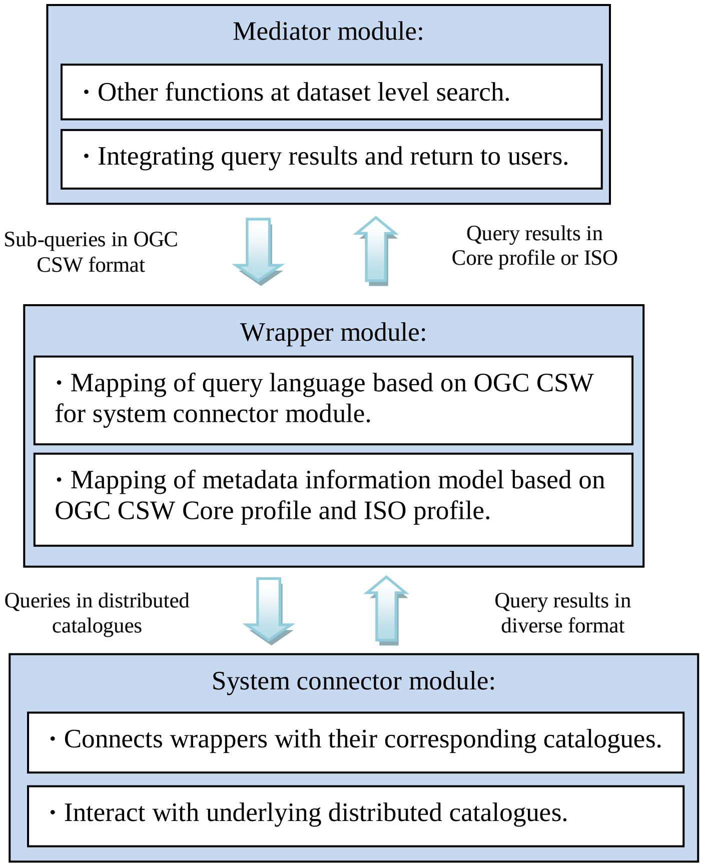

Mediator module:
• Other functions at dataset level search.
• Integrating query results and return to users.
Sub-queries in OGC CSW format
Query results in Core profile or ISO
Wrapper module:
• Mapping of query language based on OGC CSW for system connector module.
• Mapping of metadata information model based on OGC CSW Core profile and ISO profile.
Queries in distributed catalogues
Query results in diverse format
System connector module:
• Connects wrappers with their corresponding catalogues.
• Interact with underlying distributed catalogues.

Figure 22 Functions of mediator module, wrapper module and system connector module at granule level search

In summary, the granule level search is to extend the scope of EO data search by integrating distributed catalogues. The interoperability of these catalogues is also achieved through wrappers based on commonly recognized standards.

4.4.2 Wrapper and System Connector

When a query is dispatched by the mediator and sent to an individual wrapper, from the developer's point of the view, the wrapper takes standard OGC CSW query as input and translates it to a form that accommodates each underlying catalogue; from the user's point of view, the wrapper wraps heterogeneous catalogues to the OGC CSW standard. Together with the functions provided by the system connector module, the wrapper module provides standard interfaces for the mediator module to retrieve granule level metadata. The wrapper module can be considered as a translator for the input and output of the system connector module.

The search process in an underlying catalogue is enabled by a set of functions defined in system connector which varies across distributed catalogues. These functions include signing in a distributed catalogue if needed, searching in the catalogue with the query translated by a corresponding wrapper, and retrieving “raw” query results. Due to heterogeneity of distributed catalogues, multiple steps in the search process may be required. The functions to interact with underlying distributed catalogues embedded in the ICF are designed and developed in the system connector module. The relationships among mediator module, wrapper module, system connector module, and distributed catalogues are demonstrated as Figure 23.

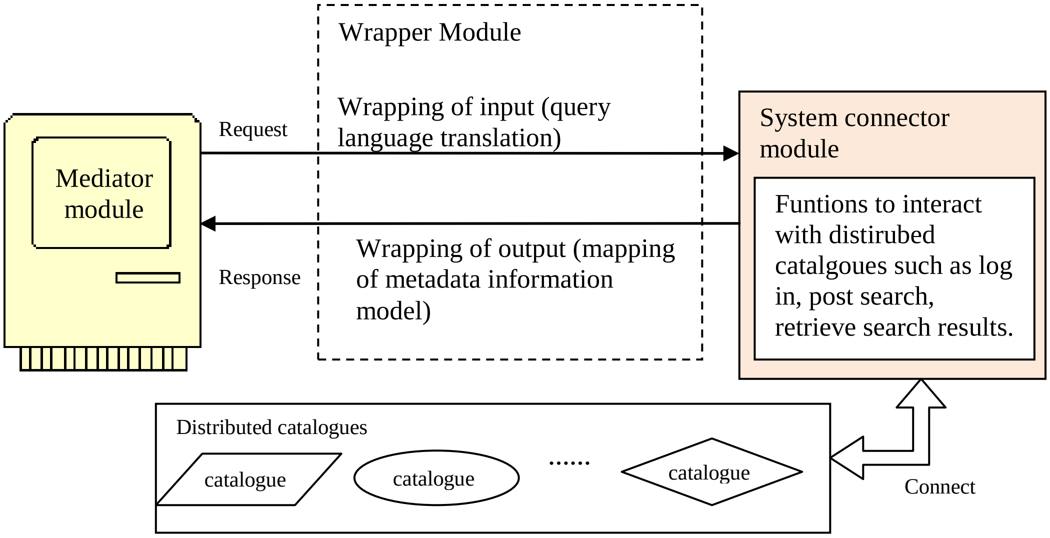
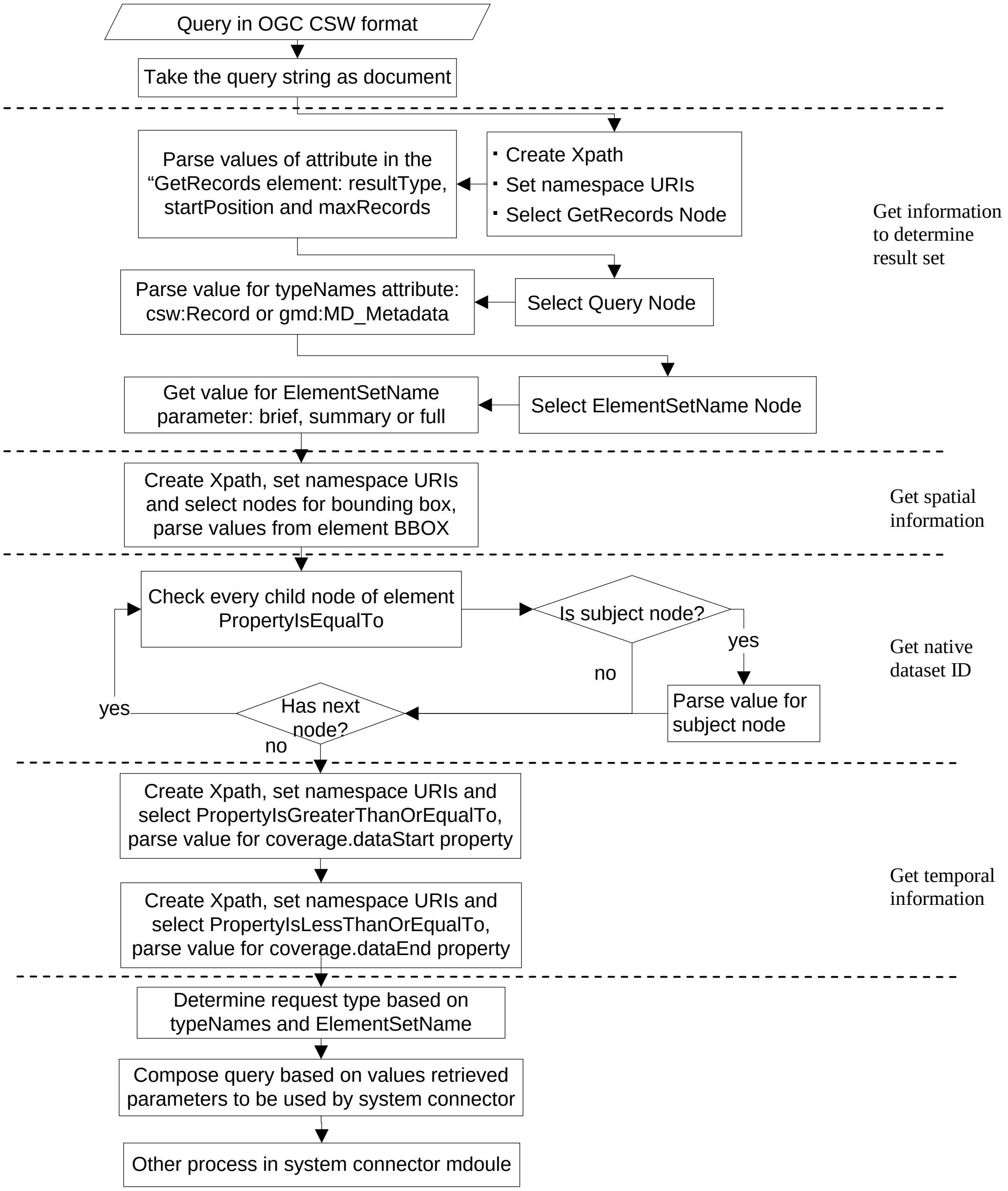

Wrapper Module
Wrapping of input (query language translation)
Request
Mediator module
Response
Wrapping of output (mapping of metadata information model)
System connector module
Funtions to interact with distirubed catalgoues such as log in, post search, retrieve search results.
Distributed catalogues
catalogue
catalogue
……
catalogue
Connect

Figure 23 Relationship among mediator module, wrapper module, system connector module and distributed catalogues

Although the result of query mapping depends on the underlying target catalogue, the process of parsing the query from mediator to retrieve search criteria is the first step consistent for every wrapper. As mentioned above, the mediator dispatches OGC CSW standard queries to the wrapper module. The process of parsing OGC CSW query in the wrapper is described in Figure 24. Based on those search criteria, a new query is composed in a format that is utilized by the target catalogue. Then the recomposed query is adopted by the system connector to interact with its catalogue. The characteristics of the interaction between system connectors and distributed catalogues vary according to the interfaces and functions provided by each catalogue. However, the basic concept in developing system connectors lies in three steps: 1) connect to a specific catalogue through its interface; 2) post the recomposed queries from wrapper to corresponding catalogue, following its search process; 3) retrieve search results returned by the catalogue.



Query in OGC CSW format
Take the query string as document
Parse values of attribute in the "GetRecords element: resultType, startPosition and maxRecords
▪ Create Xpath
▪ Set namespace URIs
▪ Select GetRecords Node
Get information to determine result set
Parse value for typeNames attribute: csw:Record or gmd:MD_Metadata
Select Query Node
Get value for ElementSetName parameter: brief, summary or full
Select ElementSetName Node
Create Xpath, set namespace URIs and select nodes for bounding box, parse values from element BBOX
Get spatial information
Check every child node of element PropertyIsEqualTo
Is subject node?
yes
no
Get native dataset ID
Parse value for subject node
Has next node?
yes
no
Create Xpath, set namespace URIs and select PropertyIsGreaterThanOrEqualTo, parse value for coverage.dataStart property
Create Xpath, set namespace URIs and select PropertyIsLessThanOrEqualTo, parse value for coverage.dataEnd property
Get temporal information
Determine request type based on typeNames and ElementSetName
Compose query based on values retrieved parameters to be used by system connector
Other process in system connector mdoule

Figure 24 Architecture of query parsing in the wrapper module

After extracting necessary information from search results from distributed catalogue, they are mapped based on the Core and the ISO metadata application profiles supported by OGC CSW. Detailed mapping for the properties extracted from query results are described in Table 5 and Table 6 in Chapter 3.

4.4.3 Query Results Integration

The functions for integrating query results are provided by the mediator module. After applying search criteria in system connectors, the query result of each relevant dataset contains several granules. There are five attributes defined in the "csw:SearchResults" element .When query results from multiple wrappers are collected by the mediator, it checks the element set name and profiles requested by users. Based on this information, the "elementSet" and "recordSchema" attributes can be defined in advance. Then the values of "numberOfRecordsMatched" and "numberOfRecordsReturned" attributes in the query result from each catalogue are checked and added; the sums for each of the attribute are calculated and recorded in the final results. Then the value for "nextRecord" attribute is determined by the start position and the final "numberOfRecordsReturned" value. After all five attributes of the search results element are retrieved, each granule record in the sub-query result sets is extracted and used as a child node for the "csw:SearchResults" element: the "csw:Record" node in OGC CSW Core profile or "gmi:MI_Metadata" node in the ISO profile. With this method, all the sub-query results are integrated and returned to users.

CHAPTER FIVE DISTRIBUTED CATALOGUE SEARCH SYSTEM

This chapter presents the implementation of a distributed catalogue search system based on the concept of ICF proposed in the previous chapter. This system demonstrates the capabilities of ICF to integrate distributed catalogues and achieve interoperability of EO data discovery and sharing. In order to build this system and demonstrate these two levels of search granularity, two distributed online catalogues, the NOAA CLASS catalogue and NASA ECHO catalogue, are utilized as data sources. This chapter first describes how to build dataset directory with these selected data sources. These catalogues are then wrapped based on the methods described in Chapter 4 and the standards described in Chapter 3. The methods for integrating those data sources are described in detail. At the end of this chapter, a use case is introduced and the results from the distributed catalogue search system are discussed.

5.1 Data Sources

5.1.1 Build Dataset Directory

There are various methods of building the dataset directory as discussed in section 4.3. Since the NOAA CLASS and NASA ECHO catalogues used as data sources in this research both have been populated in GCMD, a simple method is to harvest dataset metadata for these two catalogues directly from GCMD. GCMD offers a standard OGC CSW interface (http://gcmdsrv.gsfc.nasa.gov/csw), through which we retrieved the results. To get dataset metadata for the NOAA CLASS and NASA ECHO catalogues, a set of constraints were specified (Table 9) in the CSW requests correspondingly. Together with these constraints, “ElementSetName” in the query were set as “full” and “outputSchema” were set as “http://www.isotc211.org/2005/gmd” in order to retrieve full ISO 19115 dataset metadata.

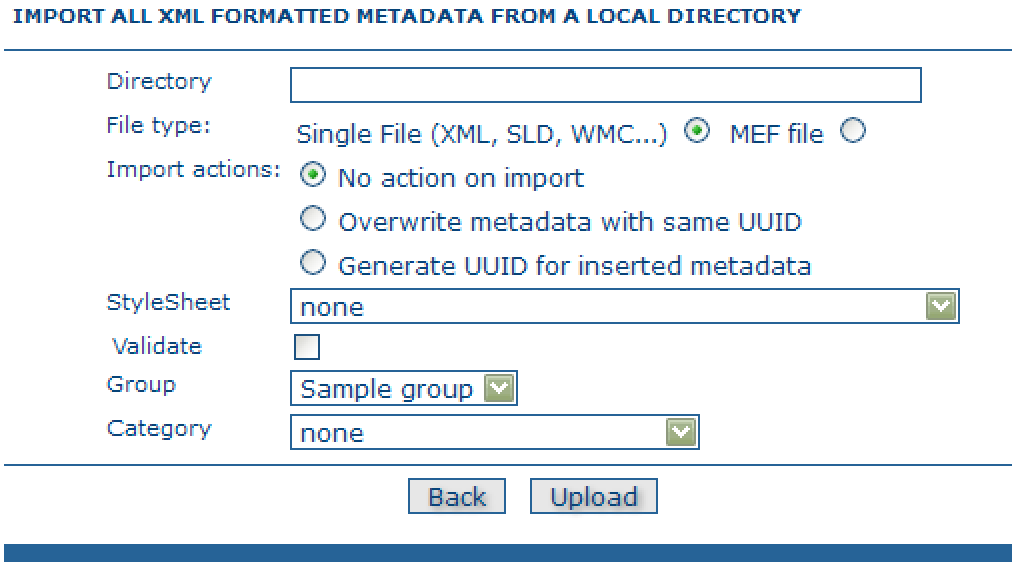
Table 9 Constraints for harvesting dataset metadata for NASA ECHO and NOAA CLASS

Catalogue	**Constraint/Filter**		
	Property Name	**Value**	**Comparison operator**
NASA ECHO	subject	USA/NASA	PropertyIsEqualTo
	subject	ECHO	PropertyIsEqualTo
NOAA CLASS	Identifier	gov.noaa.class.%	PropertyIsLike

After all relevant dataset metadata were retrieved from GCMD, the record for each individual dataset was extracted and saved as an XML file, which was named with the corresponding values of element “gmd:fileIdentifier” in the ISO 19115 metadata. A mapping file in XML format was generated to store the relationship among file ID, native dataset ID, and catalogue ID. In the mapping file, catalogue was the parent node for a set of datasets. Each catalogue was identified by an attribute “id”. In the distributed catalogue search system, there were two catalogues with their corresponding element <catalog id = “NOAA”> and <catalog id=”NASA”> saved in the mapping file. File ID and native dataset ID were specified as “entryId” and “datasetId” attributes respectively for a dataset element in the mapping file. A fragment of the mapping file is shown as below:

```
    <mappingList>
        <catalog id="NOAA">
            <dataSet datasetId="ASCAT"
entryId="gov.noaa.class.ASCAT">Advanced Scatterometer Level 1B</dataSet>
            <dataSet datasetId="AVHRR"
entryId="gov.noaa.class.AVHRR">Advanced Very High Resolution Radiometer</dataSet>
            <dataSet datasetId="CORBL" entryId="gov.noaa.class.CORBL">Coral
Bleaching Monitoring Datasets</dataSet>

            ……

        </catalog>
        <catalog id="NASA">
            <dataSet datasetId="CAMEX-3 CLOUD AND AEROSOL PARTICLE
CHARACTERIZATION (CAPAC) V1" entryId="dc8capac">CAMEX-3 Cloud and Aerosol Particle
Characterization (CAPAC)</dataSet>
            <dataSet datasetId="ACES CONTINUOUS DATA V1"
entryId="aces1cont">ACES Continuous Data</dataSet>
            <dataSet datasetId="ACES LOG DATA V1" entryId="aces1log">ACES
Log Data</dataSet>
            <dataSet datasetId="ACES TIMING DATA V1"
entryId="aces1time">ACES Timing Data</dataSet>

            ……

        </catalog>
    </mappingList>
```

With the mapping file, the relationship among file ID, native dataset ID and catalogue ID were stored for usage in query dispatching. After all the dataset metadata for the two catalogues were retrieved, they were imported into GeoNetwork through its batch import interface (Figure 25). Then the dataset directory was ready for the keyword search in the mediator module, which adopted the “AnyText” search function provided by GeoNetwork.



IMPORT ALL XML FORMATTED METADATA FROM A LOCAL DIRECTORY

Directory:

File type: Single File (XML, SLD, WMC...) ● MEF file ○

Import actions:
- ● No action on import
- ○ Overwrite metadata with same UUID
- ○ Generate UUID for inserted metadata

StyleSheet: none

Validate: □

Group: Sample group

Category: none

Figure 25 Batch import interface in GeoNetwork

For catalogues which are not registered in GCMD, necessary information of their dataset need to be extracted to create ISO 19115 metadata to build the dataset directory. These information include file identifier, language, character set, hierarchy level, contact, data stamp, metadata standard name, metadata standard version, identification information, distribution information, data quality information, metadata constraints, and metadata maintenance. The content and format of dataset metadata are specified in the ISO 19115 standard. In the next step of this research, a registration service is needed to collect dataset information of catalogues and generate metadata for them based on profiles supported by OGC CSW.

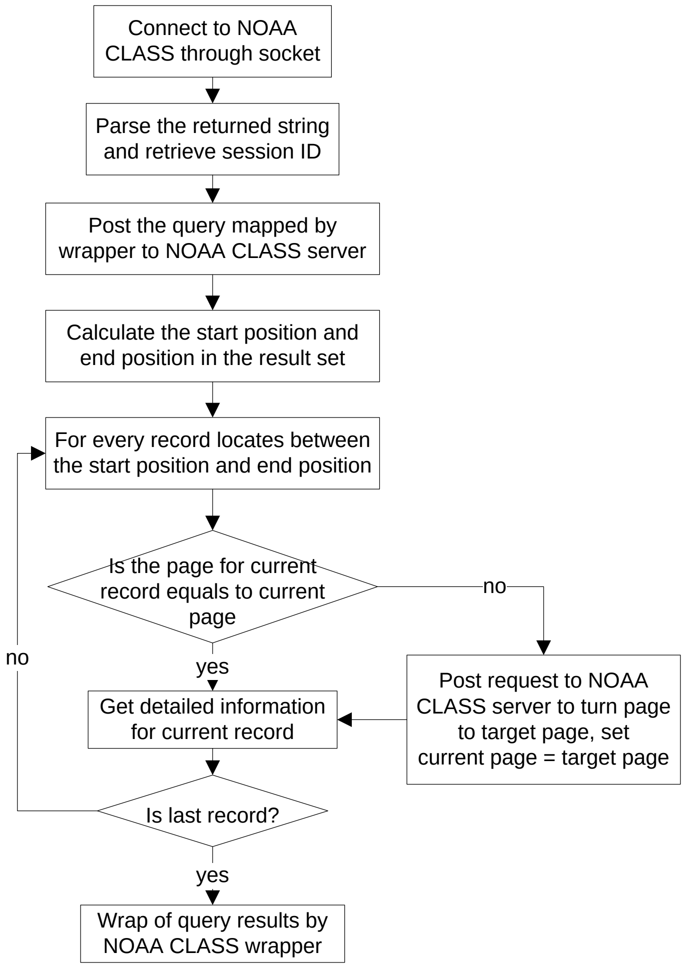
5.1.2 NOAA CLASS System

The NOAA CLASS system offers users a GUI to search their data holdings stored in its data archives. Each data archive has its own searching criteria. Some data archive, such as Microwave Integrated Retrieval System (MIRS) Daily Mapped Data may take temporal, data type, satellite and projection as searching criteria. While others may take spatial, temporal, coverage, satellite, and satellite schedule as searching criteria, such as Geostationary Operational Environmental Satellite (GOES) Satellite Data – Imager. In the NOAA CLASS system, users' requests are sent to its server through HTTP protocol after the processing of data search queries. When the required data are ready, users will receive notification emails with information about where and how to get the data. To wrap NOAA CLASS system into a standard OGC CSW service, the wrapper mapped the queries and granule level metadata contained in the search results. Using the queries in OGC CSW format from mediator, searching criteria of native dataset ID, spatial and temporal information were mapped to corresponding parameters in local queries in NOAA CLASS system (Table 10).

Table 10 Mapping betweenOGC CSW queries and NOAA CLASS queries

OGC CSW	**NOAA CLASS**
dc:subject	datatype_family
dct:coverage.dateStart	start_date, start_time
dct:coverage.dateEnd	end_date, end_time
gml:lowerCorner	slat, wlon
gml: upperCorner	nlat, elon

The system connector module communicated with NOAA CLASS catalogue by initiating a session with the NOAA CLASS server and sending queries from the wrapper to the server. An HTML web page was returned with a list of search results. However, the results needed by user might not be always on the first page. Thus, based on the start position and max records specified by user, the system connector module interpreted the values and identified the page on which the current record was on. Then it sent another request to get the target page. After that, the system connector module sent additional requests to retrieve detailed information of search results. The flowchart of system connector module for NOAA CLASS is shown as Figure 26.

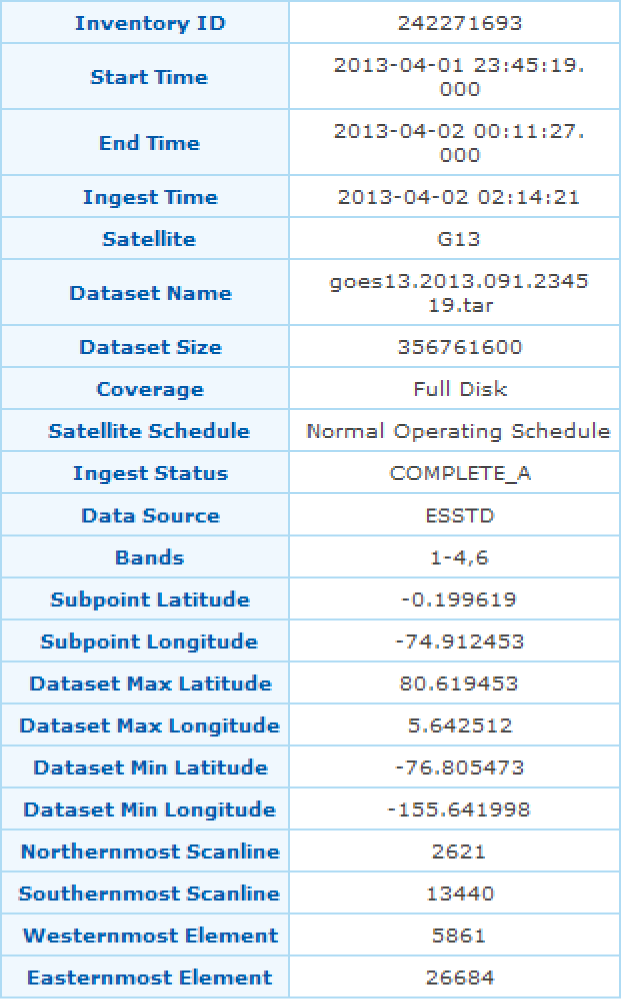


Connect to NOAA CLASS through socket
Parse the returned string and retrieve session ID
Post the query mapped by wrapper to NOAA CLASS server
Calculate the start position and end position in the result set
For every record locates between the start position and end position
Is the page for current record equals to current page
no
yes
Post request to NOAA CLASS server to turn page to target page, set current page = target page
Get detailed information for current record
Is last record?
no
yes
Wrap of query results by NOAA CLASS wrapper

Figure 26 Flowchart of system connector module for NOAA CLASS

For the NOAA CLASS system, the returned granule metadata is stored in HTML table in KVP format. Example of detailed information of a granule metadata record retrieved by the system connector in NOAA CLASS system is shown in Table 11.

Table 11 Detail information of granule in NOAA CLASS



Inventory ID	242271693
Start Time	2013-04-01 23:45:19.000
End Time	2013-04-02 00:11:27.000
Ingest Time	2013-04-02 02:14:21
Satellite	G13
Dataset Name	goes13.2013.091.234519.tar
Dataset Size	356761600
Coverage	Full Disk
Satellite Schedule	Normal Operating Schedule
Ingest Status	COMPLETE_A
Data Source	ESSTD
Bands	1-4,6
Subpoint Latitude	-0.199619
Subpoint Longitude	-74.912453
Dataset Max Latitude	80.619453
Dataset Max Longitude	5.642512
Dataset Min Latitude	-76.805473
Dataset Min Longitude	-155.641998
Northernmost Scanline	2621
Southernmost Scanline	13440
Westernmost Element	5861
Easternmost Element	26684

The wrapper module mapped the information to make them compatible with the Core and ISO profiles of OGC CSW. Take a brief record of the Core profile as an example, the mapping between OGC CSW and the NOAA CLASS information model is shown in table 12.

Table 12 Mapping between metadata information models of OGC CSW (Core profile) and NOAA CLASS

BriefRecord	**OGC CSW**	**NOAA CLASS**
	dc:identifier	datatype_family.Dataset Name
	dc:title	Dataset Name
	dc:type	Fixed String
	ows:LowerCorner	Dataset Min Longitude, Dataset Min Latitude
	ows:UpperCorner	Dataset Max Longitude, Dataset Max Latitude

With the wrapper module and the system connector module developed, the NOAA CLASS system was exposed as an OGC CSW compliant catalogue and ready for the queries dispatched by the mediator.

5.1.3 NASA ECHO

ECHO was developed by NASA to provide flexible search for NASA's EOS information and to better meet the needs of science community. It achieved this goal by 1) providing APIs for alternating user interfaces to support users' special needs in data access; 2) providing APIs for brokering data services so specialized data services can be shared across the user community; and 3) providing APIs for easy participation by a broad community of data providers (Wichmann and Pfister, 2002). The major feature of ECHO is that all interactions with it occur using XML as the base message format

(Wichmann and Pfister, 2002). This system offers a set of APIs and also provides web services for metadata repository retrieval (Wang et al., 2012). With the set of APIs provided by NASA ECHO catalogue, granule level metadata were retrieved through XML-based requests.

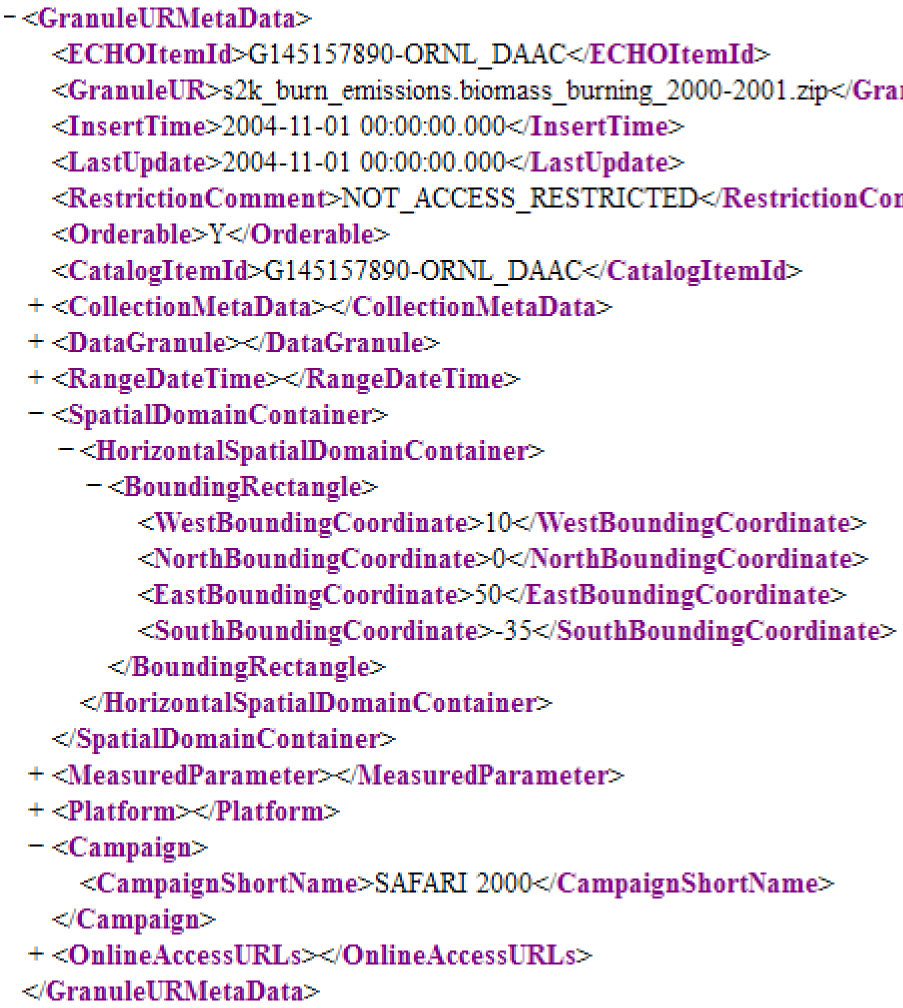
To wrap NASA ECHO into a standard OGC CSW service, the wrapper module performed mapping for both users' queries and granule level metadata contained in the search results. Based on the queries in OGC CSW format dispatched from the mediator module, the NASA ECHO catalogue wrapper module parsed the input queries as described in Figure 24. Users' searching criteria of native dataset ID, spatial, and temporal information were extracted and mapped to corresponding parameters compatible with ECHO's query format: ECHO Alternative Query Language (AQL) (Table 13).

Table 13 Mapping between CSW query format and NASA ECHO query format

OGC CSW	**NASA ECHO (AQL)**
dc:subject	<dataSetId>
dct:coverage.dateStart	<startDate><Date YYYY= ‘’ MM=’’ DD=’’></startDate>
dct:coverage.dateEnd	<stopDate><Date YYYY= ‘’ MM=’’ DD=’’></stopDate>
gml:lowerCorner	<IIMSPoint long = ‘’ lat = ‘’>
gml: upperCorner	<IIMSPoint long = ‘’ lat = ‘’>

Authentication was required before the search request was sent to the NASA ECHO catalogue through its API. The NASA ECHO system connector module logged into ECHO and then executed the AQL query through the function of

CatalogServiceLocator().getCatalogServicePort().executeQuery() in the ECHO API. After the query execution was finished, granule metadata were returned in search result. The NASA ECHO metadata model is derived directly from the EOSDIS Core System (ECS) (Mitchell et al., 2009). Example of detailed information for a NASA ECHO granule metadata record retrieved by system is shown in Figure 27.



```
- <GranuleURMetaData>
    <ECHOItemId>G145157890-ORNL_DAAC</ECHOItemId>
    <GranuleUR>s2k_burn_emissions.biomass_burning_2000-2001.zip</GranuleUR>
    <InsertTime>2004-11-01 00:00:00.000</InsertTime>
    <LastUpdate>2004-11-01 00:00:00.000</LastUpdate>
    <RestrictionComment>NOT_ACCESS_RESTRICTED</RestrictionComment>
    <Orderable>Y</Orderable>
    <CatalogItemId>G145157890-ORNL_DAAC</CatalogItemId>
  + <CollectionMetaData></CollectionMetaData>
  + <DataGranule></DataGranule>
  + <RangeDateTime></RangeDateTime>
  - <SpatialDomainContainer>
    - <HorizontalSpatialDomainContainer>
      - <BoundingRectangle>
          <WestBoundingCoordinate>10</WestBoundingCoordinate>
          <NorthBoundingCoordinate>0</NorthBoundingCoordinate>
          <EastBoundingCoordinate>50</EastBoundingCoordinate>
          <SouthBoundingCoordinate>-35</SouthBoundingCoordinate>
        </BoundingRectangle>
      </HorizontalSpatialDomainContainer>
    </SpatialDomainContainer>
  + <MeasuredParameter></MeasuredParameter>
  + <Platform></Platform>
  - <Campaign>
      <CampaignShortName>SAFARI 2000</CampaignShortName>
    </Campaign>
  + <OnlineAccessURLs></OnlineAccessURLs>
  </GranuleURMetaData>
```

Figure 27 Detailed information of granule metadata in NASA ECHO

The wrapper mapped the granule metadata from the ECS format to those defined by the Core and ISO profiles of OGC CSW. Table 14 shows a mapping table between a core profile-based brief CSW record and a record from NASA ECHO system.

Table 14 Mapping between metadata information models of OGC CSW (Core profile) and NASA ECHO

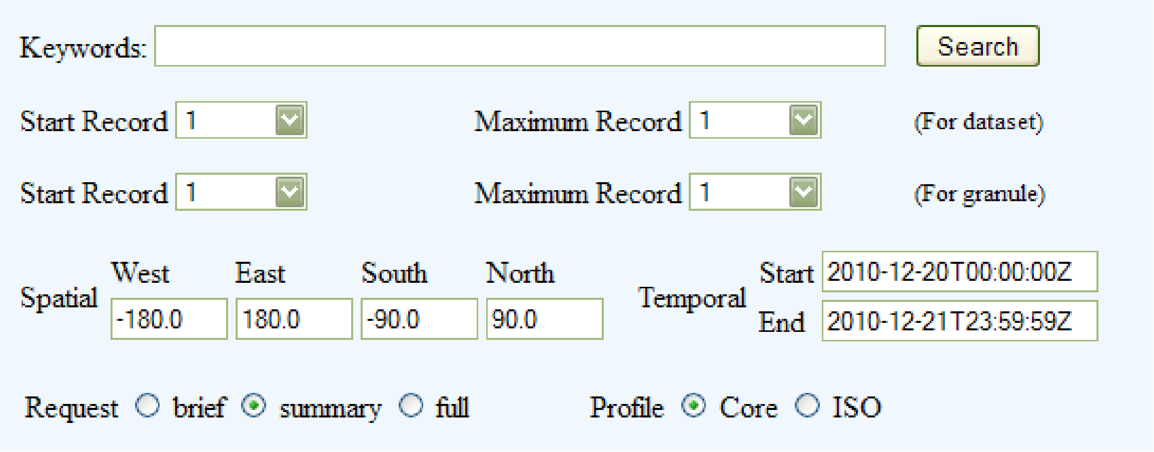
	OGC CSW	**NASA ECHO (ECS)**
BriefRecord	dc:identifier	/GranuleURMetaData/ECHOItemId
	dc:title	/GranuleURMetaData/GranuleUR
	dc:type	Fixed String
	ows:LowerCorner	/GranuleURMetaData/SpatialDomainContainer/HorizontalSpatialDomainContainer/BoundingRectangle/WestBoundingCoordinate, /GranuleURMetaData/SpatialDomainContainer/HorizontalSpatialDomainContainer/BoundingRectangle/SouthBoundingCoordinate
	ows:UpperCorner	/GranuleURMetaData/SpatialDomainContainer/HorizontalSpatialDomainContainer/BoundingRectangle/EastBoundingCoordinate, /GranuleURMetaData/SpatialDomainContainer/HorizontalSpatialDomainContainer/BoundingRectangle/NorthBoundingCoordinate

With the functions in the wrapper and system connector modules developed for NASA ECHO, the NASA ECHO catalogue was exposed as an OGC CSW-compliant catalogue and ready for queries dispatching by the mediator module.

5.2 Integrated Catalogue for Distributed EO Data Search

5.2.1 Search in Dataset Directory and Query Dispatching

As discussed above, the dataset directory was built with GeoNetwork. The mediator module interacted with GeoNetwork to apply keyword search through its “AnyText” search functions. To get search parameters from users, a GUI was developed to release users from the burden of composing OGC CSW compliant queries (Figure 28).



Keywords:

Start Record 1 Maximum Record 1 (For dataset)

Start Record 1 Maximum Record 1 (For granule)

Spatial West -180.0 East 180.0 South -90.0 North 90.0

Temporal Start 2010-12-20T00:00:00Z End 2010-12-21T23:59:59Z

Request ○ brief ● summary ○ full Profile ● Core ○ ISO

Figure 28 GUI for distributed catalogue search system

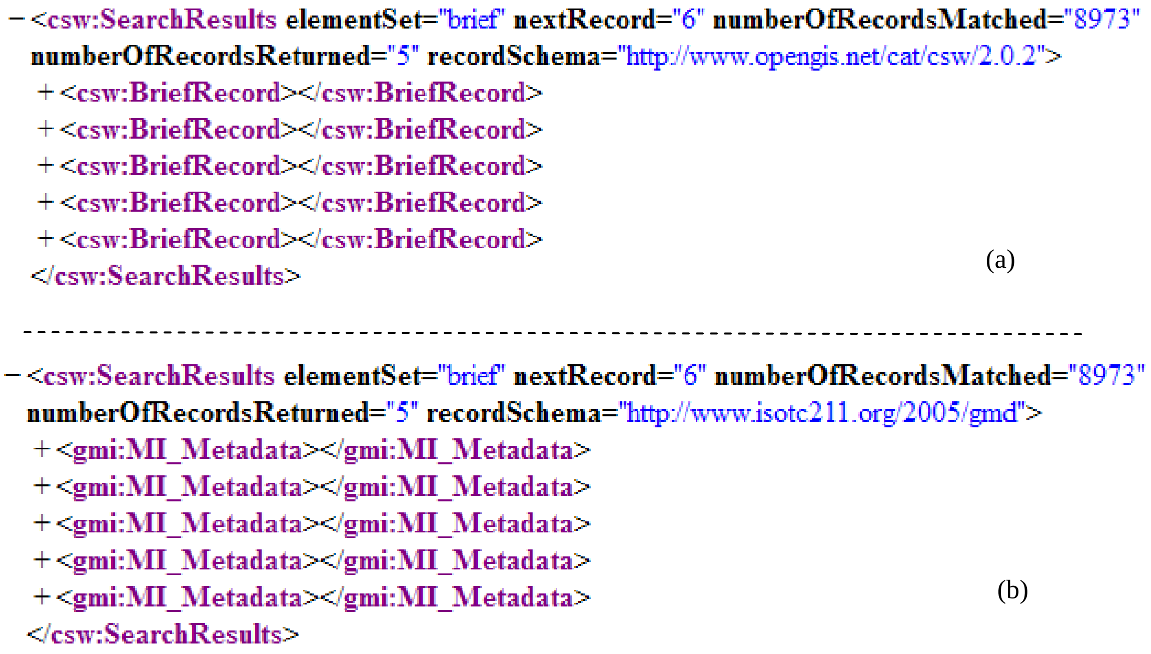
Note that the start record and maximum record in the GUI were applied at both dataset level and granule level. The distributed catalogue search system adopted the opaque pattern for query dispatching to facilitate users without domain knowledge to search EO data from the underlying NOAA CLASS and NASA ECHO catalogues. These catalogues and their interfaces were not exposed through this GUI. When search criteria were submitted to the mediator module, it followed several steps to execute the query in GeoNetwork: 1) parse the request message to extract search parameters; 2) compose an

OGC CSW-compliant query based on the search parameters extracted from step 1 and apply the keywords parameter to the "AnyText" property in it; 3) authenticate with GeoNetwork by sending a login request with username and password; 4) send the query created in step 2 to GeoNetwork to retrieve relevant dataset metadata. Since GeoNetwork uses relevance as its default sorting criteria, in the distributed catalogue search system, the first dataset is considered to be most relevant to user's query; 5) parse the results from step 4 using XPath "gmd:fileIdentifier/gco:CharacterString" for the file identifier to get a list of file IDs which are relevant to the query.

Once the list of file identifiers were retrieved, the mediator searched in the mapping file described in section 5.1.1 to get corresponding native dataset ID and catalogue ID. Together with spatial, temporal and other search criteria, several sub-queries in OGC CSW format were composed and dispatched to wrappers according to their catalogue IDs respectively.

5.2.3 Results Aggregation

With the strategies discussed in 4.4.3, through the wrapper and system connector modules, both NOAA CLASS and NASA ECHO were exposed as OGC CSW-compliant catalogues. Each granule metadata record in the sub-query result sets was extracted and used as a child node for the "csw:SearchResults" element. With this method, results from sub-queries were aggregated and returned to users. An example of the structure of aggregated records for distributed catalogue search system is shown in Figure 29: brief records in CSW Core profile (Figure 29(a)) and brief records in ISO profile (Figure 29 (b)).



```
- <csw:SearchResults elementSet="brief" nextRecord="6" numberOfRecordsMatched="8973"
  numberOfRecordsReturned="5" recordSchema="http://www.opengis.net/cat/csw/2.0.2">
   + <csw:BriefRecord></csw:BriefRecord>
   + <csw:BriefRecord></csw:BriefRecord>
   + <csw:BriefRecord></csw:BriefRecord>
   + <csw:BriefRecord></csw:BriefRecord>
   + <csw:BriefRecord></csw:BriefRecord>
  </csw:SearchResults>
```

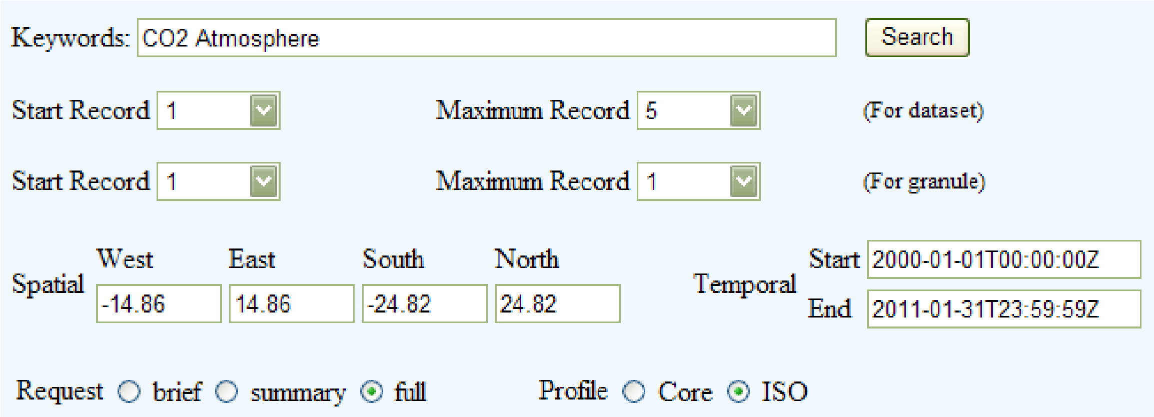
```
- <csw:SearchResults elementSet="brief" nextRecord="6" numberOfRecordsMatched="8973"
  numberOfRecordsReturned="5" recordSchema="http://www.isotc211.org/2005/gmd">
   + <gmi:MI_Metadata></gmi:MI_Metadata>
   + <gmi:MI_Metadata></gmi:MI_Metadata>
   + <gmi:MI_Metadata></gmi:MI_Metadata>
   + <gmi:MI_Metadata></gmi:MI_Metadata>
   + <gmi:MI_Metadata></gmi:MI_Metadata>
  </csw:SearchResults>
```

Figure 29 Structure of aggregated records: brief records in CSW Core profile (a) and brief records in ISO profile (b)

There are three types of element set: brief, summary, and full. For the records compatible with the OGC Core profile, the node storing metadata for each granule is the “csw:BriefRecord”, “csw:SummaryRecord” or “csw:Record” according to the type of element set, while for the ISO profile, the node is always “gmi:MI_Metadata” regardless of element set type. In summary, the integration of catalogues was achieved by wrapping distributed catalogues based on OGC CSW, building dataset directory for the datasets managed by the catalogues, dispatching queries to wrappers after analyzing them with the aid of dataset directory, and aggregating query results into final response. A search use case and its result are discussed in the following section.

5.3 Search Use Case and Results Discussion

Assuming a user wants to search EO data through keywords: “CO2” and “Atmosphere”, along with other criteria, as shown in Figure 30.



Keywords: CO2 Atmosphere

Start Record 1 Maximum Record 5 (For dataset)

Start Record 1 Maximum Record 1 (For granule)

Spatial West -14.86 East 14.86 South -24.82 North 24.82

Temporal Start 2000-01-01T00:00:00Z End 2011-01-31T23:59:59Z

Request ○ brief ○ summary ● full Profile ○ Core ● ISO

Figure 30 Searching criteria for use case

The mediator module analyzed these criteria and composed new query with GeoNetwork to get dataset records relevant to these criteria. The dataset level search in GeoNetwork showed that there were 11 records matching the query. Specified by the “Start Record” and “Maximum Records” search criteria, the first 5 dataset metadata records were retrieved and their file identifiers were extracted. Then, the mediator module extracted their corresponding catalog IDs and composed 5 sub-queries in OGC CSW format based on search criteria at the granule level search. It then dispatched them to corresponding wrappers. The underlying NOAA CLASS catalogue and NASA ECHO catalogue were wrapped as standard OGC CSW catalogues based on the method described in section 5.1.2 and 5.1.3. They thus took standard OGC CSW queries and returned OGC CSW-compliant results. In this use case, the granule level search showed that there were 3372, 78, 0, 1, 340 granule records matching their corresponding sub-queries respectively. Then based on the “Start Record” and “Maximum Records” searching criteria for granule level search, the first granule record from each dataset was retrieved (if it exists) to compose the final response to user. Then the mediator aggregated the query results by extracting each granule records in the result set and adding it as a child node of the “csw:SearchResults” element:

<csw:SearchResults elementSet="full" nextRecord="5" numberOfRecordsMatched="3791" numberOfRecordsReturned="4" recordSchema="http://www.isotc211.org/2005/gmd">. Then the aggregated query results were returned to the user.

In the whole process, users had no awareness of the underlying catalogues or their affiliated datasets. Thus they were eased from the complexity of heterogeneous interfaces, metadata models of distributed catalogues. This use case demonstrated the capabilities of interoperability enhancement and extending of granularity and scope of EO data search.

Interoperability was enhanced since the underlying NOAA CLASS catalogue and NASA ECHO catalogue which took HTTP table and ECS as metadata information models were mapped to be compatible with the Core profile and ISO profile of OGC CSW. The interfaces and query languages they used locally were mapped to the CSW format. The granularity of EO data search was extended since the dataset level search was enabled for granule level search functions provided by these two catalogues. The scope of EO data search was also extended since granules in both catalogues integrated in the distributed catalogue search system could be discovered. Besides the prototype distributed catalogue search system which demonstrated the capabilities of ICF, the Committee on Earth Observation Satellites (CEOS) Working Group on Information Systems and Services (WGISS) Integrated Catalog (CWIC) project also adopted the ICF at granule level search. Six popular catalogues including thousands of datasets were integrated in to the project to help users identify and access EO data of interest.

CHAPTER SIX CONCLUSION

This chapter summarizes the dissertation and addresses major contributions of this research on EO data search across distributed catalogues. Several issues and future work for this research are also discussed in this chapter.

6.1 Summary

In this research, a standards-based ICF is designed and developed to enhance the interoperability of EO data search across distributed catalogues. The OGC CSW specification, which had become the most widely used open standard for geospatial catalogues, was adopted in this research. OGC CSW specifies a set of interfaces between clients and catalogue services, query languages, protocol bindings, and metadata information models. Besides the Core profile of OGC CSW, the ISO application profile for OGC CSW was also applied to enhance EO data search capabilities.

The proposed ICF supports two levels of search, dataset level and granule level, to extend both granularity and scope of EO data search. In developing this framework, the mediator-wrapper architecture was adopted. The mediator module analyzes users' queries and dispatches them to corresponding wrappers; the wrapper module wraps different types of underlying distributed catalogues into OGC CSW compliant catalogues so that they can work in a plug-and-play style for the mediator.

Four modules were developed in the mediator-wrapper architecture to support search capabilities. These modules include the mediator module, the dataset directory module, the wrapper module and the system connector module. The mediator module and the dataset directory module mainly support EO data search on the dataset level. The wrapper module and the system connector module support the granule level search. The mediator module is also involved in the granule level search. The mediator module addresses query mapping, query processing, query dispatching, and granule-level search results integration from distributed catalogues. It interacts with the dataset directory to enable dataset level search. The dataset directory module is used to store dataset metadata and provides search interfaces for mediator module. The wrapper module maps queries and metadata information models for the system connector module. The system connector module interacts with distributed catalogues directly.

For dataset level search, a dataset directory was built and keyword search function was applied to retrieve relevant dataset metadata. Three methods for collecting ISO 19115 dataset metadata have been introduced: 1) collecting metadata from a registration service developed for data providers, 2) mapping dataset metadata for existing distributed catalogues, and 3) harvesting ISO 19115 metadata from GCMD. Two methods were introduced in this research to enable keyword search for dataset level search. One is to build a local metadata repository which stores the spatial, temporal, dataset descriptive text (combination of abstract and descriptive keywords sections in ISO 19115 metadata for the dataset), and other necessary information, so that the TF-IDF values for queries can be calculated and used to rank the relevance of datasets and determine the query result of keyword search. The other is to use GeoNetwork, which is an open source catalog application used to manage spatially referenced resources, to build the dataset directory, and utilize the text search functions supported by GeoNetwork to manage dataset metadata. Based on the dataset metadata returned by the dataset directory, the mediator composes sub-queries in OGC CSW format, and then dispatches the sub-queries to corresponding wrappers with the aid of mapping list file which maintains the relationship among file identifiers, native dataset identifiers, and catalogue identifiers.

For granule level search, multiple distributed catalogues were integrated in the ICF. A wrapper module and a system connector module were developed for each catalogue to be integrated in the framework. The wrapper module analyzes queries in OGC CSW format dispatched from the mediator and translates them to be compatible with corresponding catalogues. It also converts the granule metadata results returned from system connectors be compatible with the metadata information models supported by OGC CSW in this research: the Core profile and the ISO profile. The system connector module was implemented according to the interfaces provided by its corresponding catalogues.

To demonstrate the capabilities of the ICF for integrating distributed catalogues and achieving interoperability of EO data discovery, a distributed catalogue search system was developed. The NOAA CLASS catalogue and NASA ECHO catalogue were used as data sources for this system. Take a keyword search of “CO2 atmosphere” as an example, the system processes the query on the dataset level to get relevant dataset IDs to compose sub-queries in OGC CSW format for granule level search, dispatches those sub-queries to wrappers, integrates and sends results to user. The major intent of this research is not to evaluate relevance and accuracy of search results of the sample distributed catalogue system, but to demonstrate capabilities of enhancing interoperability, extending granularity and scope of EO data search in the ICF.

6.2 Contribution

This research contributes to the EO data search across distributed catalogues in the following aspects.

First, it presented a framework that could integrate distributed catalogues, supporte two levels of data search, and enhance the interoperability for distributed EO data search. This framework offers a harmonized and standard-compliant interface for users to access distributed catalogues and releases users from the complexity of underlying heterogeneous geospatial catalogues services. Most existing catalogues were developed specifically for their data providers or spatial agencies to serve granule level data, thus their search capability is limited. By standardizing and integrating these catalogues in the framework, the EO data search ability is improved; and the interoperability of these catalogues is enhanced.

Second, a solution for integrating distributed catalogues was provided to extend the scope of EO data search. The mediator-wrapper architecture was adopted in this research to manage query dispatching and standardization of distributed catalogues. Existing distributed catalogues use different query interfaces and information models to manage their data products; most of the metadata are customized by spatial agencies due to the heterogeneity of the EO data they collected. To hide users from various heterogeneous query interfaces, the wrapper wraps queries for distributed catalogues based on OGC CSW so that they can offer standard CSW interface for EO data search. To hide users from various heterogeneous metadata information models adopted by different catalogues, the wrapper wraps the query results based on the application profiles supported by OGC CSW, such as CSW Core profile and ISO profile in this research so that interoperability is achieved by standardization of distribute catalogues. Once catalogues are standardized, they can be easily integrated into to the framework and ready to receive queries from the mediator thus the scope for EO data search in enlarged.

Third, strategies for enabling dataset level search for distributed catalogues serving granule level data were provided to extend the granularity of EO data search. The concept of these strategies is first collecting dataset information to build a local dataset directory, on top of which the dataset level search is applied. In this research, the dataset information is mapped to ISO 19115 to achieve interoperability. The standardized dataset metadata can be easily recognized and reused by other catalogues and applications. To help users without much domain knowledge, keyword search function was developed for dataset level search in the ICF proposed in this research. Two concrete methods were provided in this research to support keyword search. Since most existing catalogues serves granule level data, the granularity of search is extended by adding dataset level search functionality for them.

6.3 Future Work

Some future work is needed to further improve the ICF. The current framework only includes keyword search functionality on the dataset level. Also, this search process is currently only based on the static keywords match without the full exploration of underlying semantics. For example if “water” is used as the keyword, then data products for “ocean”, “sea” and other “water”-related data should also be considered as relevant search results under certain circumstances. The relationship among those keywords, or ontology, used in EO data search needs to be built. Semantic query processing methods can be added to extend the integrated catalogue search function and to improve the discovery capability of EO data search.

When integrating heterogeneous distributed catalogues, there is always a limitation in the query mapping, that is, query mismatching. The ICF offers a set of common queryable properties defined in the OGC CSW query interfaces, which can promote the interoperability of EO data search. However, these properties cannot always match the query interfaces provided by the underlying catalogues exactly, thus the advanced search features provided by each individual catalogue will not be preserved. Take the NOAA CLASS for example, for some datasets, users can specify from which satellite their required data is. While through ICF, users cannot specify this parameter in the search criteria. Another example is that ICF provides interfaces for users to specify spatial and temporal search criteria while the underlying catalogues might not support these search criteria. Due to the heterogeneity in distributed catalogues, query mismatching and the balance between interoperability and features remain as open issues in the integration process.

The ICF provides two methods to fulfill the keyword search function. The second method is to use the text search functions provided by GeoNetwork directly. GeoNetwork uses Lucene accuracy for searches to determine the similarity between the query and search results. According to the mapping table for Lucene searchable fields to CSW queryable elements, any property of type “gco:CharacterString” will be searched when “AnyText” search is applied. To further refine the search area and make the relevance of search results measurable, future investigation in Lucene is needed.

The distributed catalogue search system developed in this research can be further expanded by integrating more catalogues to it. To fulfill this object, corresponding system connectors and wrappers need to be developed for each new catalogue and the mediator component need to be adjusted with the updates by adding the new wrapper into its dispatching targets. A registration service is needed for data providers to offer information about their data products to facilitate the process of collecting ISO 19115 dataset metadata into the dataset directory.

The opaque pattern of query dispatching was adopted in the distributed catalogue search system to help users without domain knowledge to search EO data. In the opaque pattern, users have no awareness of underlying catalogues. For users who already know which catalogues and datasets their required data can be searched from, the transparent pattern of query dispatching can be exploited by exposing interfaces to interact with underlying distributed catalogues. OGC CSW-compliant interfaces for underlying catalogues are supported in the ICF by the functions of wrappers. The GUI of the distributed catalogue search system can be further extended by exposing these interfaces to users. Other future investigations of distributed catalogue search of EO data can be made based on the ICF proposed in this research.

REFERENCES

REFERENCES

Alameh, N., (2003). *Service Chaining of Interoperable Geographic Information Web Services*. IEEE Internet Computing 7 (5): 22-59.

Aditya, B., Bhalotia, G., Chakrabarti, S., Hulgeri, A., Nakhe, C., Parag, P., & Sudarshan, S. (2002, August). *Banks: Browsing and keyword searching in relational databases. In Proceedings of the 28th international conference on Very Large Data Bases* (pp. 1083-1086). VLDB Endowment.

Agrawal, S., Chaudhuri, S., & Das, G. (2002). *DBXplorer: A system for keyword-based search over relational databases.* In Data Engineering, 2002. Proceedings. 18th International Conference on (pp. 5-16). IEEE.

Alameh, N., 2003. *Service Chaining of Interoperable Geographic Information Web Services.* IEEE Internet Computing 7 (5): 22-59.

Amato, G., Bolettieri, P., Gennaro, C., & Rabitti, F. (2013). *Quick and Easy Implementation of Approximate Similarity Search with Lucene.* In Digital Libraries and Archives (pp. 163-171). Springer Berlin Heidelberg.

Bai, Y., L. Di, A. Chen, Yang L., Y. Wei, (2007), *Towards a Geospatial Catalogue Federation Service.* Photogrammetric Engineering & Remote Sensing 73 (6): 699–708.

Bakhtouchi, A., Chakroun, C., Bellatreche, L., & Aït-Ameur, Y. (2012). *Mediated Data Integration Systems Using Functional Dependencies Embedded in Ontologies.* Recent Trends in Information Reuse and Integration, 227-256.

Baldini, A., Boldrini, E., Santoro, M., & Mazzetti, P. (2010, May). *GeoNetwork powered GI-cat: a geoportal hybrid solution.* In EGU General Assembly Conference Abstracts (Vol. 12, p. 11436).

Beneventano, D., Gennaro, C., Bergamaschi, S., & Rabitti, F. (2011). *A mediator-based approach for integrating heterogeneous multimedia sources.* Multimedia Tools and Applications, 1-24.

Bergamaschi, S., Domnori, E., Guerra, F., Lado, R. T., & Velegrakis, Y. (2011, June). *Keyword search over relational databases: a metadata approach.* In Proceedings of the 2011 international conference on Management of data (pp. 565-576). ACM.

Bigagli, L., Nativi, S., Mazzetti, P., & Villoresi, G. (2005). *Mediation to deal with information heterogeneity.* In Geophysical Research Abstracts (Vol. 7, p. 06509).

Bishr, Y. 1998. *Overcoming the semantic and other barriers to G1S interoperability.* International Journal of Geographical Information Science 12 (4): 299-314

Carvalho, F. G., Trevisan, D. G., & Raposo, A. (2012). *Toward the design of transitional interfaces: an exploratory study on a semi-immersive hybrid user interface.* Virtual Reality, 1-18.

Chang, Y., Cheng, H., (2010). *A Metadata Classification Assisted Scientific Data Extraction Architecture.* Advances in Grid and Pervasive Computing 6140/2010: 679-688

Chen, A., Di, L., Wei, Y., Liu, Y., Bai, Y., Hu, C., & Mehrotra, P. (2005). Grid *Computing enabled geospatial catalogue web service.* American Society for Photogrammetry and Remote Sensing 2005, 7-11.

Chen, A., Di. L., Bai, Y., Wei, Y., & Liu, Y. (2010). *Grid computing enhances standards-compatible geospatial catalogue service.* Computers & Geosciences 36, Issue 4: 411-421

Chen, A., L. Di, Y. Bai, Y. Wei, (2007). *Grid-enabled Web Services for Geospatial Interoperability.* American Geophysical Union (AGU) 2006 Joint Assembly. May 23-26, 2006. Baltimore, Maryland, USA.

DCMI, 2010. DCMI Metadata Terms,
URL: http://dublincore.org/documents/dcmi-terms/, Dublin Core Metadata Initiative (Date issued 2010-10-11).

de Andrade, F. G., Baptista, C. D. S., & Leite Jr, F. L. (2011). *Using Federated Catalogs to Improve Semantic Integration among Spatial Data Infrastructures.* Transactions in GIS, 15(5), 707-722.

Di, L., & Ramapriyan, H. K. (2010). *Standards-based data and information systems for Earth observations–An introduction.* Standard-Based Data and Information Systems for Earth Observation, 1-6.

Dragicevic, S. (2004). *The potential of Web-based GIS.* Journal of Geographical Systems, 6(2), 79-81.

Fielding, R. T., Gettys, J., Mogul, J. C., Nielsen, H. F., Masinter, L., Leach, P., and Berners-Lee, T. (1997). *Hypertext transfer protocol---HTTP/1.1.* Internet RFC 2616, Jan. 1997.

Fuger, S., F. Najmi, N. Stojanovic, (2005). *OASIS ebXML Registry Information Model Version 3.0,* http://docs.oasis-open.org/regrep/v3.0/specs/regrep-rim-3.0-os.pdf. *OASIS Standard*, May 2nd, 2005

Ganeshan, K. V., Sarda, N. L., & Gupta, S. (2010, November). *Keyword search in geospatial database.* In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems (pp. 538-539). ACM.

Gao, R., Li, D., Li, W., & Dong, Y. (2012). *Application of Full Text Search Engine Based on Lucene.* Advances in Internet of Things, 2(4), 0-0.

Garcia-Molina, H., Papakonstantinou, Y., Quass, D. Rajaraman, A., Sagiv, Y., Ullman, J. ,Vassalos, V., Widom, J., (1997). *The TSIMMIS Approach to Mediation: Data Models and Languages.* Journal of Intelligent Information Systems 8 (2): 117-132.

GeoNetwork (2013) *GeoNetwork developer manual* http://geonetwork-opensource.org/manuals/2.8.0/eng/developer/GeoNetworkDeveloperManual.pdf

GeoNetwork (2013) *GeoNetwork user manual* http://geonetwork-opensource.org/manuals/2.8.0/eng/users/GeoNetworkUserManual.pdf

Govedarica, M., Bošković, D., Petrovački, D., Ninkov, T., Ristić, A., (2010). *Metadata Catalogues in Spatial Information Systems,* Geodetski list 64 (4): 313-334.

Gray, J. (1996), *Data Management: Past, Present, and Future.* Technical Report MSR-TR-96-18. IEEE computer 29(10): 38-46

Grill, S., & Schneider, M. (2009). *Geonetwork open source as an application for SDI and education.* GIS Ostrava, 25-28.

Heery, R. and M. Patel, (2000). *Application profiles: mixing and matching metadata schemas,* Ariadne, issue 25, 2000.
URL: http://www.ariadne.ac.uk/issue25/app-profiles/intro.html

Hoschek, W., Martinez, J. J., Samar A., Stockinger, H., and Stockinger, K. (2000). *Data Management in an International Data Grid Project.* Lecture Notes in Computer Science Volume 1971, 2000, pp 77-90

Hristidis, V., & Papakonstantinou, Y. (2002, August). *Discover: Keyword search in relational databases.* In Proceedings of the 28th international conference on Very Large Data Bases (pp. 670-681). VLDB Endowment.

Hua, H., & Weiss, B. (2011, December). *Strategies for Infusing ISO 19115 Metadata in Earth Science Data Systems.* In AGU Fall Meeting Abstracts (Vol. 1, p. 04).

Hubner, S., Spittel, R., Visser, U., and Vogele, T.J.,(2004). *Ontology-based search for interactive digital maps.* Intelligent Systems, IEEE 19 (3) 80-86

Hulgeri, A., Bhalotia, G., Nakhe, C., Chakrabarti, S., & Sudarshan, S. (2001). *Keyword search in databases.* IEEE Data Engineering Bulletin, 24(3), 22-31.

Jafari, A., P. Padmanabh, G. banhazl, D. Rapka, F. Lee, K. Andapally, (2010). *A web-based renewable energy monitoring and management system.* Proceedings of the 12th International Conference on Information Integration and Web-based Applications & Services 2010. NY, USA.

Jones, K. S. (1972). *A statistical interpretation of term specificity and its application in retrieval.* Journal of documentation, 28(1), 11-21.

Kojima, I., Kimoto, M., & Matono, A. (2010, February). *OGC catalog service for heterogeneous earth observation metadata using extensible search indices.* In Proceedings of the 6th Workshop on Geographic Information Retrieval (p. 9). ACM.

Kumar, M., & Vig, R. (2013). *Focused Crawling Based Upon TF-IDF Semantics and Hub Score Learning.* Journal of Emerging Technologies in Web Intelligence, 5(1), 70-77.

Landrau, V. M. C. (2002). *Accessing geospatial data to support research, monitoring and environmental management activities.* Master thesis, International Institute for Geo-Information Science and Earth Observation Enschede, the Netherlands. http://www.itc.nl/library/papers/msc_2002/gim/cuadrado_landrau.pdf

Langegger, A., Wöß, W., & Blöchl, M. (2008). *A semantic web middleware for virtual data integration on the web.* The Semantic Web: Research and Applications, 493-507.

Laurini, R. (1998). *Spatial multi-database topological continuity and indexing: a step towards seamless GIS data interoperability.* International Journal of Geographical Information Science 12 (4): 373-402

Li, W., Yang, C., & Raskin, R. (2008). *A semantic enhanced search for spatial web portals.* In AAAI Spring Symposium Technical Report, SS-08-05 (pp. 47-50).

Liu, X., Maly, K., Zubair, M., Hong, Q., Nelson, M. L., Knudson, F., & Holtkamp, I. (2006). *Federated searching interface techniques for heterogeneous OAI repositories.* Journal of Digital information, 2(4).

Luo, Y., Lin, X., Wang, W., & Zhou, X. (2007, June). *Spark: top-k keyword query in relational databases.* In International Conference on Management of Data: Proceedings of the 2007 ACM SIGMOD international conference on Management of data (Vol. 11, No. 14, pp. 115-126).

Lynden, S.J., Pahlevi, S.M., Kojima, I., (2008). *Service-based data integration using OGSA-DQP and OGSA-WebDB.* In: GRID 2008, 160–167

Manso, M. Á., Wachowicz, M., & Bernabé, M. Á. (2009). *Towards an integrated model of interoperability for spatial data infrastructures.* Transactions in GIS, 13(1), 43-67.

Marlin, J. W., Knudson, R. L., Ruehle, T. M., Stuart, A. F., & Hughes III, E. T. (1998). *Table driven graphical user interface. U.S. Patent No. 5,778,377.* Washington, DC: U.S. Patent and Trademark Office.

McCurley, K. S. (2001, April). *Geospatial mapping and navigation of the web.* In Proceedings of the 10th international conference on World Wide Web (pp. 221-229). ACM.

Miller, L.L., and S. Nusser, (2003). *An Infrastructure for Supporting Spatial Data Integration.* Federal Committee on Statistical Methodology Conference, Novemeber 17-19. Arlinton, Virginia

Mitchell, A., Ramapriyan, H., & Lowe, D. (2009, July). *Evolution of web services in EOSDIS—Search and order metadata registry (ECHO).* In Geoscience and Remote Sensing Symposium, 2009 IEEE International, IGARSS 2009 (Vol. 5, pp. V-371). IEEE.

Myers, B., Hudson, S. E., & Pausch, R. (2000). *Past, present, and future of user interface software tools.* ACM Transactions on Computer-Human Interaction (TOCHI), 7(1), 3-28.

Nativi, S., & Bigagli, L. (2009). *Discovery, mediation, and access services for earth observation data.* Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of, 2(4), 233-240.

Naumann, F., Leser, U., Freytag, J. C., (1999). *Quality-driven Integration of Heterogeneous Information Systems.* Proceedings of the International Conference on Very Large Data Bases 25: 447-458

Nebert, D., Whiteside, A., and Vretanos P., (2007). *OpenGIS Catalog services implementation specification*, Version 2.0.2, OGC 07-006r1.

Nogueras-Iso, J. (2005). *OGC catalog services: a key element for the development of spatial data infrastructures.* Computers & Geosciences, 31(2): 199–209.

Olfat, H., Kalantari, M., Rajabifard, A., Senot, H., & Williamson, I. P. (2012). *Spatial Metadata Automation: A Key to Spatially Enabling Platform.* International Journal of Spatial Data Infrastructures Research, 7, 173-195.

Özsu, M. T., & Valduriez, P. (2011). *Principles of distributed database systems.* Springer.

Ramos, J. (2003, December). *Using tf-idf to determine word relevance in document queries.* In Proceedings of the First Instructional Conference on Machine Learning.

Salton, G., & Buckley, C. (1988). *Term-weighting approaches in automatic text retrieval.* Information processing & management, 24(5), 513-523.

Segev, A., & Toch, E. (2009). *Context-based matching and ranking of web services for composition.* Services Computing, IEEE Transactions on, 2(3), 210-222.

Senkler, K., Uwe V., Albert R., (2004). *An ISO 19115/19119 Profile For OGC Catalogue Service CSW 2.0. 10th EC GI & GIS Workshop,* June 23-25, 2004, Warsaw, Poland.

Shao, Y., Warnock, A., & Kang, L. (2012, February). *CWIC data partner guide* http://www.ceos.org/images/WGISS/CWIC/CWIC%20Data%20Partner%20Guide.doc

Shen, S., Zhang, T., Wu, H., & Liu, Z. (2012). *A Catalogue Service for Internet GIServices Supporting Active Service Evaluation and Real ‐ Time Quality Monitoring.* Transactions in GIS, 16(6), 745-761.

Shvaiko, P., Ivanyukovich, A., Vaccari, L., Maltese, V., and Farazi, F., (2010). *A semantic geo-catalogue implementation for a regional SDI,* in Proceedings of the INSPIRE conference, 2010.

Singh, G., Bharathi, S., Chervenak, A., Deelman, E., Kesselman, C., Manohar, M., Patil, S., Pearlman, L., (2003). *A Metadata Catalog Service for Data Intensive Applications,* Supercomputing, 2003 ACM/IEEE conference: 33

Suwanmanee, S., Benslimane, D., & Thiran, P. (2005, March). *OWL-based approach for semantic interoperability. In Advanced Information Networking and Applications, 2005.* AINA 2005. 19th International Conference on (Vol. 1, pp. 145-150). IEEE.

Tata, S., & Lohman, G. M. (2008, June). *SQAK: doing more with keywords.* In Proceedings of the 2008 ACM SIGMOD international conference on Management of data (pp. 889-902). ACM.

Tateishi, R., Sumantyo, S., & Tetuko, J. (2012, July). *Development of geospatial data sharing/overlay system-CEReS Gaia.* In Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International (pp. 558-561). IEEE.

Vaccari, L., Shvaiko, P., & Marchese, M. (2009). *A geo-service semantic integration in spatial data infrastructures.* Journal of Spatial Data Infrastructures Research, 4, 24-51.

Visser, U., Stuckenschmidt, H., (2002). *Interoperability in GIS - Enabling Technologies.* In: Proc. 5th AGILE Conference on Geographic Information Science (2002): 291-297

Voges, U. and Senkler, K., (2007). *OpenGIS Catalogue Services Specification 2.0.2 - ISO Metadata Application Profile,* version 1.0, OGC 07-045.

Wang, H., Shao, Y., Di, L., & Kang, L. (2012, August). Discovery of agriculture data using federated catalogue service. In Agro-Geoinformatics (Agro-Geoinformatics), 2012 First International Conference on (pp. 1-5). IEEE.

Wang, Y. (2012, April). *Research on web data integration framework based on cloud computing.* In Consumer Electronics, Communications and Networks (CECNet), 2012 2nd International Conference on (pp. 2823-2826). IEEE.

Wei, Y. Di, L. Zhao, B. Liao, G. Chen, A. Bai, Y. Liu, Y., (2005).*The Design and Implementation of a Grid-enabled Catalogue Service.* International Geoscience and Remote Sensing Symposium 6: 4224

Wichmann, K., & Pfister, R. (2002, June). *ECHO–a message-based framework for metadata and service management*. In Earth Science Technology Conference, Pasadena, CA.

Wiederhold, G. (1999). *Mediation to deal with heterogeneous data sources.* Interoperating Geographic Information Systems, 1-16.

Wiederhold, G., (1992). *Mediators in the architecture of future information systems.* IEEE Computer 25 (3):3849

Wu, H. C., Luk, R. W. P., Wong, K. F., & Kwok, K. L. (2008). *Interpreting tf-idf term weights as making relevance decisions.* ACM Transactions on Information Systems (TOIS), 26(3), 13.

Yang, C., Wong, D. W., & Miao, Q. (2011). *Advanced geoInformation science*, CRC Press (2010), 608p.s

Yang, R., Ramapriyan, H. K., & Meyer, C. B. (2011). *Data Access and Data Systems. Advanced Geoinformation Science,* C. Yang, et aI., Eds., ed: CRC Press, 201(1), I27-I37.

Yue, P., Di, L., Zhao, P., Yang, W., Yu, G., & Wei, Y. (2006, July). *Semantic augmentations for geospatial catalogue service.* In Proceedings of the 2006 IEEE International Geoscience and Remote Sensing Symposium (IGARSS06) Vol. 31, pp. 3486-3489.

Yue, P., Gong, J., Di, L., He, L., & Wei, Y. (2011). *Integrating semantic web technologies and geospatial catalog services for geospatial information discovery and processing in cyberinfrastructure.* Geoinformatica, 15(2), 273-303.

CURRICULUM VITAE

Huilin Wang received her Bachelor of Science in Software Engineering from Wuhan University in 2006.