International Journal of Geo-Information

MDPI

Article

Advanced Cyberinfrastructure to Enable Search of Big Climate Datasets in THREDDS

Juozas Gaigalas, Liping Di * and Ziheng Sun

Center for Spatial Information Science and Systems, George Mason University, Fairfax, VA 22030, USA; juozasgaigalas@gmail.com (J.G.); zsun@gmu.edu (Z.S.)

* Correspondence: ldi@gmu.edu

Received: 30 September 2019; Accepted: 31 October 2019; Published: 2 November 2019

Abstract: Understanding the past, present, and changing behavior of the climate requires close collaboration of a large number of researchers from many scientific domains. At present, the necessary interdisciplinary collaboration is greatly limited by the difficulties in discovering, sharing, and integrating climatic data due to the tremendously increasing data size. This paper discusses the methods and techniques for solving the inter-related problems encountered when transmitting, processing, and serving metadata for heterogeneous Earth System Observation and Modeling (ESOM) data. A cyberinfrastructure-based solution is proposed to enable effective cataloging and two-step search on big climatic datasets by leveraging state-of-the-art web service technologies and crawling the existing data centers. To validate its feasibility, the big dataset served by UCAR THREDDS Data Server (TDS), which provides Petabyte-level ESOM data and updates hundreds of terabytes of data every day, is used as the case study dataset. A complete workflow is designed to analyze the metadata structure in TDS and create an index for data parameters. A simplified registration model which defines constant information, delimits secondary information, and exploits spatial and temporal coherence in metadata is constructed. The model derives a sampling strategy for a high-performance concurrent web crawler bot which is used to mirror the essential metadata of the big data archive without overwhelming network and computing resources. The metadata model, crawler, and standard-compliant catalog service form an incremental search cyberinfrastructure, allowing scientists to search the big climatic datasets in near real-time. The proposed approach has been tested on UCAR TDS and the results prove that it achieves its design goal by at least boosting the crawling speed by 10 times and reducing the redundant metadata from 1.85 gigabytes to 2.2 megabytes, which is a significant breakthrough for making the current most non-searchable climate data servers searchable.

Keywords: climate science; metadata; web cataloging service; big geospatial data; geospatial cyberinfrastructure

1. Introduction

Cyberinfrastructure plays an important role in today's climate research activities [1–6]. Climate scientists search, browse, visualize, and retrieve spatial data using web systems on a daily basis, especially as data volumes from observation and model simulation grow to large amounts that personal devices cannot hold entirely [7,8]. The big data challenges of volume, velocity, variety, veracity, and value (5Vs), have pushed geoscientific research into a more collaborative endeavor that involves many observational data providers, cyberinfrastructure developers, modelers, and information stakeholders [9]. Climate science has developed for decades and produced tens of petabytes of data products, including stationary observations, hindcast, and reanalysis, which are stored in distributed data centers in different countries around the globe [10]. Individuals or small

groups of scientists face big challenges when they attempt to efficiently discover the data they require. Currently, most scientists acquire their knowledge about datasets via conferences, colleague recommendations, textbooks, and search engines. They become very familiar with the datasets they use, and every time they want to retrieve the data, they go directly to the dataset website to download the data falling within the requested time and spatial windows. However, these routines are less sustainable as the sensors/datasets become more varied, models evolve more frequently, and new data pertaining to their research is available somewhere else [9].

In most scenarios, metadata is the first information that researchers see, before they access and use the actual Earth observation and modeling data the metadata describes [11,12]. Based on the metadata, they decide whether or not the actual data will be useful in their research. For big spatial data, metadata is the key component backing up all kinds of users' daily operations, such as searching, filtering, browsing, downloading, displaying, etc. Currently, two of the fundamental problems in accessing and using big spatial data are the volume of metadata and the velocity of processing metadata [13,14]. Through manual investigation of Unidata THREDDS data repository (a metadata source we take as an example of typical geodata storage patterns) [15,16], it reveals that most of the metadata are highly redundant. The vast majority of metadata records contain identical information and only key fields representing spatial and temporal characteristics are regularly updated. However, there exists a regular pattern to how the redundant information is structured and how new information is added to the repository—but, the pattern varies according to data organization hierarchy and changes with the type of data being delivered (for example: Radar station vs. satellite observation vs. regular forecast model output).

To overcome these big data search challenges, we must confront practical problems in the information model, information quality, and technical implementation of information systems. Our study follows the connection between fundamental scientific challenges and existing implementations of geoscience information systems. This study aims to build a cataloging model capable of fully describing real-time heterogeneous metadata whilst simultaneously reducing data volume and enabling search within big Earth data repositories. This model can be used to efficiently represent redundant data in the original metadata repository and to perform lossless compression of information for lightweight efficient storing and searching. The model can shrink the huge amount of metadata (without sacrificing information complexity or variety available in the original repositories) and reduce the computational burden on searching among them. The model defines two types of objects: Collections and granules. It also defines their lifecycle and relationship to the upstream THREDDS repository data. Collection contains content metadata (title, description, authorship, variable/band information, etc.). Each collection contains one or more granules. Each granule contains only the spatiotemporal extent metadata. We prototyped the model as an online catalog system within EarthCube CyberConnector [17–20]. We have made the final system available online at: http://cube.csiss.gmu.edu/CyberConnector/web/covali. The system provides a near real-time replica of the source catalog (e.g., THREDDS), optimizes the metadata storage, and enables searching capability which was not available before. The system is like a clearinghouse with its own metadata database. Currently, the system is mainly used for searching operational time-series observations/simulations collected/derived from field sensors. Other datasets, like remote sensing datasets and airborne datasets, can be foreseen to be supported in the near future. The novelty of this research is that it turns the legacy data center repositories into lightweight flexible catalog services, which are more manageable by providing searching capabilities for petabytes of datasets. The work provides important references to people operating the operation of big climate data centers and advises on further improvements in those operational climate data centers to better serve the climate science community. This paper is organized as follows. Section 2 describes the background knowledge and history. Section 3 introduces related work. Section 4 introduces the proposed model. Section 5 shows the implementation of the model and the required cyberinfrastructure. Section 6 demonstrates the experiment results. Section 7 discusses the results of our approach. Section 8 concludes the paper.

The study described in this paper is an attempt to contribute to the global scientific endeavor on understanding and predicting the impacts of climate change. Understanding climate change and its impacts requires understanding Earth as a complex system of systems with behaviors that emerge from the interaction and feedback loops that occur on a range of temporal and spatial scales. However, new advances in these studies are obstructed by the challenges of interdisciplinary collaboration and the difficulty of data and information collaboration [21–27]. The difficulties of information collaboration can be understood in terms of long-standing big data problems of variety (complexity) and volume/velocity.

2. Background

Metadata is a powerful tool for dealing with big data challenges. We discuss the background work on metadata and interoperability of metadata catalogs as critical components of advanced cyberinfrastructure that we envision.

2.1. Metadata

The topic of metadata has been approached by two distinct scholarly traditions. Understanding them helps us clarify our approach to metadata in cyberinfrastructure. Library information scientists have described the metadata bibliographic control approach. Bibliographic principles allow information users to describe, locate, and retrieve information-bearing entities. The basic metadata unit is the "information surrogate" that derives its usefulness from being locatable (by author, title, and subject), accurately describing the information object (the data of the metadata) and identifying how to locate the object. The second (complementary) view of metadata originates in the computer science discipline and is called the data management approach. Complex and heterogeneous data (textual, graphical, relational, etc.) is not separated into information units, but is instead described by data models and architectures that represent "additional information that is necessary for data to be useful" [27]. The key difference is the bibliographic approach works with distinct information entities of limited types, while the data management approach works with models of data/information structures and their relationships.

This distinction between bibliographic and data management approaches is important in the context of ongoing efforts of metadata standardization [28–31]. The second approach is not conducive for standardization because the data management models are as complex and heterogeneous as the structures of the data being modeled. Consequently, in accordance with existing standards, the currently available metadata for large climate datasets follows the first approach, which provides bibliographic information and does not describe the data structures in a way that may permit new capacities of advanced cyberinfrastructure. Our paper describes the work to supplement and transform the existing bibliographical metadata with a custom metadata management model resulting in new applications for the existing data. Metadata standardization is a prerequisite for interoperability, which is a prerequisite for building distributed information systems capable of handling complex Earth system data [32].

2.2. Interoperability, Data Catalogs, Geoinformation Systems, and THREDDS

Data and information collaboration across disciplines is critical for advanced Earth science. Unfortunately, there is no strongly unified practice for data recording, storage, transmission, and processing that the entire scientific community follows [33–37]. Disparate fields and traditions have their own preferred data formats, software tools, and procedures for data management. However, Earth system studies generally work with data that follow a geospatial–temporal format [38–42]. All of the data can be meaningfully stored on a 4D (3 spatial and one temporal) dimension grid. This basic commonality has inspired standardization efforts with the goal of enabling wider interoperability and collaboration.

Following organic outgrowth from the community, the standardization efforts are now headed by Open Geospatial Consortium (OGC) and the International Organization for Standardization Technical Committee 211 (ISO TC 211) and have yielded successful standards in two areas relevant to us [43–47]. First is the definition of NetCDF as one of the standard data formats for storing geospatial data. The second is the metadata standardization. Those efforts are extremely relevant to our research and are further discussed in the Related Works section. For background, it is important to mention that the standard geospatial metadata models developed by OGC are still evolving capabilities for describing the heterogeneous, high-volume, or high-velocity big data we are studying. The commonly used OGC/ISO 19* series metadata standards have relatively limited relational features (aggregation only) and, in the repository we studied, each XML encoded metadata record contains mostly redundant information (for example, two metadata objects that represent two images from a single sensor mostly contain duplicated information that describes sensor characteristics). However, there are multiple lines of work that ISO TC 211 is pursuing that addresses these issues and suggests a trend for expanding the applicability of standardized metadata models and the integration of a greater variety of information.

The standard geographical metadata model was developed in conjunction with a standard distributed catalog registration model titled Catalog Services for the Web (CSW) [48,49]. The CSW standard is widely known and many Earth system data providers offer some information about their data holdings via the CSW interface. However, the CSW standard is also poorly suited to support big data collaborative studies for the Earth system. CSW follows the basic OGC metadata model in a way that makes it challenging to capture valuable structure and semantics of existing data holdings without storing extremely redundant information—which exhausts computing resources without taking advantage of the true value of large and complex Earth big data. However, OGC metadata stored in CSW is the existing standard that governs not only data distribution practices, but also how researchers think about data collaboration.

The next item this study works with is the UCAR Unidata THREDDS Data Server (TDS). The University Corporation for Atmospheric Research (UCAR) Unidata is a geoscience data collaboration community of diverse research and educational institutions. It provides the real-time heterogeneous Earth system data that this study targets. THREDDS is Unidata's Thematic Real-Time Environmental Distributed Data Services. TDS is a web server that provides metadata and data access for scientific datasets to climate researchers. TDS provides its own rudimentary hierarchical catalog service that is not searchable and does not support the CSW standard. However, it does support the OGC geospatial metadata standard—although not consistently or comprehensively. In order to make data hosted by TDS searchable, the TDS metadata must be copied to another server and a searchable catalog must be created for the metadata. This task is performed by a customized web crawler developed by this study.

This study attempts to build upon the existing infrastructure with its available resources and limitations to provide new capabilities. The limitations of the existing systems are two-fold. First is the limits of the CSW metadata registration model (it does not naturally support registering information about metadata lifecycle or sufficiently detailed aggregation information), and second is the incompleteness of information within metadata provided by THREDDS. This study attempts to erase the limits by first interpolating information to improve the quality of the existing metadata model and then by extending the model to provide advanced capabilities. It demonstrates how to integrate TDS metadata with CSW software and proposes several practical solutions that work around the limitations of the CSW metadata registration model. We do this to show that improvements in metadata and catalog capabilities can also reduce the challenges of big data in variability, volume, and velocity.

3. Related Work

This paper brings together several existing lines of work to confront the problems of integrating and searching vast and diverse climate science datasets. Existing research in areas of metadata modeling, geospatial information interoperability, geospatial cataloging, web information crawling,

and search indexing provides the building blocks for our work to demonstrate and evaluate advanced climate data cyberinfrastructure capabilities.

3.1. Metadata Models

There are many studies exploring the fundamental relationship between metadata models and information capabilities. There exists diverse work in other areas that deal with the same basic issues and demonstrates that the creation of novel metadata models can be used as a method for solving information challenges. For example, Spéry et al. [50] have developed a metadata model for describing the lineage of changes of geographical objects over time. They used a direct acyclic graph and a set of elementary operations to construct their model. The model supports new application of querying historical cadastral data and minimizes the size of geographical metadata information. Spatiotemporal metadata modeling can be generalized as a description of objects in space and time, and relationships between objects conceived as flows of information, energy, and material to model interdependent evolution of objects in a system [51]. Provenance ("derivation history of data product starting from its original sources" [52]) modeling is an important part of metadata study. Existing metadata models and information systems have been experimentally extended with provenance modeling capabilities to enable visualization of data history and analysis of workflows that derive data products used by scientists [53,54]. An experiment to re-conceptualize metadata as a practice "knowledge management" yielded a metadata model that can support the needs of spatial decision-making by identifying issues of entity relationships, integrity, and presentation [55]. The proposed metadata model allows communicating more complex information about spatial data. This metadata model makes it possible to build an original geographic information application, named Florida Marine Resource Identification System, that extends the use of the existing environment and civil data to empower users with higher-level knowledge for analysis and planning. Looking outside the geospatial domains, we still observe that the introduction of specialized metadata approaches and models permits the development of new capabilities.

3.2. Geospatial Metadata Standardization, Interoperability, and Cataloging

The diversity of metadata models and formats developed by research has enabled new powerful geoinformation systems, but has also introduced a new set of problems of data reuse and interoperability. Public and private research, administrative, and business organizations have accumulated growing stores of geoinformation and data, but this data has not become easier to discover and access for users outside limited organization jurisdictions. This has led to significant resource wastage and duplication of effort for data producers and consumers. Cataloging has grown increasingly challenging because of this heterogeneity. In response, new spatial data infrastructures have been developed. They have attempted to integrate and standardize multiple metadata models and develop shared semantic vocabulary models to enable discovery by employing the "digital library" models of metadata. In this process, syntactic and semantic interoperability challenges have been identified. Syntactic operability refers to information portability—the ability of systems to exchange information. Semantic interoperability refers to domain knowledge that permits information services to understand how to meaningfully use the data from other systems [56].

Various techniques for achieving metadata interoperability have been explored [57]. Two related families of techniques can be identified. One approach attempts to create standard and universal models, the other creates mappings between several metadata representations of the same data. Transformation between several metadata models requires that syntactic, structural, and semantic heterogeneities can be reconciled. The reconciliation is accomplished with techniques called metadata crosswalks. A crosswalk is "a mapping of the elements, semantics, and syntax from one metadata schema to another". Once mappings are developed, they can be used to apply multiple metadata schemas to existing data [58].

The possibilities for interoperability have been advanced by the efforts led by the International Organization for Standardization Technical Committee 211 (ISO TC 211) to standardize metadata representation. It introduced the ISO 19* series of geospatial metadata standards for describing geographic information by the means of metadata [54,59–61]. The standards define mandatory and optional metadata elements and associations among elements. For example, spatiotemporal extent, authorship, and general description of datasets are required and recommended by the standard. Other kinds of information like sequencing of datasets in a collection, aggregation, and other relational data are optional in the standard. The ISO 19* series of standards also provides an XML schema for the representation of the metadata in XML [62].

Looking at existing metadata interoperability work, we see a recurrence of similar problems such as diversity of metadata representations and complexity of mapping between them. Several authors discuss the practical challenges of developing software and systems for translation [27,59,63]. There exists a proliferation of study efforts and results that advance the goals of interoperability by identifying key understanding of the challenges of interoperability and demonstrating systems, services, and models that address common challenges. Our work attempts to preserve existing interoperability advances while exploring the possibilities of expanding existing metadata models to support new possibilities use of existing data.

Standardized metadata is often stored and made available using catalog services. Catalogs allow users to find metadata using queries that describe the desired spatial, temporal, textual, and other information characteristics of the searched data [64]. The OGC Catalog Service for the Web (CSW) is one of the widely used catalog models in the geoscience domain to describe geographic information holdings [6,65].

3.3. Web Harvesting and Crawling

One critical capacity of metadata cyberinfrastructure is the ability to integrate metadata from remote web repositories. The process of finding and importing web linked data in a metadata repository is called “crawling” and is accomplished using a software system called “metadata web crawler”. A web crawler is a computer program that browses the web in a “methodical, automatic manner or in an orderly fashion” [66]. A crawler is an internet bot, it is a program that autonomously and systematically retrieves data from the world wide web. It automatically discovers and collects different resources in an orderly fashion from the internet according to a set of built-in rules. Patil and Patil [66] summarize this general architecture of web crawlers and also provide a definition of several types of web crawlers. A focused crawler is a type that is designed to eliminate unnecessary downloading of web data by incorporating an algorithm for selecting which links to follow. An incremental crawler first checks for changes and updates to pages before downloading their full data. It necessarily involves an index table of page update dates and times. We follow these two strategies in the design of our crawler. The authors also outline common strategies for developing distributed and parallelized crawlers. Our crawler runs on a single machine, but we use a multithreaded process model with a shared queue mechanism—a common parallelization strategy identified by the authors [67].

A fairly recent review collected by Desai et al. [68] shows that web crawler research is an active area of work—however, most of this work is focused on the needs of general web search engine index construction. There exists an area of research called “vertical crawling” which contends with the problems of crawling non-traditional web data: News items, online shopping lists, images, audio, video. There does not appear any publications regarding efficient crawling of heterogeneous Earth system metadata.

There exists substantial previous work to show the feasibility of crawling this metadata. One recent paper summarizes the state of the art. Li et al. [69] present a heterogenous Earth system metadata crawling and search system named PolarHub—a web crawling tool capable of conducting large-scale search and crawling of distributed geospatial data. It uses existing textual web search engines (Google) to discover OGC standards-compliant geospatial data services. It presents an interactive interface

that allows users to find a large variety and diversity of catalogs and related data services. It has a sophistical distributed multi-threaded software system architecture. PolarHub shows that it is possible to present data from many sources in a single place. However, it does not present datasets, only endpoints that users must further explore on their own. It does not download, summarize, or harmonize the metadata stored on the remote catalogs. It shows the feasibility of cyberinfrastructure that integrates a variety of data based on interoperable standards but does not discuss data volume and velocity challenges that arise when deeper and fuller crawling is done. PolarHub users can find a large number of catalogs and services that contain, for instance, "surface water temperature" data but they cannot use metadata crawler following this catalog hub strategy to discover datasets that hold "surface water temperature inside X spatial and temporal extent with Y spatial and temporal resolution".

A complementary strategy is discussed by Pallickara et al. [70], who present a metadata crawling system named GLEAN, which provides a new web catalog for atmospheric data based on the extraction of fine-grained metadata from existing large-scale atmospheric data collections. It solves the data volume problem by introducing a new metadata scheme based on custom synthetic datasets that represent collections (or subsets or intersections) of multiple existing datasets. This reduces metadata overhead greatly and permits high performance and precise discovery and access of specific datasets inside vast atmospheric data holdings. Unlike PolarHub, GLEAN avoids the data variety challenge by limiting its processing to one type of data format used in atmospheric science. They also do not contend with the interrelated velocity and near real-time access problems—in GLEAN crawling, the discovery of updated datasets is initiated by manual user request. They do not use the OGC catalog or metadata standards to support interoperability.

BCube project (part of EarthCube initiative) attacks similar problems with another approach [71]. EarthCube is a National Science Foundation initiative to create open community-based cyberinfrastructure for all researchers and educators across the geosciences. EarthCube cyberinfrastructure must integrate heterogeneous data resources to allow forecasting the behavior of the complex Earth system. EarthCube is composed of many building blocks. Our work is part of the EarthCube Cyberway building block. BCube (The Brokering Building Block) offers a different approach for heterogeneous geodata interoperability. BCube adopts a brokering framework to enhance cross-disciplinary data discovery and access. A broker is a third party online data service that contains a suite of components, called accessors. Each accessor is designed to interface with a different type of geodata repository. A broker allows users to access multiple repositories with a single interface without requiring data providers to implement interoperability measures. BCube supports metadata brokering. It can search, access, and translate heterogeneous metadata from multiple sources. It demonstrates deeper interoperability than other approaches discussed here, but does not attempt to solve data volume or velocity problems [72]. The BCube approach is very relevant to us; however, BCube has very few documents available and the system is inaccessible. We were unable to compare some of the details of our different approaches.

Song and Di [73] studied the same problem with the same example repository: Unidata TDS. The authors determined the volume and velocity characteristics of the target repository metadata. Like our study, they propose modeling it with concepts of collection and granule. They implemented a crawler that is able to crawl some of the TDS archive. Their work is the previous progress in the same project as ours and is highly relevant to this study. However, their approach did not perform well using real-world TDS data, which led us to take it in a different direction. We rebuilt their work to demonstrate real-time search and the possibility of processing all of TDS by using a more sophisticated metadata model, and a more advanced integrated search client and indexing service that permits true real-time search.

Reviewing existing work reveals tremendous advances toward solving the challenges of creating interoperable Earth system cyberinfrastructures that can practically process a large volume and variety of observation and model data that are generated in high-velocity data production processes. Lines of work in metadata modeling, standardization, interoperability, repository crawling, and processing

provide the basis for the materials for our study. Our contribution is to synthesize these approaches to explore how interoperability and performance could be achieved simultaneously.

4. Materials and Methods

To enable searching of big climate data, we propose a new big data cataloging solution, which includes the following steps. (1) Analyze the target geodata repository that provides a good example of data challenges for cross-disciplinary Earth system scientific collaboration. (2) Analyze the qualities and characteristics of the data in the selected repository. (3) Construct a model of the repository. (4) Use the repository model to construct an efficient metadata resource model. (5) Develop a crawler system that uses repository and metadata resource models to optimize its crawling algorithm and metadata representation. (6) Demonstrate advanced interoperable big geodata search and access capabilities that our approach permits. The completed cyberinfrastructure model and system architecture (derived from our metadata model) is shown in Figure 1.

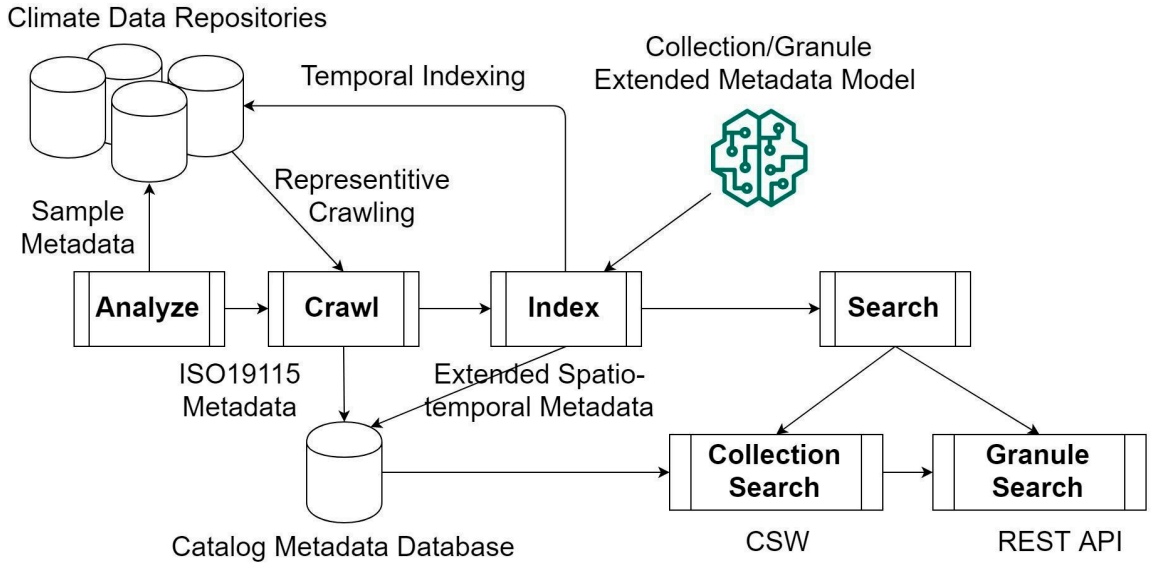


Figure 1. The proposed big climate data cataloging solution. Abbreviations: CSW, Catalog Services for the Web; REST API, Representational State Transfer Application Programming Interface.

4.1. Metadata Repository Selection

We took Unidata THREDDS Data Server (TDS) as our example target geodata repository platform. TDS was chosen because it is widely used by atmospheric and other related Earth science fields. It supports a good variety of open metadata and data standards and there exist many data centers that use TDS. It supports basic catalog features but lacks advanced search capabilities. It gives users and administrators large latitude of how the data is organized and updated inside the TDS catalog. The geodata stored across many TDS installations meets our broad criteria for real-world data variety, volume, and velocity.

A single TDS instance was selected as a target for our experiment. UCAR Unidata TDS (thredds.ucar.edu) repository was determined as a suitable target system and a good example of diverse uses of TDS. Unidata TDS contains a requisite variety of data. It has near real-time data that demonstrate the data velocity challenge. It contains a variety of data granularity and a good range in the size and complexity of datasets available. The volume of data and the volume of metadata is sufficiently challenging. The catalog structure is heterogeneous—different types of data are organized on different principles. On initial inspection, Unidata TDS was determined to be a great example of the challenges we wanted to explore.

Using manual inspection and basic statistical analysis via custom Python scripts, we started mapping out the characteristics of the Unidata TDS information system. We tried to answer the following questions: (a) What is the hierarchical structure of data organization in this repository?; (b) how frequently are new records are added and removed?; (c) which parts of the catalog exhibit regular patterns in the information structure that can be generalized and which parts contain unique information?; (d) what are the size and content of the metadata resources stored in the catalog?; (e) how is information in metadata resources related to metadata resources location within the hierarchy of catalog structure?; and (f) what are the data transmission qualities of the Unidata TDS network system—what portion of the TDS information can be transferred and copied to our system?

4.2. Repository Analysis

The following figures show some of the surface structure of the Unidata TDS catalog retrieved using a web browser from http://thredds.ucar.edu/thredds/catalog.html. Figure 2 shows the top level of the catalog hierarchy. Each listed item is a folder (a catalog). Most catalogs contain several levels of nested catalogs (Figure 3) in a tree-like hierarchy similar to a file system. At the bottom (leaf) tree level (Figure 4), the catalogs contain a list of data resources. Catalogs are presented in two formats. First is the HTML format, suitable for manual web browsing. Second is the XML format that contains additional metadata about the catalogs and the data resources. The XML representation follows THREDDS Client Catalog Specification. The specification extends the basic filesystem-like structure with temporal, spatial, and data variable description metadata annotations [74].

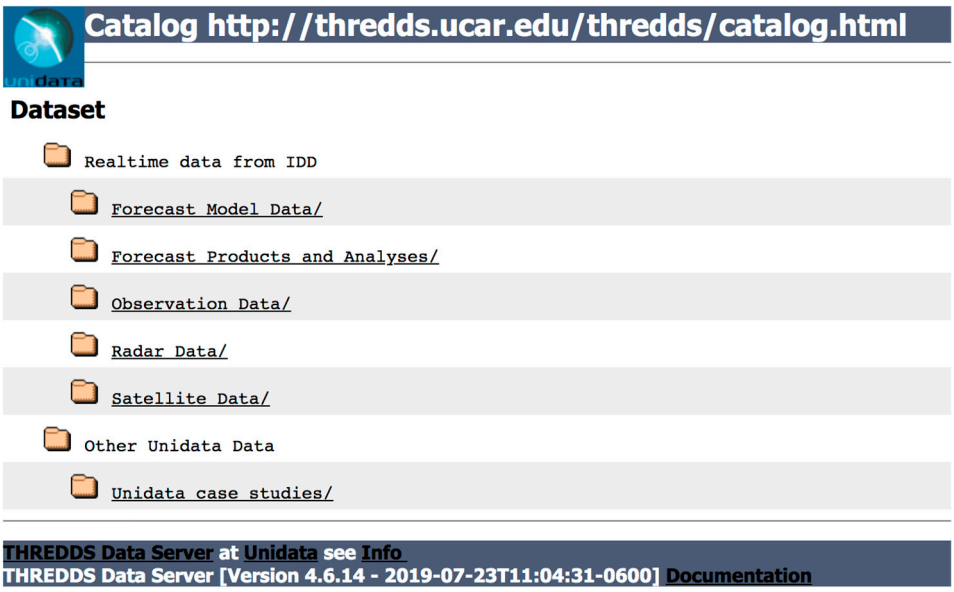


Figure 2. Top level Unidata THREDDS Data Server (TDS) catalog listing.

The TDS catalog provides a powerful general catalog hierarchy model. However, the practical use of this model by scientists who produce geodata is what determines the possibility of data collaboration and harmonization—as well as the specific shapes and possible solutions for big data problems. Email correspondence with Unidata explained that the data placed in different sub-catalogs is produced and organized by different teams of scientists. Although Unidata TDS acts as a unified repository for diverse Earth data, there are no mandatory overarching organizing principles to enable data harmonization [75].

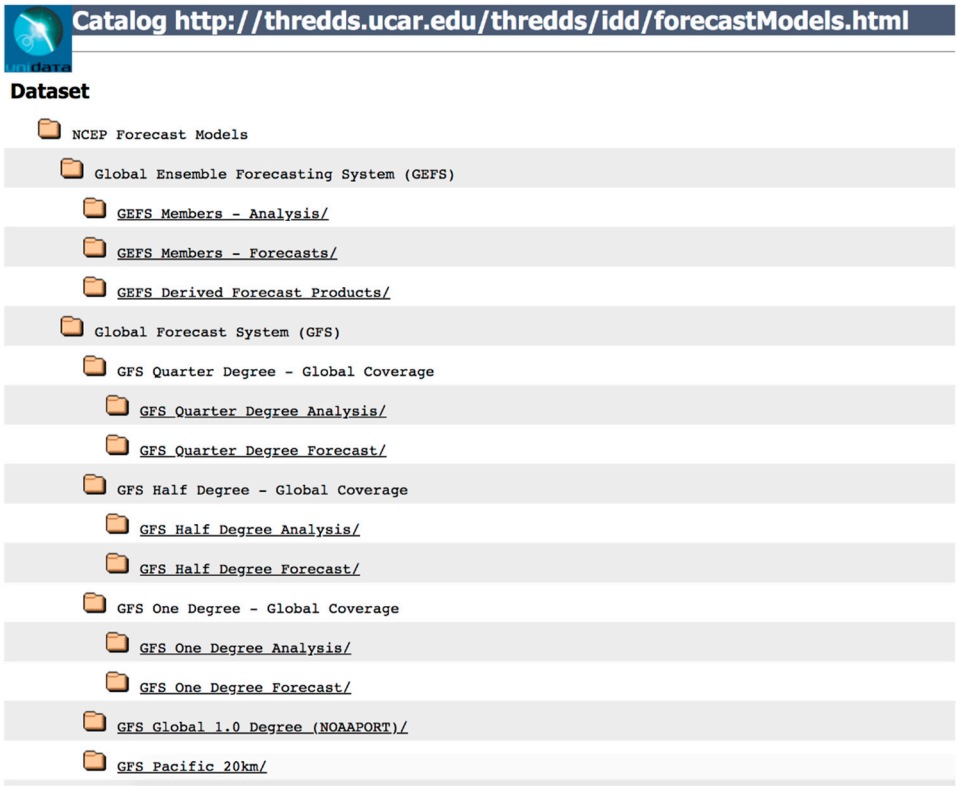


Figure 3. Nested catalogs. Only first 11 entries shown here. More entries omitted.

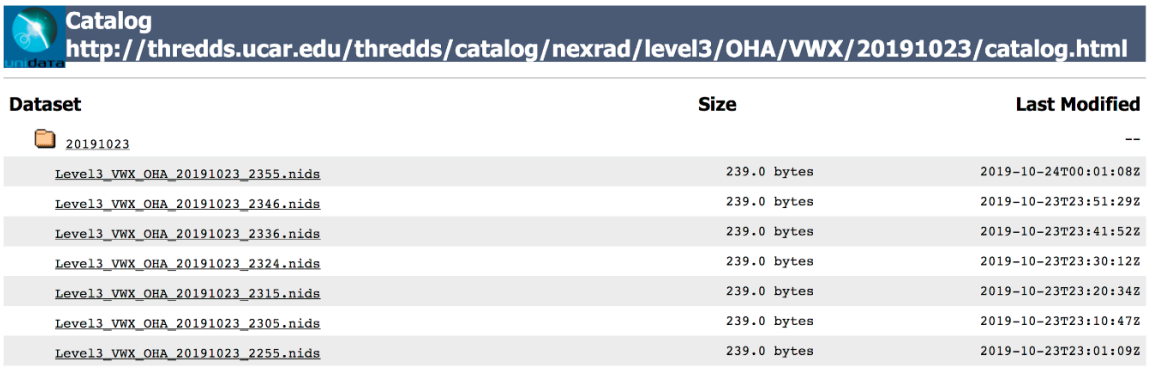


Figure 4. Data resource (dataset) listing at the bottom of the catalog hierarchy. Only first seven entries shown here. More entries omitted.

That being the case, the next step was to understand and describe different sub-structures organically adopted by different teams. After manual inspection and basic statistical analysis performed with custom Python scripts, the following information was compiled to broadly describe the different patterns of sub-catalog utilization (Table 1).

Table 1. Unidata TDS subcatalog big data characteristics.

Catalog	Sub-Catalogs/ Granules	Estimated Catalogs Size	New Granule Production	Granularity	Regularity
NCEP Forecast	5300/5300	500 MB	6 h	Coarse	Regular
Observations	8/186	20 MB	Irregular	Coarse	Irregular
Satellite	1500/71,000	150 MB	10 min	Fine	Regular
Radar	25,000/7 million	25 GB	5–10 min (irregular)	Very fine	Irregular

Four general types of data are simultaneously held in the Unidata TDS repository: (1) Forecast model output, (2) observations (time series from in-situ instruments), (3) satellite imagery, and (4) radar imagery from stationary radar network (NEXRAD, Next Generation Weather Radar) [76]. Each type contains much additional variety in its own hierarchy of sub-catalogs, but at this level, there are some clear and useful broad differences in data qualities that can guide our experiment.

In Table 1, the estimated catalog size is the total size of the metadata held in the catalog. Most of this metadata is completely redundant, but without knowing the deeper structure of this data, we would have to mirror all of this data in order to enable search and discovery capabilities that THREDDS does not support. We calculated maximum data transfer throughput of 4 MB/s or 5 min to load 1 GB of catalog data. It appears to be possible to mirror the entire Unidata TDS metadata catalog in several hours, but data throughputs we observed were not consistent, often slowing down by one order of magnitude. Furthermore, the speed of data processing (indexing and registering with a standard-compliant OGC CSW catalog) is also very time, and compute and storage resource, consuming. We do not have the capabilities to register and search millions of records mostly containing redundant information. Furthermore, the Unidata TDS data were added in near real-time according to specific patterns and structure in the sub-catalogs. If we attempted to copy and register all of that metadata, then we would not have been able to provide near real-time capabilities.

The last two columns in Table 1 show two critical qualities that determine what approach we needed to take to integrate that metadata into our systems.

If final datasets have "coarse" granularity, that means each dataset is a very file and the size of metadata is small in relation to the data size—for "coarse" datasets, we can copy, harvest, and index the metadata into our search system. "Fine" datasets stretch technical capabilities to transfer and process metadata. "Very fine" records are too numerous (the data files too small) for us to be able to effectively synchronize or process their metadata.

If datasets are produced in a regular way (predictable spatiotemporal attributes), then we can harvest minimal information and model the entire catalog. However, for NEXRAD radar metadata, there is no regular pattern to metadata production. A new record could be added every 5 min or every 15 min—and their regularity/irregularity also varies in time and depending on different radar sites (different sub-catalogs). This fine-grained irregular data is the most challenging, because it can neither be harvested wholesale nor modeled in an accurate way. It requires a targeted combination approach. Additional considerations arise when tracking what datasets have expired and been removed—ideally, this should be accomplished without performing an expensive full scan of the TDS repository.

Further examination of the sub-catalogs structure for irregular (and regular) highly granular data revealed additional useful structural information. Some catalogs are "dynamic" (or "live" or "streaming")—they are updated with new data resources with regular (or irregular) frequency. Other catalogs are archival—they can be assumed to never change (until they expire and are deleted entirely). Three distinct types of sub-catalogs can be identified:

- Pure archival directories: These folders only contain old collections and granules and will never be updated or deleted.
- Mixed archival directories: Some of the sub-folders contain archive material, some contain live, streaming near real-time data granules and collections.
- Daily archival directories: Folders that contain streaming data for a given day; when the day passes, this directory becomes an archive folder and does not need to be mirrored again. When daily archives expire, all the data resources for that day are deleted together.

4.3. Crawling

Big data catalogs normally need to complete a lot of crawling tasks to grab metadata files, and repeat scanning to capture metadata of the newly observed datasets on a regular basis. Crawling is the fundamental information source of metadata, and how to intelligently crawl is one of the largest challenges in big data searching due to the repeated computational burden and the complexity of

the content. When designing our crawling strategy, we considered the observation update frequency, time window, observatory network organization, and made the crawler only touch the folders of those updated sensors at collection (sensor) level. Although a sensor has millions of metadata records, we only crawl the metadata at the sensor level. In other words, only one metadata is crawled for each sensor (or instrument). Using this strategy, we can save numerous hours in crawling and metadata transferring over network, especially when the network is unstable. After applying a parallel worker mechanism, we can have dozens of crawlers working on scanning and capturing new/updated metadata of petabytes of climate datasets.

Our crawler is different from most existing crawlers in the literature, because it is not a general-purpose search engine crawler. Typical crawlers download the entire web page, find links to follow, and add those links to the work queue. We cannot do a similar thing, because the web content we are crawling (TDS catalog) contains vastly redundant information that is not possible to download and process in its entirety without overloading available computing and network resources. There are various sensors in the climate monitoring networks and the sensors are dynamically changing, with new sensors added or old sensors removed. We had to crawl the THREDDS Data Server to make sure all the observations were fully synchronized in our catalog. Our crawler design must incorporate knowledge of metadata and metadata structure its processing and queueing algorithms in order to download only essential information.

4.4. Indexing

The third step is indexing, which extracts the spatiotemporal information from the crawled metadata and creates indexes for data granules of times series by each instrument. CSW provides the basic metadata registration and query model. However, the large granularity of metadata objects (and lack of aggregation/relational capabilities) makes CSW inefficient for storing and querying large numbers of datasets and that have only small variations in their metadata. A more efficient model is needed. This is a long explored and essentially solved problem in computer science and informatics. Theodoridis et al. [77] summarize the basic approach. For a time-evolving spatiotemporal object, a snapshot of its evolution can be represented by a triplet $\{o\_id, s_i, t_i\}$—object id, space-stamp, and time-stamp. This information allowed us to create a "repository production model" (Figure 5). We identified patterns in the catalog hierarchical structure that allowed us to identify which paths in the catalog folder hierarchy are "live" and which ones are "archival". In our crawler implementation (discussed in the next section), we used the structure path patterns to drive the crawler algorithm in two stages—"full sync" stage, which copies the archival data, and "update" stage, which monitors and refreshes the listing from "live" catalog paths.

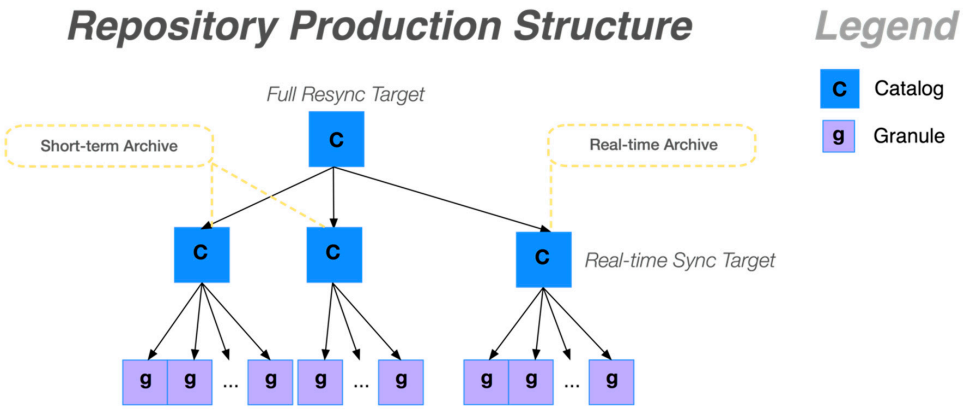


Figure 5. Repository metadata production model.

The repository production model allows targeted crawling—however, the number of metadata resources remains too large to harvest, process, and index in its entirety, even when done in two stages

to avoid redundant harvesting. We needed a second model that encompasses the metadata information structure (Figure 6). There are two issues we needed to solve: First is that most of the metadata in the catalog is completely redundant; second is that metadata information scope is not consistent in the catalog. The two issues have the same source: Catalogs, and sub-catalogs and data granules all can have metadata attached.

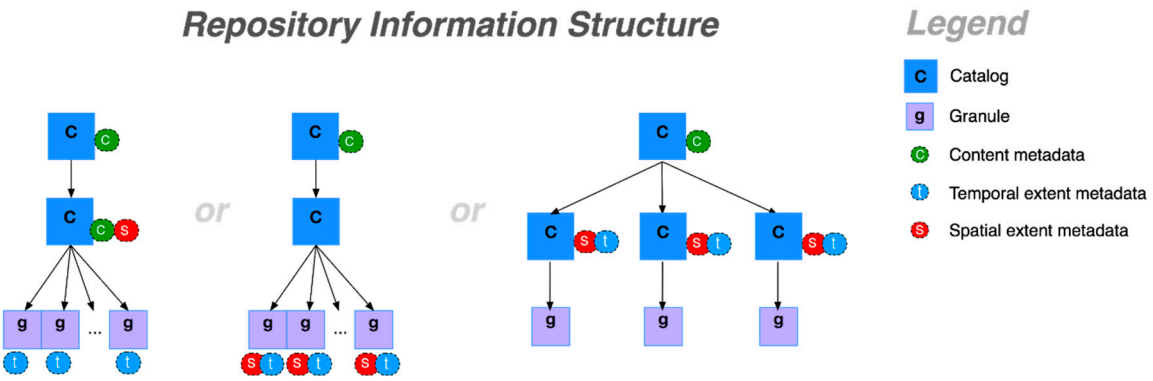


Figure 6. Repository information model.

In these examples from Unidata TDS, we see that metadata is attached to the hierarchical catalog structure in various ways. In the first example, a catalog contains some content metadata (for example: Authorship), the sub-catalog contains additional content metadata (ex: Variable names) and spatial metadata, while each granule contains temporal metadata. In the next two examples, the distribution of metadata between catalogs and granules is different. The last example is a case where each catalog only contains a single data record (granule). In some cases, the metadata is simply duplicated between several catalog levels, while in others, one specific layer contains all metadata. Another important detail is that the catalog hierarchy, the names of parent catalogs is also metadata for the data resources.

When combined, these two perspectives (information change model and information structure model) produce a model of the Unidata TDS repository that can be used to develop efficient (non-redundant) harvesting and representation of all contained metadata. By applying the production model to our crawler design, we were able to harvest only the information we know had changed. Knowing the structure of data changes also allowed us to perform targeted incremental harvesting for near real-time discovery capability. We defined two types of objects: Collections and granules. Collection contains content metadata (title, description, authorship, variable/band information, etc.). Each collection contains one or more granules. Each granule contains only the spatiotemporal extent metadata. The OGC CSW catalog standard does not support the composition of collections and granules, so we used CSW to represent collections only, while granules had to be stored externally. We used popular PyCSW software to hold collection metadata. We extended PyCSW with a PostgreSQL relational database to store relations between collections and granules and granule metadata (Figure 7).

4.5. Two-Step Search Process

When the metadata is harvested into PyCSW and temporal granule index is saved in PostgreSQL, the search clients can use these two data sources to retrieve final results for access. The search process takes place in two steps. Initially, the client searches the PyCSW store using standard search methods and queries. This returns a list of collection level results. To get a list of granules, the search client sends a second query to the crawler service. The crawler service queries the granule index, refreshes the index with the latest granules if needed, and returns a list of granules for the requested collection. The search client can then use the collection level CSW record and combine it with selected granule information to produce granule level CSW information. Figures 1 and 8–10 show these interactions from systems architecture and event sequence perspectives.

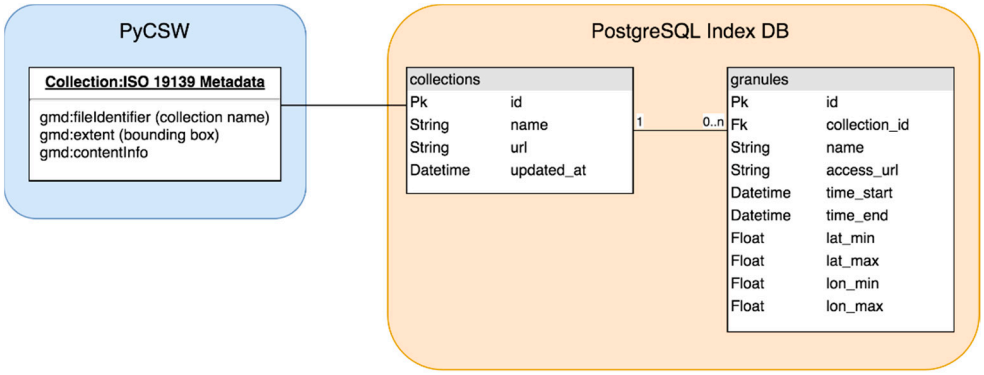


Figure 7. Metadata collection and granule resources stored in referentially linked PyCSW and PostgreSQL databases. PyCSW gmd: fileIdentifier corresponds as a key to collections table name field in the SQL database.

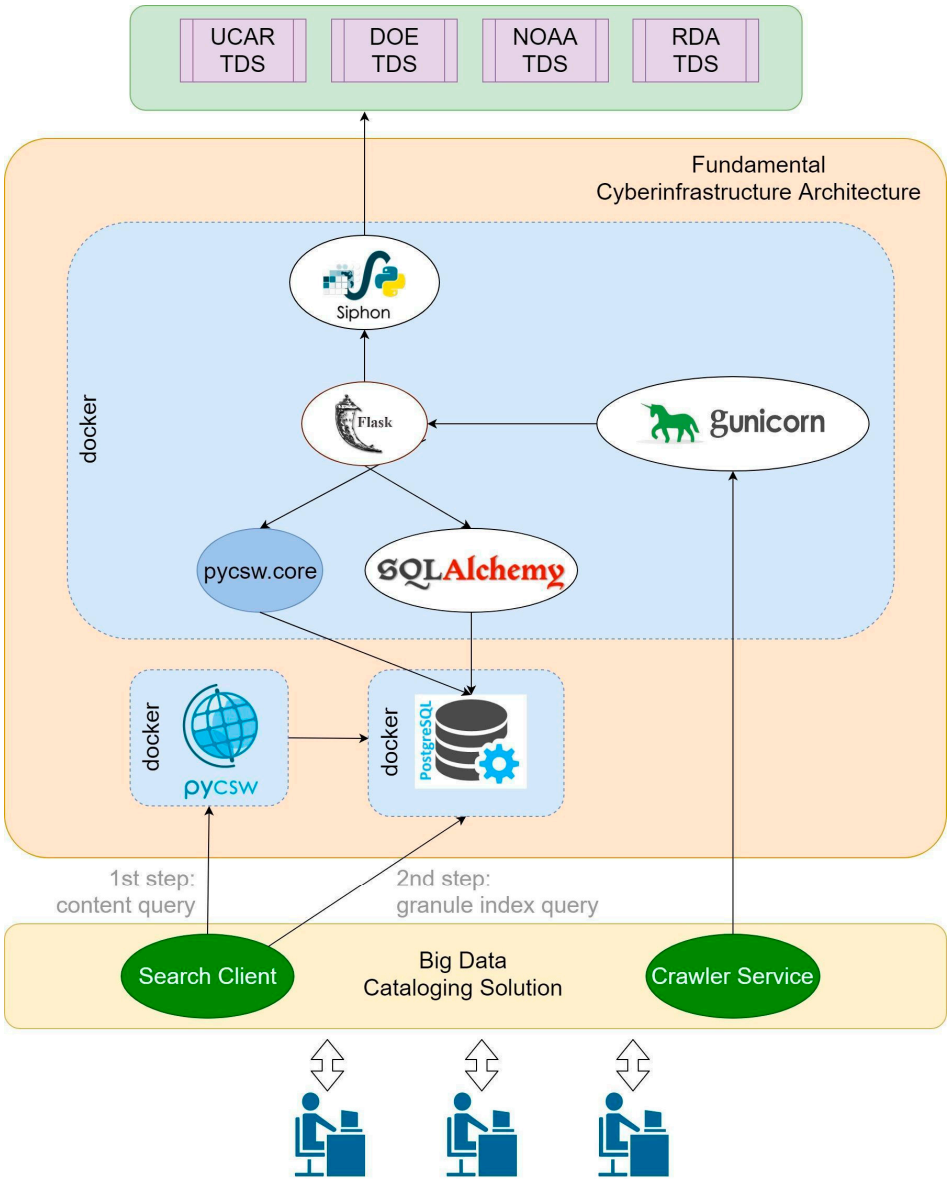


Figure 8. Implementation architecture for searching big data served via Unidata THREDDS Data Server (TDS). Although in our study only UCAR TDS is used, the system is designed to support any TDS repository as a data source.

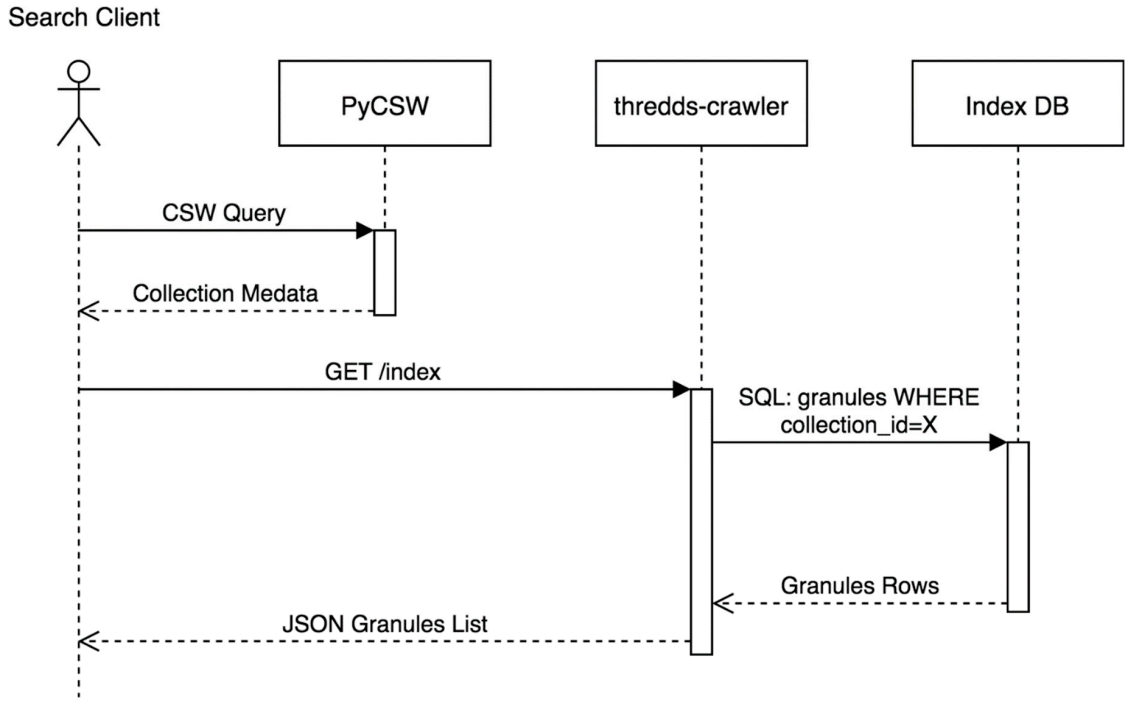


Figure 9. Simple granule index retrieval during search.

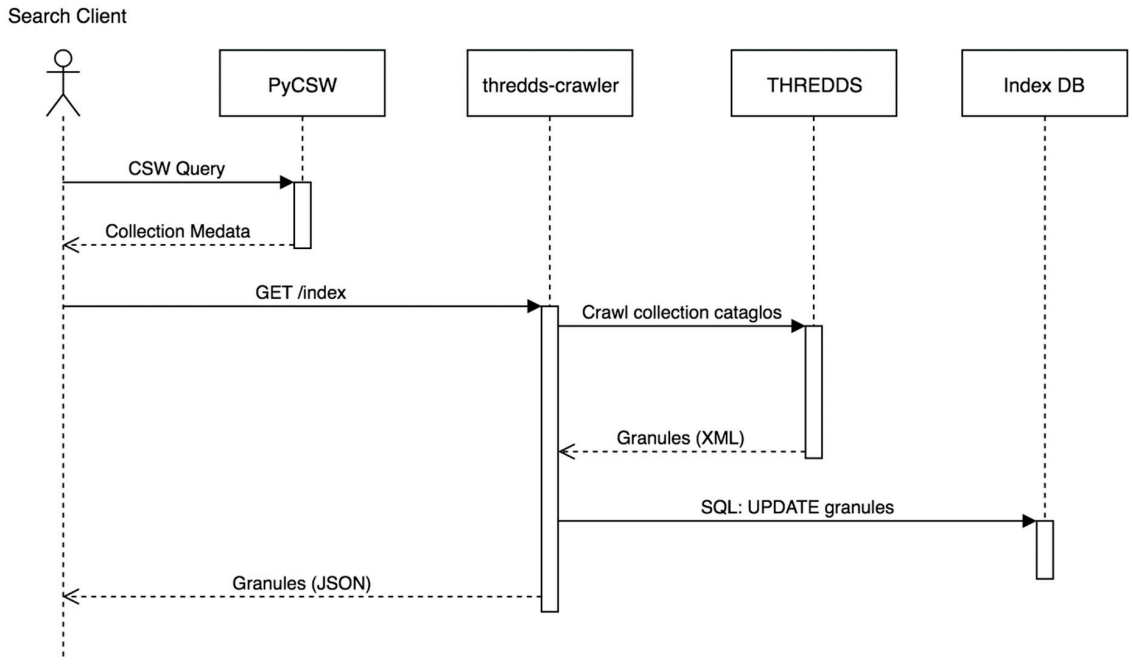


Figure 10. Granule index retrieval when granule temporal range is outside the range stored in the index. Search triggers an additional step that immediately crawls the TDS catalog and updates the real-time granule metadata. Abbreviations: DB, Database.

So far, we have analyzed the Unidata TDS repository structure, built a model of the repository that can inform an effective crawling strategy, and defined the model for the product output for the crawler. We have also described how a search client should function. To complete our experiment, we built a crawler that follows our metadata model and demonstrates a web search capability for the entire contents of Unidata TDS.

5. Implementation

We implemented a module system within EarthCube CyberConnector [17,78,79] to realize the proposed mode (Figure 8). The implementation included the searching server system and the client system. We will introduce the searching capabilities enabled by these systems.

5.1. Crawler Service Implementation

We built a web crawler that traverses Unidata TDS and extracts and stores essential metadata without using unnecessary resources. It is named 'thredds-crawler' and the source code is available via a public GitHub repository: https://github.com/CSISS/thredds-crawler.

The crawler is written in Python. It was built using common open source libraries for HTTP API interaction (Flask [https://www.fullstackpython.com/flask.html], Gunicorn [https://gunicorn.org/]), general XML processing (libxml [https://lxml.de/}]), and database abstraction (SQLAlchemy [https://www.sqlalchemy.org/]). It uses native Python threading libraries to support concurrency. For traversing the Unidata THREDDS catalog and retrieving metadata, it uses the Unidata provided Python Siphon library [https://github.com/Unidata/siphon].

To support our big-data experiment requirements, the crawler is tightly integrated with a catalog software PyCSW [https://pycsw.org/] and a PostgreSQL [https://www.postgresql.org/] database. The crawler, PyCSW, and the database each run in a separate Docker [https://www.docker.com/] container. For the sake of this demonstration, all three services run on the same machine and communicate over the local network. The Docker-compose tool is used to connect and orchestrate the three containers. This architecture allows simple scaling out to multiple machines using containers, which allows for potential substantial improvement in system performance.

The crawler docker container runs as a web service hosted by Gunicorn—a python HTTP server widely used for hosting web applications. It serves three HTTP API endpoints that perform the following functions: Harvest, create index, and read index.

The harvest function loads the Unidata Catalog XML from specified catalog_url using Siphon library. The catalog contains a list of datasets. TDS has a feature to translate its dataset metadata into the ISO/OGC-compatible XML format. For each dataset being harvested, the harvester constructs a query to TDS to retrieve ISO/OGC metadata for the dataset. The ISO/OGC metadata returned by TDS is, however, often incomplete, inaccurate, or inconsistent in some way. The crawler harvesting process then applies a chain of XML filters to the ISO/OGC metadata to rectify it with information from native TDS dataset metadata. Once the metadata is downloaded and processed, it is saved in the PyCSW database directly by using a PyCSW compatibility library.

Indexing is similar to harvesting, but involves a strategy for targeting datasets to be harvested and additional processing steps. During index creation, TDS catalogs and datasets are turned into collections and granules in our model. For each TDS dataset encountered, we determine the collection name. Crucially, the collection name is not the name of the catalog containing the dataset. We found that catalog names are inconsistent, but that TDS dataset ids contain consistent identification information. In the TDS, dataset ids are kept unique by including timestamps in the dataset id. For example, in a dataset with id "NWS/NEXRAD3/PTA/YUX/20190830/Level3_YUX_PTA_***20190830_1713***.nids", the portion "***20190830_1713***" is a timestamp. To turn TDS catalogs with datasets into collections with granules, we remove the temporal information to construct the collection id. Then, we download the dataset in the ISO/OGC XML format and transform its XML content with a filter function called "collection builder". This function updates the dataset metadata to turn it into a more general form that describes the collection. It changes identifiers stored in the metadata. It also adds standard-compliant additional fields that identify the metadata for describing "series" ("series" in ISO/OGC model, "collection" in our model). This process needs to be done only for the first dataset encountered for each collection. When processing additional datasets, the existing collection is reused. In TDS, the dataset spatiotemporal extent information is part of the catalog metadata, which means that we only need to download a single dataset metadata to build the collection metadata and we can index the remainder of granules

from catalog metadata. This solves the redundancy issue that previously prevented TDS from being searchable. We also correct TDS identifiers to ensure that the namespace authority portion of the identifier is correctly set.

The following tables help illustrate the process of extract collections from a granule identifier for multiple types of data. Table 2 shows the catalog paths of TDS datasets. Table 3 shows how the collection identifier is generated, and Table 4 shows the final result collection name.

Table 2. Catalog paths of TDS datasets for three types of data. The catalog path hierarchy is marked in green. The dataset filename is marked in red.

Data Type	Example TDS Catalog Path
RADAR	Radar Data › NEXRAD Level III Radar › PTA › YUX › 20190830 › **Level3_YUX_PTA_20190830_1713.nids**
Model	Forecast Model Data › GEFS Members › Analysis › GEFS_Global_1p0deg_Ensemble_ana_20190731_0000.grib2 › **GEFS_Global_1p0deg_Ensemble_ana_20190731_0000.grib2**
Satellite	Satellite Data › GOES West Products › CloudAndMoistureImagery › Mesoscale-2 › Channel16 › 20190831 › **OR_ABI-L2-CMIPM2-M6C16_G17_s20192430003570 _e20192430003570_c20192430003570.nc**

Table 3. TDS dataset identifiers for three types of data. The portion of the identifiers that contain temporal information is highlighted.

Data Type	Example TDS Dataset Identifier
RADAR	NWS/NEXRAD3/PTA/YUX/ **20190830** /Level3_YUX_PTA_ **20190830_1713** .nids
Model	grib/NCEP/GEFS/Global_1p0deg_Ensemble/members-analysis/ GEFS_Global_1p0deg_Ensemble_ana_ **20190731_0000** .grib2
Satellite	goes-west-products/CloudAndMoistureImagery/Mesoscale-2/Channel16/ 20190831/OR_ABI-L2-CMIPM2-M6C16_G17 _ **s20192430003570_e20192430003570_c20192430003570** .nc

Table 4. Collection identifiers for the example dataset IDs. They are calculated by removing temporal information and prefixing an authority namespace field.

Data Type	Calculated Example Dataset Collection Identifier
RADAR	edu.ucar.unidata:NWS/NEXRAD3/PTA/YUX/Level3_YUX_PTA.nids
Model	edu.ucar.unidata:grib/NCEP/GEFS/Global_1p0deg_Ensemble/members-analysis/ GEFS_Global_1p0deg_Ensemble_ana.grib2
Satellite	edu.ucar.unidata:goes-west-products/CloudAndMoistureImagery/Mesoscale-2/ Channel16/OR_ABI-L2-CMIPM2-M6C16_G17.nc

When the index harvesting is complete, the collection information (OGC/ISO 19139 XML metadata format) is stored in PyCSW. The granule information is stored in a compact SQL index store (Figure 7). Once the index is created, it can be retrieved from the crawler web service using HTTP API (GET/index). These requests take a collection name and temporal extent as parameters. Although our data model includes granule spatial and temporal extent, at the time of publication, only temporal index queries were implemented. It checks the index data store to see if the latest available granules are newer than the requested time extent. If more recent granules are not required, the crawler returns a list of granules in compact JSON format (Figure 9). However, if the index does not contain recent enough granules, then the index service performs a partial "refresh" indexing of the TDS repository. It uses the TDS catalog link stored in our PyCSW collection and re-runs the index process described here. (Figure 10). However, as we discussed in the Experiment section, the TDS catalog is organized with some sub-catalogs storing archival information, while others contain near real-time "live data". The crawler index refresh process takes advantage of that structure. It ignores the old sub-catalogs and only

indexes those that contain more recent and unknown data. This makes near real-time index retrieval fast and efficient.

Both harvesting and index creation use the same multi-threaded queue strategy to achieve higher performance. Normally, most of the time is spent waiting for data to be transmitted over the network. By using many threads, we can increase the saturation of both the network and local computer and memory resources, which allows the metadata to become available much faster.

5.2. Search System Implementation

The search system is implemented based on the previously developed EarthCuber CyberConnector infrastructure building block [17]. CyberConnector is a Java-based web application that supports discovery and visualization of data from CSW catalogs [17]. We extended CyberConnector to support accessing metadata harvested and indexed by the thredds-crawler described in the previous section. We modified the CyberConnector Search Client to perform a two-stage search. The web application user selects "Search" function (Figure 11). They select a time range, which is used by the thredds-crawler index service to determine if granule refresh is needed. The web browser sends an AJAX request to the CyberConnector web application with search parameters. CyberConnector queries thredds-crawler PyCSW service for collections that match query parameters. It returns a list of collections. To see the granules available in a collection, the user clicks the "List Granules" button (Figure 12). This issues another request to CyberConnector for a granules list in the specified temporal extent. CyberConnector web application proxies the granules list request to thredds-crawler indexing service, which returns a list of granules (Figure 9); or thredds-crawler harvests TDS to update the index and then returns a list of granules (Figure 10). The client receives a list of granules, which can then be downloaded or visualized (Figure 13).

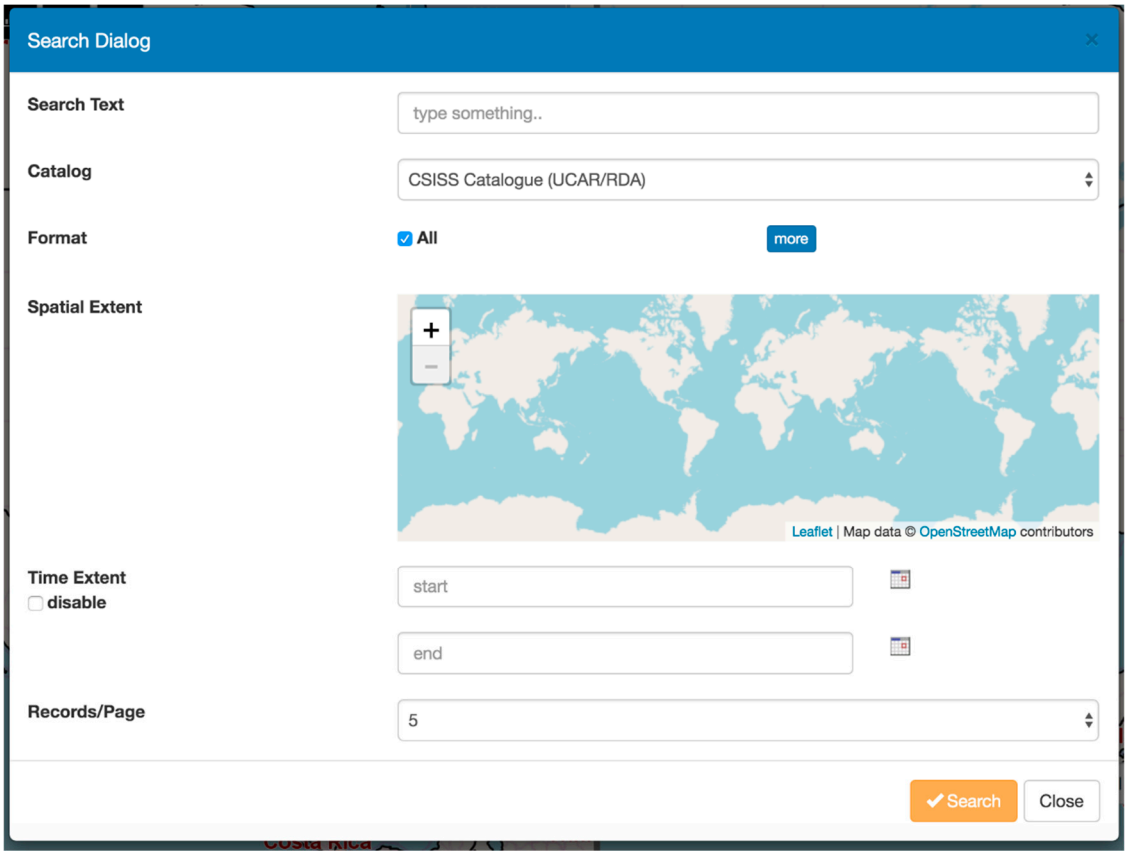


Figure 11. Search client web interface.

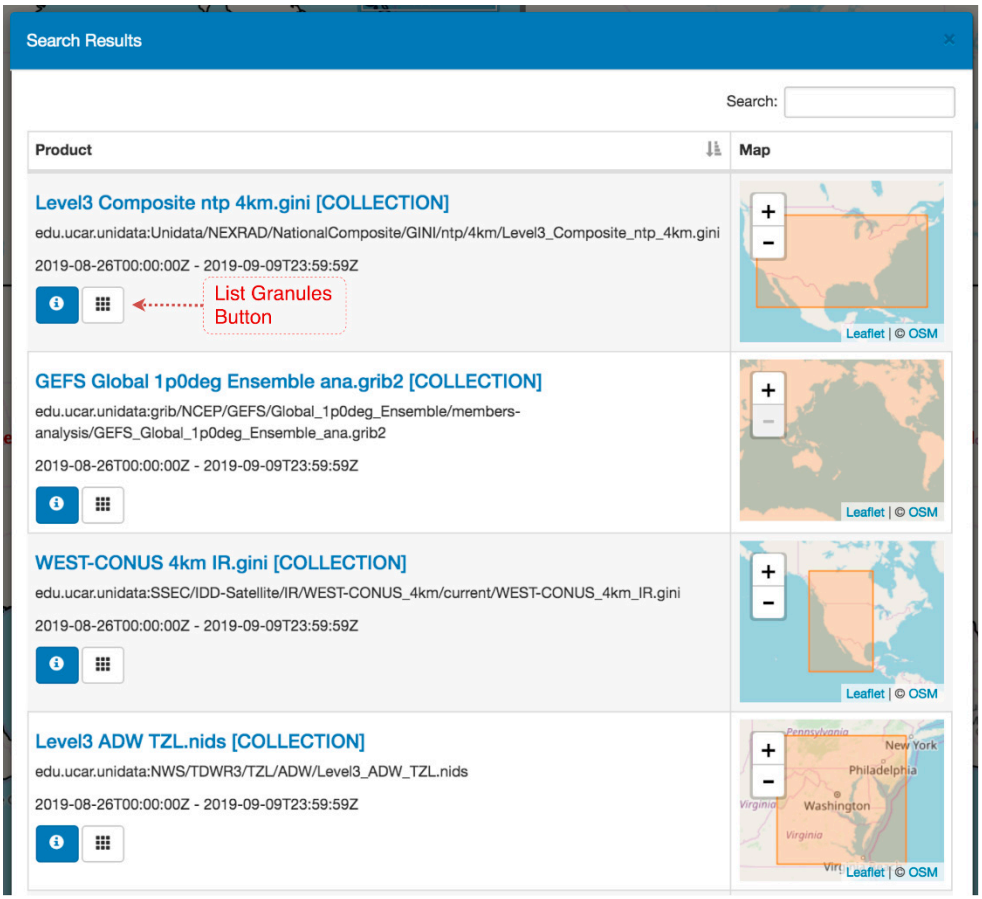


Figure 12. Search results with "List Granules" button.

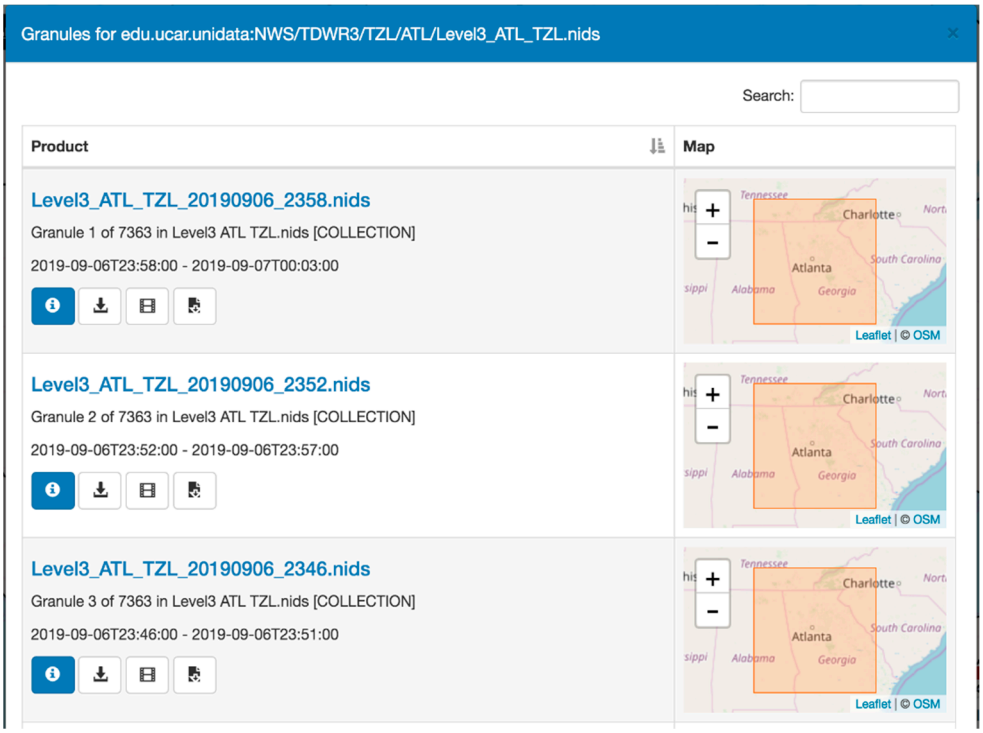


Figure 13. Granules list for a collection. The buttons allow users to view metadata, download the dataset, or visualize it.

6. Experiment and Results

Based on the implemented catalog system, we conducted several experiments to validate the feasibility of the proposed approach. The datasets for climate science are generally very large because of their long-term running and high temporal resolution. We took the UCAR NEXRAD dataset [80] and the RDA ASR dataset (53.09 terabytes) [81] as our demonstration examples. The searching capabilities on the two datasets were established in the EarthCube CyberConnector. We made a complete set of tests on the searcher and the results are introduced below.

6.1. Searching the NEXRAD Dataset

NEXRAD is a very important dataset for climate science research. It currently comprises 160 sites throughout the United States and selected overseas locations (as shown in Figure 14). The basic original datasets, including three meteorological base data quantities: Reflectivity, mean radial velocity, and spectrum width, are called Level II. The derived products are called Level III, which include numerous meteorological analysis products. All NEXRAD Level-II data are available via NCEI, as well as NOAA big data plan cloud providers, Amazon web service (http://thredds-aws.unidata.ucar.edu/thredds/catalog.html) and Google Cloud (https://cloud.google.com/storage/docs/public-datasets/nexrad). UCAR provides the near real-time observed data via their THREDDS data server (http://thredds.ucar.edu). Unfortunately, all these data repositories are still non-searchable at present, because it is a huge challenge for any catalog to index and search such big amount of metadata files for the frequently updated radar data records (every 6 min). We used this dataset to prove that the proposed cataloging approach can work well on frequently updated big datasets.

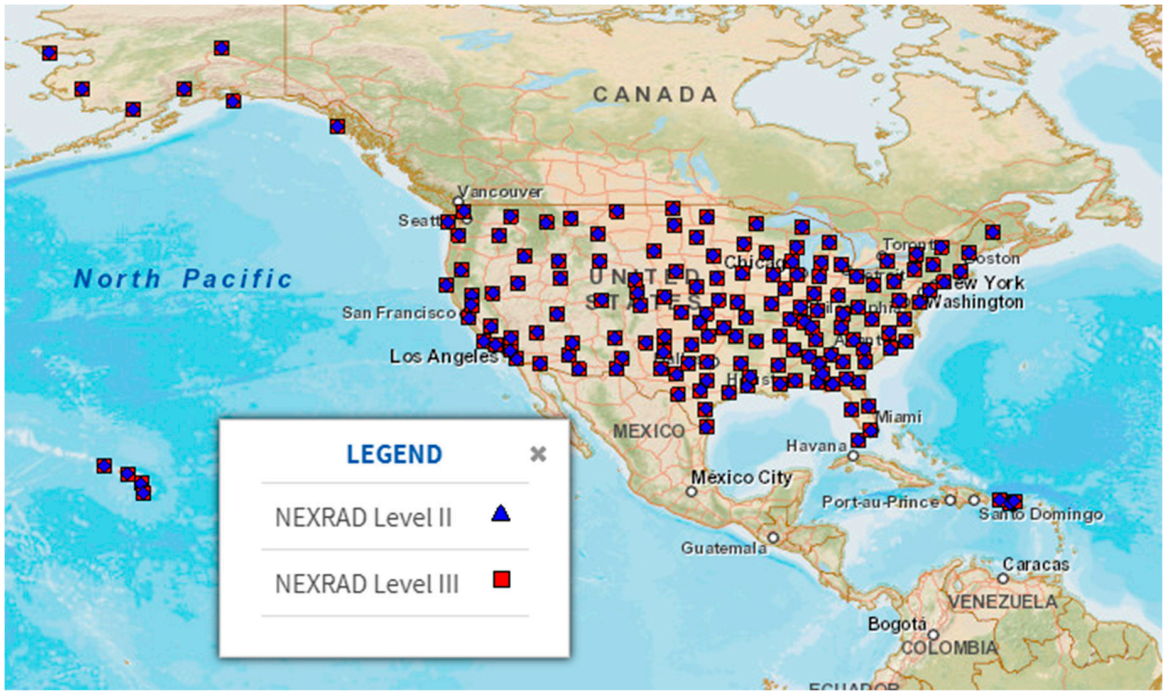


Figure 14. NOAA NCDC Radar Data Map (NEXRAD Level II and III).

The completed system consists of the harvester/indexer service and the search client that is available to the user as a web application. As a result, users are able to search diverse heterogeneous Earth system observation and modeling datasets simultaneously. Once the metadata is found, users can use the CyberConnector visualization system to simultaneously visualize near real-time NEXRAD radar, satellite observation, and forecast simulation model product data. The system performance

characteristics of this approach are significantly improved over the existing naive method of harvesting all of the datasets' metadata.

6.2. Searching UCAR RDA (Research Data Archive) TDS Repository

NSF-funded NCAR CISL (Computational & Information System Lab) maintains Research Data Archive (RDA), which stores over 11,000 terabytes of climate datasets in its high-performance data storage system.

RDA hosts many climate datasets at present, and the Arctic System Reanalysis (ASR) is one of them. ASR is a demonstration regional reanalysis for the greater Arctic developed by Ohio State University. The ASR version 2 dataset (the latest version) is served via RDA with a total volume of 53.04 terabytes. The horizontal resolution is 15 km and the temporal coverage is from 2000 to 2016. It has 34 pressure levels (71 model levels), 31 surface (including 3 soil variables), and 11 upper air analysis variables, 71 surface (including 3 soil variables), and 17 upper air forecast variables.

RDA provides TDS for most of its archived datasets. We harvested the metadata of ASR from its TDS and made them publicly available in CyberConnector. As shown in Figure 15, scientists can search the ASR dataset by providing keywords, spatial extent, or temporal range. The ASR data is in NetCDF format, which is displayable in COVALI. We demonstrated searching ASR dataset in COVALI and visualized the temperature at 2 m above the surface within 12 h. COVALI and RDA were deployed in two remotely distributed facilities. The interactions between COVALI and RDA big data storage were conducted via the standard service interface and over the network. The experiment proves that the proposed solution works well for enabling search on remote big data.

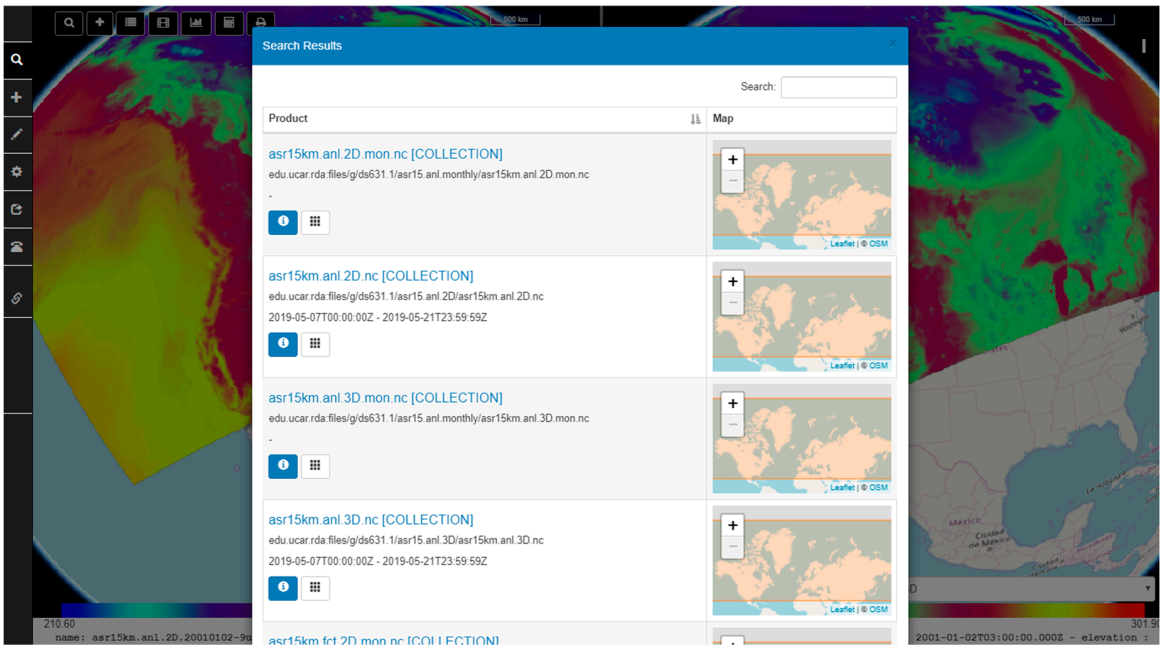


Figure 15. ASR (Arctic System Reanalysis) search results and visualization of the temperature at 2 m above the surface in the CyberConnector COVALI visualization system.

6.3. Performance Evaluation

The traditional approach for cataloging climate datasets is fully harvesting all the metadata files of every single data record. We implemented the searcher using the traditional method before, but the performance was very slow and sustained operation not possible for the practical scenario for big data cataloging. After we applied the new cataloging strategy, we tested it by crawling several hundreds and thousands of records from UCAR THREDDS Data Server. We tested using different sets of parallel workers: 40 workers, 20 workers, 10 workers, 5 workers, and a single worker, respectively, to measure

the improvements of parallel crawling. Figure 16 displays the time cost of the test to compare the performance of the traditional approach and the proposed approach. The results demonstrate that the proposed approach outperforms the traditional approach at least ten times on the overall time cost (from ~10 to ~1 s) and has significant improvements on harvesting speed, storage use, and search speed based on the number of datasets being processed.

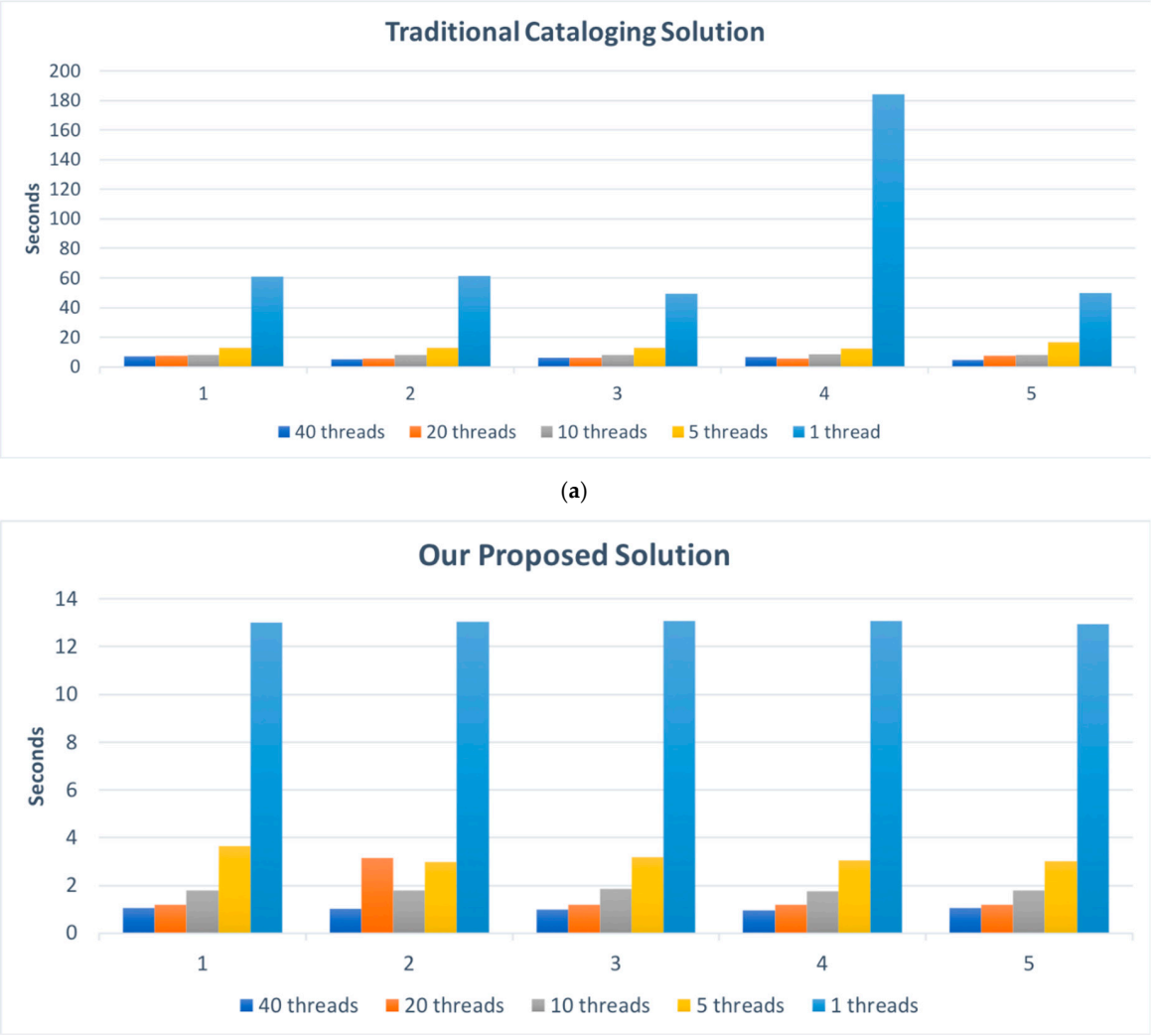
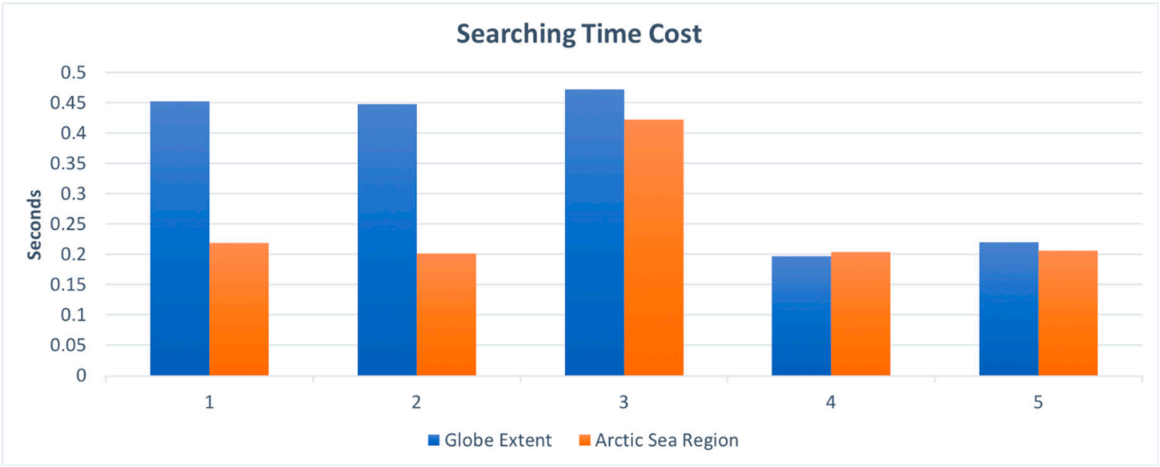
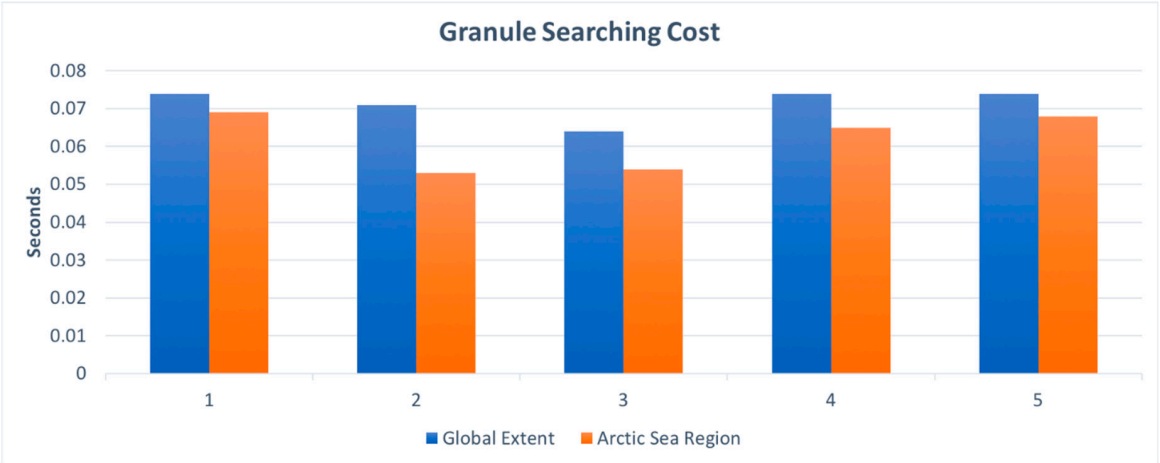


Figure 16. Performance comparison (time in seconds) of the traditional harvesting approach (**a**) and our approach (**b**), sampled 5 times for crawling 125 records.

Search time cost has two components. The time to search for collections in the catalog and the time to retrieve the granules list from the granule index. Figure 17 shows that search result retrieval is extremely fast in our system.



(**a**)



(**b**)

Figure 17. Search performance (time in seconds) with two different spatial extent parameters. (**a**) Time to query the catalog (collection searching, step 1); (**b**) time to query the granule index (granule searching, step 2).

The search currently supports filters, including keywords, data format, and spatiotemporal extents. All of them are fixed filters with less uncertainties. Therefore, the returned results stay the same as long as the metadata base does not add new records or delete existing records. The result completeness is 100% accurate because correct records match the filter conditions. Users can narrow down spatiotemporal extent based on their interest and provide one or more keywords which could match the data field names. The first page relevance of the search results depend on the relationships between the inputted keywords and the metadata field values. Based on our experiences with climate scientists, we find that they normally do not input any keywords and only use spatiotemporal filters to check out what can be searched in a catalog. Once they have a region of interest or a time window, they then have an impression about what possible data is available. They come to the catalog just to find the access URL to download or visualize the data files. Normally, results from our search client are numerous because of the loose filters the scientists give. The results on the first page are usually very well related to the scientists' needs. A more intelligent search, such as a semantics-based search, which could find more accurate first-page results with higher relevance, will be studied in the next stage of work.

7. Discussion

Our solution to the big data volume, variety, and velocity challenges discussed in the paper consists of a novel metadata model, and cyberinfrastructure architecture and implementation that is derived from the model. The metadata model combines the description of metadata content (the "information model") with the description of metadata repository structure and behavior. The cyberinfrastructure consists of a crawler service that takes advantage of the metadata model to optimize THREDDS crawling strategy to eliminate the transfer and processing of redundant metadata information. Additionally, the metadata repository model permits the crawler service to perform incremental metadata transfer, which enables real-time search capability. The demonstrated cyberinfrastructure also includes an interoperable catalog service that uses the metadata model to minimize the storage of redundant information. Finally, a search client that uses the catalog and the crawler services is implemented.

7.1. Can the Proposed Solution Address the Volume Challenge?

Metadata volume is ~25 GB for the UCAR RADAR dataset. The traditional method for harvesting metadata (as discussed in Section 6.3) is able to process approximately one record (with an approximate size of 100 KB) per second. To completely ingest all of THREDDS RADAR metadata at the observed harvesting rate, it would take 250,000 s or ~70 h. By using the proposed metadata model and cataloging system, we observe harvesting rates that are at least 10 times faster. This permits daily synchronization of all Unidata TDS metadata.

7.2. Can the Proposed Solution Address the Velocity Challenge?

We determined that new (live) RADAR metadata is being generated at 330 records per minute. Our maximum harvest capacity (constrained by Unidata THREDDS network capacity) is 60 records per minute. Using the traditional method, we cannot keep up with the data velocity. Using the indexing harvester approach, we can process up to 1400 records per minute. This exceeds the velocity of THREDDS data production. Additionally, by using incremental index update during the client search request exchange, we can target the indexing harvest process to the exact sub-catalog containing the updated information and thus provide real-time search capability for this high-velocity data.

7.3. Can the Proposed Solution Reduce Metadata Crawling Redundancy?

The solution demonstrated here is able to reduce redundancy in crawling and storage resource consumption. For example, using the traditional method with Forecast Models catalog, ~7000 records are downloaded. The total storage used is 1.85 GB. The same metadata can be processed using our approach by downloading only 45 sample metadata records (2.2 MB) that represent collection level information. This represents a 99% reduction in data transmission and storage costs.

7.4. What Are the Benefits and Drawbacks of the Proposed Solution Compared to Other Big Data Searching Strategies?

The solution demonstrates the expected benefits described at the beginning of this study. The main drawback of this solution is the model and software system complexity. Custom software has to be developed to intelligently process catalogs as they are being harvested. To get complete and accurate results, the ingested metadata must be cleaned and transformed to fill in missing pieces of information and to make it conform to our model. Although our approach is general enough to work with multiple TDS repositories, in practice, inconsistencies and additional varieties from each repository must be reconciled using custom code. Our work demonstrates that it is possible to build a unified and highly efficient searchable catalog system for large and heterogeneous Earth system data repositories that supports real-time queries; however, every solution has its limitations and costs. In this case, the costs are complexity in software and systems architecture, which means increased software development and maintenance costs.

8. Conclusions

This paper proposed and demonstrated a novel cyberinfrastructure-based cataloging solution to enable an efficient two-step search on big climatic datasets by leveraging the existing data centers and state-of-art web service technologies. We used the huge datasets served by UCAR THREDDS Data Server (TDS), which serves Petabyte-level ESOM data and updates hundreds of terabytes of data every day, as our study dataset to validate its feasibility. We analyzed the metadata structure in TDS and created an index for data parameters. A developed metadata registration model, which defines constant information, delimits variable information, and exploits spatial and temporal coherence in metadata, was constructed. The model derives a sampling strategy for a high-performance concurrent web crawler bot which is used to mirror the essential metadata of the big data archive without overwhelming network and computing resources. The metadata model, crawler, and standard-compliant catalog service form an incremental search cyberinfrastructure, allowing scientists to near real-time search in big climatic datasets. We experimented with the approach on both UCAR TDS and NCAR RDA TDS, and the results prove that the proposed approach achieves its design goal, which is a significant breakthrough for the current most non-searchable climate data servers. The solution identified redundant information and determined the sampling frequencies to keep unpredictable parts of the source catalog synchronized with our downstream mirror catalog. An automated hierarchical crawler–indexer and a complimentary search system using the pre-existing EarthCube CyberConnector were implemented. Metadata crawling and access performance validates our integrated approach as an effective method for dealing with big data challenges posed by heterogeneous, real-time Earth System Observation and Model data. However, although the proposed approach outperforms the traditional searching solution for big data, it is still time-consuming in both crawling and searching processes, and may be out of pace dealing with real-time streaming data. In the future, we will study to further reduce the time spent in crawling redundant metadata and to find a high-performance method for rapid and intelligent search.

Author Contributions: Conceptualization, Liping Di and Ziheng Sun; methodology, Ziheng Sun and Juozas Gaigalas; software, Juozas Gaigalas and Ziheng Sun; validation, Juozas Gaigalas, Ziheng Sun; formal analysis, Liping Di, Ziheng Sun, and Juozas Gaigalas; investigation, Juozas Gaigalas; resources, Liping Di; data curation, Juozas Gaigalas, Ziheng Sun, and Liping Di; writing—original draft preparation, Juozas Gaigalas; writing–review an Editing, Liping Di and Ziheng Sun; visualization, Juozas Gaigalas and Ziheng Sun; supervision, Liping Di and Ziheng Sun; project administration, Liping Di and Ziheng Sun; funding acquisition, Liping Di.

Funding: This research was funded by the National Science Foundation, grant number AGS-1740693 & CNS-1739705; PI: Liping Di.

Acknowledgments: We sincerely thank UCAR, UCAR Unidata Support Team, and the authors of the software, libraries, tools, and datasets that we have used in this work.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wright, D.J.; Wang, S. The emergence of spatial cyberinfrastructure. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 5488–5491. [CrossRef] [PubMed]
2. *CyberinfrastruCture Vision for 21st Century DisCoVery*; National Science Foundation Cyberinfrastructure Council: Arlington, VA, USA, 2007.
3. Yang, C.; Goodchild, M.; Gahegan, M. Geospatial Cyberinfrastructure: Past, present and future. *Comput. Environ. Urban Syst.* **2010**, *34*, 264–277. [CrossRef]
4. Yue, P.; Gong, J.; Di, L.; Yuan, J.; Sun, L.; Sun, Z.; Wang, Q. GeoPW: Laying Blocks for the Geospatial Processing Web. *Trans. GIS* **2010**, *14*, 755–772. [CrossRef]
5. Di, L. Geospatial Sensor Web and Self-adaptive Earth Predictive Systems (SEPS). In Proceedings of the Earth Science Technology Office (ESTO)/Advanced Information System Technology (AIST) Sensor Web Principal Investigator (PI) Meeting, San Diego, CA, USA, 13 February 2007.

6. Zhao, P.; Yu, G.; Di, L. Geospatial Web Services. In *Emerging Spatial Information Systems and Applications*, 1st ed.; IGI Global: Hershey, PA, USA, 2006; pp. 1–35.
7. Shukla, J.; Palmer, T.N.; Hagedorn, R.; Hoskins, B.; Kinter, J.; Marotzke, J.; Miller, M.; Slingo, J.; Shukla, J.; Palmer, T.N.; et al. Toward a New Generation of World Climate Research and Computing Facilities. *Bull. Am. Meteorol. Soc.* **2010**, *91*, 1407–1412. [CrossRef]
8. Sherretz, L.A.; Fulker, D.W. Unidata: Enabling Universities to Acquire and Analyze Scientific Data. *Bull. Am. Meteorol. Soc.* **1988**, *69*, 373–376. [CrossRef]
9. Schnase, J.L.; Duffy, D.Q.; Tamkin, G.S.; Nadeau, D.; Thompson, J.H.; Grieg, C.M.; McInerney, M.A.; Webster, W.P. MERRA Analytic Services: Meeting the Big Data challenges of climate science through cloud-enabled Climate Analytics-as-a-Service. *Comput. Environ. Urban Syst.* **2017**, *61*, 198–211. [CrossRef]
10. Khan, M.A.; Uddin, M.F.; Gupta, N. Seven V's of Big Data understanding Big Data to extract value. In Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education, Bridgeport, CT, USA, 3–5 April 2014; pp. 1–5.
11. Habermann, T. Metadata Life Cycles, Use Cases and Hierarchies. *Geosciences* **2018**, *8*, 179. [CrossRef]
12. Greenberg, J. Metadata and the World Wide Web. *Encycl. Libr. Inf. Sci.* **2003**, *3*, 1876–1888.
13. Li, S.; Dragicevic, S.; Castro, F.A.; Sester, M.; Winter, S.; Coltekin, A.; Pettit, C.; Jiang, B.; Haworth, J.; Stein, A.; et al. Geospatial big data handling theory and methods: A review and research challenges. *ISPRS J. Photogramm. Remote Sens.* **2016**, *115*, 119–133. [CrossRef]
14. Bernard, L.; Mäs, S.; Müller, M.; Henzen, C.; Brauner, J. Scientific geodata infrastructures: Challenges, approaches and directions. *Int. J. Digit. Earth* **2014**, *7*, 613–633. [CrossRef]
15. Domenico, B.; Caron, J.; Davis, E.; Kambic, R.; Nativi, S. *Thematic Real-Time Environmental Distributed Data Services (THREDDS): Incorporating Interactive Analysis Tools into NSDL*; Multimedia Research Group, University of Southampton: Southampton, UK, 1997; Volume 2.
16. John Caron, U.; Davis, E. UNIDATA's THREDDS data server. In Proceedings of the 22nd International Conference on Interactive Information Processing Systems for Meteorology, Oceanography, and Hydrology, Atlanta, GA, USA, 27 January–3 February 2006.
17. Sun, Z.; Di, L.; Hao, H.; Wu, X.; Tong, D.Q.; Zhang, C.; Virgei, C.; Fang, H.; Yu, E.; Tan, X.; et al. CyberConnector: A service-oriented system for automatically tailoring multisource Earth observation data to feed Earth science models. *Earth Sci. Inform.* **2018**, *11*, 1–17. [CrossRef]
18. Di, L.; Sun, Z.; Yu, E.; Song, J.; Tong, D.; Huang, H.; Wu, X.; Domenico, B. Coupling of Earth science models and earth observations through OGC interoperability specifications. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 3602–3605.
19. Di, L.; Sun, Z.; Zhang, C. Facilitating the Easy Use of Earth Observation Data in Earth system Models through CyberConnector. In Proceedings of the AGU Fall Meeting, New Orleans, LA, USA, 11–15 December 2017. Abstract #IN21D-0072.
20. Sun, Z.; Di, L. CyberConnector COVALI: Enabling inter-comparison and validation of Earth science models. In Proceedings of the AGU Fall Meeting, Washington, DC, USA, 10–14 December 2018. Abstract #IN23B-0780.
21. Schellnhuber, H.J. 'Earth system' analysis and the second Copernican revolution. *Nature* **1999**, *402*, C19–C23. [CrossRef]
22. Calvin, K.; Bond-Lamberty, B. Integrated human-earth system modeling—State of the science and future directions. *Environ. Res. Lett.* **2018**, *13*, 063006. [CrossRef]
23. Hurrell, J.W.; Holland, M.M.; Gent, P.R.; Ghan, S.; Kay, J.E.; Kushner, P.J.; Lamarque, J.-F.; Large, W.G.; Lawrence, D.; Lindsay, K.; et al. The Community Earth system Model: A Framework for Collaborative Research. *Bull. Am. Meteorol. Soc.* **2013**, *94*, 1339–1360. [CrossRef]
24. Reid, W.V.; Bréchignac, C.; Tseh Lee, Y. Earth system research priorities. *Science* **2009**, *325*, 245. [CrossRef] [PubMed]
25. Lovelock, J. Gaia: The living Earth. *Nature* **2003**, *426*, 769–770. [CrossRef]
26. Holm, P.; Goodsite, M.E.; Cloetingh, S.; Agnoletti, M.; Moldan, B.; Lang, D.J.; Leemans, R.; Moeller, J.O.; Buendía, M.P.; Pohl, W.; et al. Collaboration between the natural, social and human sciences in Global Change Research. *Environ. Sci. Policy* **2013**, *28*, 25–35. [CrossRef]
27. Burnett, K.; Ng, K.B.; Park, S. A comparison of the two traditions of metadata development. *J. Am. Soc. Inf. Sci.* **1999**, *50*, 1209–1217. [CrossRef]

28. Di, L.; Moe, K.L.; Yu, G. Metadata requirements analysis for the emerging Sensor Web. *Int. J. Digit. Earth* **2009**, *2*, 3–17. [CrossRef]
29. Di, L.; Schlesinger, B.M.; Kobler, B.U.S. *Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata*; Federal Geographic Data Committee: Reston, VA, USA, 2000.
30. Yue, P.; Gong, J.; Di, L. Augmenting geospatial data provenance through metadata tracking in geospatial service chaining. *Comput. Geosci.* **2010**, *36*, 270–281. [CrossRef]
31. Di, L.; Kobler, B. NASA Standards for Earth Remote Sensing Data. *Int. Arch. Photogramm. Remote Sens.* **2000**, *33*, 147–155.
32. Yue, P.; Sun, Z.; Gong, J.; Di, L.; Lu, X. A provenance framework for Web geoprocessing workflows. In Proceedings of the 2011 IEEE International Geoscience and Remote Sensing Symposium, Vancouver, BC, Canada, 24–29 July 2011; pp. 3811–3814.
33. Sun, Z.; Yue, P.; Di, L. GeoPWTManager: A task-oriented web geoprocessing system. *Comput. Geosci.* **2012**, *47*, 34–45. [CrossRef]
34. Sun, Z.; Yue, P.; Lu, X.; Zhai, X.; Hu, L. A Task Ontology Driven Approach for Live Geoprocessing in a Service-Oriented Environment. *Trans. GIS* **2012**, *16*, 867–884. [CrossRef]
35. Sun, Z.; Peng, C.; Deng, M.; Chen, A.; Yue, P.; Fang, H.; Di, L. Automation of Customized and Near-Real-Time Vegetation Condition Index Generation Through Cyberinfrastructure-Based Geoprocessing Workflows. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 4512–4522. [CrossRef]
36. Tan, X.; Di, L.; Deng, M.; Huang, F.; Ye, X.; Sha, Z.; Sun, Z.; Gong, W.; Shao, Y.; Huang, C. Agent-as-a-service-based geospatial service aggregation in the cloud: A case study of flood response. *Environ. Model. Softw.* **2016**, *84*, 210–225. [CrossRef]
37. Sun, Z.; Di, L.; Zhang, C.; Fang, H.; Yu, E.; Lin, L.; Tang, J.; Tan, X.; Liu, Z.; Jiang, L.; et al. Building robust geospatial web services for agricultural information extraction and sharing. In Proceedings of the 2017 6th International Conference on Agro-Geoinformatics, Fairfax, VA, USA, 7–10 August 2017; pp. 1–4.
38. Tan, X.; Di, L.; Deng, M.; Chen, A.; Sun, Z.; Huang, C.; Shao, Y.; Ye, X. Agent-and Cloud-Supported Geospatial Service Aggregation for Flood Response. *ISPRS Ann Photogramm. Remote Sens. Spat. Inf. Sci.* **2015**, *2*, 13–18. [CrossRef]
39. Jiang, L.; Sun, Z.; Qi, Q.; Zhang, A. Spatial Correlation between Traffic and Air Pollution in Beijing. *Prof. Geogr.* **2019**, *71*, 654–667. [CrossRef]
40. Liang, L.; Geng, D.; Huang, T.; Di, L.; Lin, L.; Sun, Z. VCI-based Analysis of Spatio-temporal Variations of Spring Drought in China from 1981 to 2015. In Proceedings of the 2019 8th International Conference on Agro-Geoinformatics (Agro-Geoinformatics), Stanbul, Turkey, 16–19 July 2019; pp. 1–6.
41. Zhong, S.; Di, L.; Sun, Z.; Xu, Z.; Guo, L. Investigating the Long-Term Spatial and Temporal Characteristics of Vegetative Drought in the Contiguous United States. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 836–848. [CrossRef]
42. Zhong, S.; Xu, Z.; Sun, Z.; Yu, E.; Guo, L.; Di, L. Global vegetative drought trend and variability analysis from long-term remotely sensed data. In Proceedings of the 2019 8th International Conference on Agro-Geoinformatics (Agro-Geoinformatics), Istanbul, Turkey, 16–19 July 2019; pp. 1–6.
43. Bai, Y.; Di, L. Providing access to satellite imagery through OGC catalog service interfaces in support of the Global Earth Observation System of Systems. *Comput. Geosci.* **2011**, *37*, 435–443. [CrossRef]
44. Chen, Z.; Chen, N. Use of service middleware based on ECHO with CSW for discovery and registry of MODIS data. *Geo-Spat. Inf. Sci.* **2010**, *13*, 191–200. [CrossRef]
45. Bai, Y.; Di, L.; Wei, Y. A taxonomy of geospatial services for global service discovery and interoperability. *Comput. Geosci.* **2009**, *35*, 783–790. [CrossRef]
46. Chen, N.; Di, L.; Yu, G.; Gong, J.; Wei, Y. Use of ebRIM-based CSW with sensor observation services for registry and discovery of remote-sensing observations. *Comput. Geosci.* **2009**, *35*, 360–372. [CrossRef]
47. Di, L.; Yu, G.; Shao, Y.; Bai, Y.; Deng, M.; McDonald, K.R. Persistent WCS and CSW services of GOES data for GEOSS. In Proceedings of the 2010 IEEE International Geoscience and Remote Sensing Symposium, Honolulu, HI, USA, 25–30 July 2010; pp. 1699–1702.
48. Hu, C.; Di, L.; Yang, W. The research of interoperability in spatial catalogue service between CSW and THREDDS. In Proceedings of the 2009 17th International Conference on Geoinformatics, Fairfax, VA, USA, 12–14 August 2009; pp. 1–5.

49. Bai, Y.; Di, L.; Chen, A.; Liu, Y.; Wei, Y. Towards a Geospatial Catalogue Federation Service. *Photogramm. Eng. Remote Sens.* **2007**, *73*, 699–708. [CrossRef]
50. Spéry, L.; Claramunt, C.; Libourel, T. A Spatio-Temporal Model for the Manipulation of Lineage Metadata. *Geoinformatica* **2001**, *5*, 51–70. [CrossRef]
51. She, J.; Feng, X.; Liu, B.; Xiao, P.; Wang, P. *Conceptual Data Modeling on the Evolution of the Spatiotemporal Object*; Chen, J., Pu, Y., Eds.; International Society for Optics and Photonics: The Hague, The Netherlands, 2007; Volume 6753, p. 67530H.
52. Simmhan, Y.L.; Plale, B.; Gannon, D. A survey of data provenance in e-science. *ACM SIGMOD Rec.* **2005**, *34*, 31–36. [CrossRef]
53. Sun, Z.; Yue, P.; Hu, L.; Gong, J.; Zhang, L.; Lu, X. GeoPWProv: Interleaving Map and Faceted Metadata for Provenance Visualization and Navigation. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 5131–5136.
54. Di, L.; Shao, Y.; Kang, L. Implementation of Geospatial Data Provenance in a Web Service Workflow Environment with ISO 19115 and ISO 19115-2 Lineage Model. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 5082–5089.
55. West, L.A.; Hess, T.J. Metadata as a knowledge management tool: Supporting intelligent agent and end user access to spatial data. *Decis. Support Syst.* **2002**, *32*, 247–264. [CrossRef]
56. Nogueras, J.; Zarazaga, F.J.; Muro, R.P. Interoperability between metadata standards. In *Geographic Information Metadata for Spatial Data Infrastructures*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 89–127.
57. Zhao, P. *Geospatial Web Services: Advances in Information Interoperability: Advances in Information Interoperability*; IGI Global: Hershey, PA, USA, 2010.
58. Haslhofer, B.; Klas, W. A survey of techniques for achieving metadata interoperability. *ACM Comput. Surv.* **2010**, *42*, 7. [CrossRef]
59. Wei, Y.; Di, L.; Zhao, B.; Liao, G.; Chen, A. Transformation of HDF-EOS metadata from the ECS model to ISO 19115-based XML. *Comput. Geosci.* **2007**, *33*, 238–247. [CrossRef]
60. Di, L. The development of remote-sensing related standards at FGDC, OGC, and ISO TC 211. In Proceedings of the 2003 IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2003), Toulouse, France, 21–25 July 2003; Volume 1, pp. 643–647.
61. Di, L. Distributed geospatial information services-architectures, standards, and research issues. *Int Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2004**, *35 Pt 2*.
62. ISO. *ISO 19115: Geographic Information—Metadata*; ISO: Geneva, Switzerland, 2013.
63. Bhattacharya, A.; Culler, D.E.; Ortiz, J.; Hong, D.; Whitehouse, K.; Culler, D. *Enabling Portable Building Applications through Automated Metadata Transformation*; University of California at Berkeley: Berkeley, CA, USA, 2014.
64. Nogueras-Iso, J.; Zarazaga-Soria, F.J.J.; Be´jarbe´jar, R.; Lvarez, P.J.A.; Muro-Medrano, P.R.R.; Béjar, R.; Álvarez, P.J.; Muro-Medrano, P.R.R. OGC Catalog Services: A key element for the development of Spatial Data Infrastructures. *Comput. Geosci.* **2005**, *31*, 199–209. [CrossRef]
65. Sun, Z.; Di, L.; Gaigalas, J. SUIS: Simplify the use of geospatial web services in environmental modelling. *Environ. Model. Softw.* **2019**, *119*, 228–241. [CrossRef]
66. Singh, G.; Bharathi, S.; Chervenak, A.; Deelman, E.; Kesselman, C.; Manohar, M.; Patil, S.; Pearlman, L. A Metadata Catalog Service for Data Intensive Applications. In Proceedings of the 2003 ACM/IEEE Conference on Supercomputing, Phoenix, AZ, USA, 15–21 November 2003.
67. Tan, X.; Di, L.; Deng, M.; Fu, J.; Shao, G.; Gao, M.; Sun, Z.; Ye, X.; Sha, Z.; Jin, B. Building an Elastic Parallel OGC Web Processing Service on a Cloud-Based Cluster: A Case Study of Remote Sensing Data Processing Service. *Sustainability* **2015**, *7*, 14245–14258. [CrossRef]
68. Desai, K.; Devulapalli, V.; Agrawal Asst, S.; Kathiria Asst, P.; Patel Professor, A. Web Crawler: Review of Different Types of Web Crawler, Its Issues, Applications and Research Opportunities. *Int. J. Adv. Res. Comput. Sci.* **2017**, *8*, 1199–1202.
69. Li, W.; Wang, S.; Bhatia, V. PolarHub: A large-scale web crawling engine for OGC service discovery in cyberinfrastructure. *Comput Environ Urban Syst.* **2016**, *59*, 195–207. [CrossRef]
70. Pallickara, S.L.; Pallickara, S.; Zupanski, M.; Sullivan, S. Efficient Metadata Generation to Enable Interactive Data Discovery over Large-scale Scientific Data Collections. In Proceedings of the 2010 IEEE Second International Conference on Cloud Computing Technology and Science, Indianapolis, IN, USA, 30 November–3 December 2010.

71. Lopez, L.A.; Khalsa, S.J.S.; Duerr, R.; Tayachow, A.; Mingo, E. The BCube Crawler: Web Scale Data and Service Discovery for EarthCube. In Proceedings of the AGU Fall Meeting, San Francisco, CA, USA, 15–19 December 2014. Abstracts IN51C-06.
72. Khalsa, S.J.S. Data and Metadata Brokering–Theory and Practice from the BCube Project. *Data Sci. J.* **2017**, *16*, 1–8. [CrossRef]
73. Song, J.; Di, L. Near-Real-Time OGC Catalogue Service for Geoscience Big Data. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 337. [CrossRef]
74. Unidata THREDDs Client Catalog Spec 1.0.7. Available online: https://www.unidata.ucar.edu/software/tds/current/catalog/InvCatalogSpec.html (accessed on 26 August 2019).
75. Unidata THREDDS Support [THREDDS #BIA-775104]: Unidata THREDDs Metadata Structure and Volume. Juozasgaigalas@gmail.com. Gmail. Available online: https://mail.google.com/mail/u/0/#search/Unidata+THREDDs+metadata+structure+and+volume/FMfcgxvwzcCgSZmpPZsQFqdjLlCkPNfm (accessed on 26 August 2019).
76. Ansari, S.; Del Greco, S.; Kearns, E.; Brown, O.; Wilkins, S.; Ramamurthy, M.; Weber, J.; May, R.; Sundwall, J.; Layton, J.; et al. Unlocking the Potential of NEXRAD Data through NOAA's Big Data Partnership. *Bull. Am. Meteorol. Soc.* **2018**, *99*, 189–204. [CrossRef]
77. Theodoridis, Y.; Sellis, T.; Papadopoulos, A.N.; Manolopoulos, Y. Specifications for Efficient Indexing in Spatiotemporal Databases. In Proceedings of the Tenth International Conference on Scientific and Statistical Database Management, Capri, Italy, 3 July 1998.
78. Zhang, C.; Di, L.; Sun, Z.; Lin, L.; Yu, E.G.; Gaigalas, J. Exploring cloud-based Web Processing Service: A case study on the implementation of CMAQ as a Service. *Environ. Model. Softw.* **2019**, *113*, 29–41. [CrossRef]
79. Aronson, E.; Ferrini, V.; Gomez, B. *Geoscience 2020: Cyberinfrastructure to Reveal the Past, Comprehend the Present, and Envision the Future*; National Science Foundation: Alexandria, VA, USA, 2015.
80. Heiss, W.H.; McGrew, D.L.; Sirmans, D. Nexrad: Next generation weather radar (WSR-88D). *Microw. J.* **1990**, *33*, 79–89.
81. Bromwich, D.H.; Wilson, A.B.; Bai, L.; Liu, Z.; Barlage, M.; Shih, C.-F.; Maldonado, S.; Hines, K.M.; Wang, S.-H.; Woollen, J.; et al. The Arctic System Reanalysis, Version 2. *Bull. Am. Meteorol. Soc.* **2018**, *99*, 805–828. [CrossRef]