Developing an open data portal for the ESA climate change initiative

Article

Published Version

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

To link to this article DOI: http://dx.doi.org/10.5334/dsj-2020-016

Publisher: Codata

www.reading.ac.uk/centaur

RESEARCH PAPER

Developing an Open Data Portal for the ESA Climate Change Initiative

Philip Kershaw[1,2], Kevin Halsall[3], Bryan N. Lawrence[4,5], Victoria Bennett[1,2], Steve Donegan[1,2], Alan Iwi[1,4], Martin Juckes[1,4], Eduardo Pechorro[6], Ruth Petrie[1,4], Joe Singleton[1], Ag Stephens[1,4], Alison Waterfall[1,2], Antony Wilson[7] and Alexander Wood[8]

[1] Centre for Environmental Data Analysis, RAL Space, STFC Rutherford Appleton Laboratory, Chilton, Didcot, UK

[2] National Centre for Earth Observation, UK

[3] Telespazio VEGA UK Ltd, 350 Capability Green, Luton, Bedfordshire, UK

[4] National Centre for Atmospheric Science, UK

[5] Department of Computer Science, University of Reading, Reading, UK

[6] ESA Climate Office, ESCAT, Harwell Campus, Chilton, Didcot, UK

[7] Scientific Computing, STFC Rutherford Appleton Laboratory, Chilton, Didcot, UK

[8] CGI Space, Defence and Intelligence, CGI, Leatherhead, Surrey, UK

Corresponding author: Philip Kershaw (philip.kershaw@stfc.ac.uk)

We introduce the rationale for, and architecture of, the European Space Agency Climate Change Initiative (CCI) Open Data Portal (http://cci.esa.int/data/). The Open Data Portal hosts a set of richly diverse datasets – 13 "Essential Climate Variables" – from the CCI programme in a consistent and harmonised form and to provides a single point of access for the (>100 TB) data for broad dissemination to an international user community. These data have been produced by a range of different institutions and vary across both scientific and spatio-temporal characteristics. This heterogeneity of the data together with the range of services to be supported presented significant technical challenges.

An iterative development methodology was key to tackling these challenges: the system developed exploits a workflow which takes data that conforms to the CCI data specification, ingests it into a managed archive and uses both manual and automatically generated metadata to support data discovery, browse, and delivery services. It utilises both Earth System Grid Federation (ESGF) data nodes and the Open Geospatial Consortium Catalogue Service for the Web (OGC-CSW) interface, serving data into both the ESGF and the Global Earth Observation System of Systems (GEOSS). A key part of the system is a new vocabulary server, populated with CCI specific terms and relationships which integrates OGC-CSW and ESGF search services together, developed as part of a dialogue between domain scientists and linked data specialists. These services have enabled the development of a unified user interface for graphical search and visualisation – the CCI Open Data Portal Web Presence.

1. Introduction

The European Space Agency (ESA) Open Data Portal (http://cci.esa.int/data/) has been developed to meet the objective of disseminating data outputs from the ESA Climate Change Initiative (CCI) programme (Hollmann et al., 2013). The CCI programme was initiated as a contribution towards the goal of the United Nations Framework Convention on Climate Change to create a database of Essential Climate Variables (ECVs) and to more fully exploit the long-term global archives of earth observation data available from ESA and its member states. In particular, the programme aims to provide stable, long-term, satellite based data products
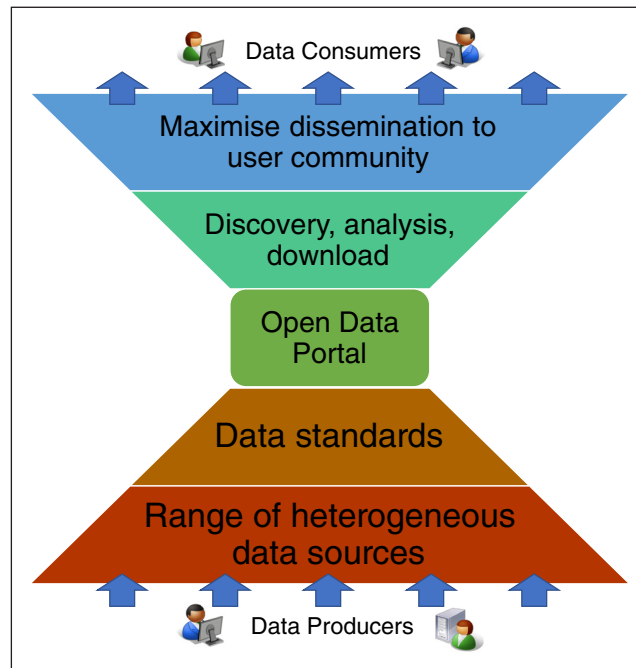
Figure 1: The double challenge of reducing the scope for both producers and consumers in their development and exploitation of essential data variables: from heterogeneous sources to broad dissemination.

for climate modellers and researchers. Much of the programme is devoted to the development and maintenance of the ECV data products, but a key activity is the archival and dissemination of those data via the *CCI Open Data Portal* (at the time of this project, hosting 13 ECVs from 130 datasets involving 40 institutions).

The main contribution of the work described in this paper is to provide a detailed description of how the portal software was architected to meet the dual challenges of heterogeneous data management at scale and delivering a prescribed set of services for discovery and open access. The former requires curating and hosting a complex and varied set of climate data products in a harmonised and consistent manner. The latter requires high performance dissemination to the international user community and the effective integration of independent technologies with different conceptual approaches to solve similar problems. Both data and service delivery depend on standards, yet begin with diversity leading to fundamental challenges associated with narrowing the scope and managing the expectations of both producers and consumers (**Figure 1**). The entire activity depends on the data management and curation workflow as will be seen.

In the remainder of this paper we introduce the ECVs, the CCI programme and context. We discuss the available technologies and relevant prior work. We then present the CCI portal architecture and the necessary workflow. We describe the services and the Web Presence itself. Finally in conclusion, we discuss the issues which arose and the lessons learned. There is much future work required.

2. Background

The Open Data Portal delivers on the CCI programme goal to support modellers by providing ECV data in a format suitable for use in the development and evaluation of climate models, for example, in seasonal to decadal prediction (Goddard et al., 2013) or wider earth system model evaluation (Eyring et al., 2016). This section provides background information about the Essential Climate Variables and context for the Open Data Portal (requirements on interoperation, curation and deployment environment).

2.1. Data Inputs: Essential Climate Variables

The Climate Change Initiative data collection consists of Essential Climate Variables derived from satellite observations, covering marine, terrestrial and atmospheric domains. The products are mostly global gridded fields, but also include points or profiles along satellite orbit tracks, Shapefiles[1] (e.g. glacier outlines) and images. The datasets span time periods from a few years to multiple decades. The products have been generated using state of the art scientific algorithms developed by the project teams and are typically updated each year, as algorithms are improved and/or new satellite data are incorporated.

---

[1] Shapefile https://www.esri.com/library/whitepapers/pdfs/shapefile.pdf.

Over 130 discrete datasets have been delivered to date, with some ECVs producing many variants (Hollmann et al., 2013) such as products using different satellite datasets, or different algorithms, or the resulting data projected onto different grids, or different spatial or temporal resolutions. The majority of data products are gridded but with the notable exceptions of glaciers and some of the icesheet data which are vector based. Products are available to a range of the standard earth observation processing levels: Level 2 – geophysical variables derived from measurements but in spatio-temporal scale of the original acquisition; Level 3 – variables mapped to a uniform spatio-temporal scale and Level 4 – datasets created from the analysis of lower level data that result in gridded, gap-free products. Depending on the ECV or product, individual dataset sizes are between a few gigabytes and tens of terabytes. The data are used widely in the Earth Observation and Climate community, alone or in combination with model data to increase understanding of the earth system. Satellite observations of climate variables are increasingly used to evaluate or initialise climate models, as well as provide a global view of key quantities and their spatial and temporal evolution.

2.2. Open Data Portal Context

ESA set the requirements for the Open Data Portal, following consultation with the climate science and earth observation communities, mandating a single repository and point of access for the data working in the context of a pre-existing system, the Earth System Grid Federation (ESGF) but also the adoption of a range of community standards in order to facilitate broad dissemination of the data. One such standard is the Open Geospatial Consortium (OGC) Catalog Services for the Web (CSW, Nebert et al., 2016), which defines a means for catalogue and search of geospatial data records. By providing such a service it would be possible to integrate the Open Data Portal with initiatives such as the Global Earth Observation System of Systems (GEOSS).

ESGF and GEOSS differ markedly: while both provide access to data, the former uses consistent software and a globally federated system of index and data nodes to provide discovery and download services respectively (WIlliams et al., 2011), the latter provides a brokerage service utilising high level discovery before passing users down to other portals which themselves deliver a range of services (Bai et al., 2012). Both are predicated on two key services: discovery and delivery, but the semantics and nature of those services differ substantially.

In ESGF, data discovery utilises faceted search using a customised metadata schema. Discovery metadata is replicated between nodes in the federation. This enables distributed download from multiple sites in the federation. The primary workflow is data discovery and bulk data download via wget[2] or GridFTP (Allcock et al., 2005), although more sophisticated services such as the Live Access Server (LAS, Hankin et al., 2001) and OPeNDAP (Cornillon et al., 2003) are available for some datasets.

The GEOSS Portal[3] leverages the CSW formalism to aggregate discovery metadata from multiple data repositories into a single web-hosted search interface. A brokering approach (Nativi et al., 2013) provides the bridge between heterogeneous underlying technologies, data standards and disciplines, albeit with fundamental limitations in the level of integration (whilst standardisation of discovery metadata allows search across multiple catalogues, the differences in access conditions, semantics and structure for file access limit what can be achieved). As a consequence, it is generally not possible to deliver federated data services from any of the constituent portals. However, most GEOSS portals support one or both of the core OGC protocols: Web Map Server (WMS, Beaujardiere 2006) and Web Coverage Service (WCS, Peter Baumann 2012) for some or all datasets.

In practice the requirement for the Open Data Portal working in the context of both the ESGF and GEOSS has meant the need to support both interfaces: the Open Data Portal needs to integrate into the Earth System Grid Federation and broker into the GEOSS via the OGC-CSW. ESGF conformance also required conforming to the requirements of the Obs4MIPS project (Teixeira et al., 2014).

2.3. Archival and Curation

Delivering a portal to data relies on having data available and conforming to the necessary formats and standards. This involves having an ingest process into a repository which deals with multiple versions of data sourced from the producers, validates formats and prepares any necessary additional metadata.

The Centre for Environmental Data Analysis (CEDA) provides the necessary archival and curation environment – archival provides reliable data availability including multiple copies of data and curation provides active management of the data, including format and metadata migration (if necessary). Systems have been built up over decades (Pepler and Callaghan, 2015) to deliver reliable data services and complex metadata systems are in place (Parton et al., 2015) which exploit a wide range of information categories from archive metadata, to the necessary browse and discovery metadata (Lawrence et al., 2009). These data systems are

---

[2] GNU Wget, https://www.gnu.org/software/wget/.
[3] GEOSS Portal http://www.geoportal.org/.

supported by data scientists who have expertise in the relevant science areas and can actively curate the metadata content to maximise utility as systems and target communities evolve.

2.4. Deployment Environment

The target environment for the system is the JASMIN data analysis facility (Lawrence et al., 2013), a collaborative environment for the environmental sciences community combining access to the archive with user-managed storage, a cloud and a batch processing system. JASMIN uses a tenancy model to allocate its computing resources to different projects and programmes. The CCI Open Data Portal was deployed using cloud and virtualisation services on JASMIN. The CCI data itself was integrated into CEDA's curated data archive on JASMIN. This consists of a broad range of earth observation and atmospheric science datasets which CEDA provides access to as part of a wider service on behalf of NERC, the Natural Environment Research Council.

3. Architecture

The baseline architecture consists of a central archive, metadata systems, data services, and a web front-end (hereafter referred to as the *Web Presence*) – with data and metadata services consistent with ESGF and GEOSS requirements. This is delivered with customised implementations of the two main ESGF components, the *Index Node* and the *Data Node*.

The Index Node includes an identity provider for user authentication, a metadata catalogue, a portal and search service. The Data Node provides a data publisher, and services for data download and access control, built on the Unidata THREDDS Data Server (TDS)[4] augmented with GridFTP for bulk data download. The key innovation has been to integrate additional OGC services for download and visualisation and reconcile alternative strategies for data discovery with ESGF Search and CSW.

The architecture is summarised in **Figure 2**, there are two CCI user facing interfaces: the Open Data Portal Web Presence itself, and the CCI Toolbox user interface (section 5.4). Additional interfaces include data
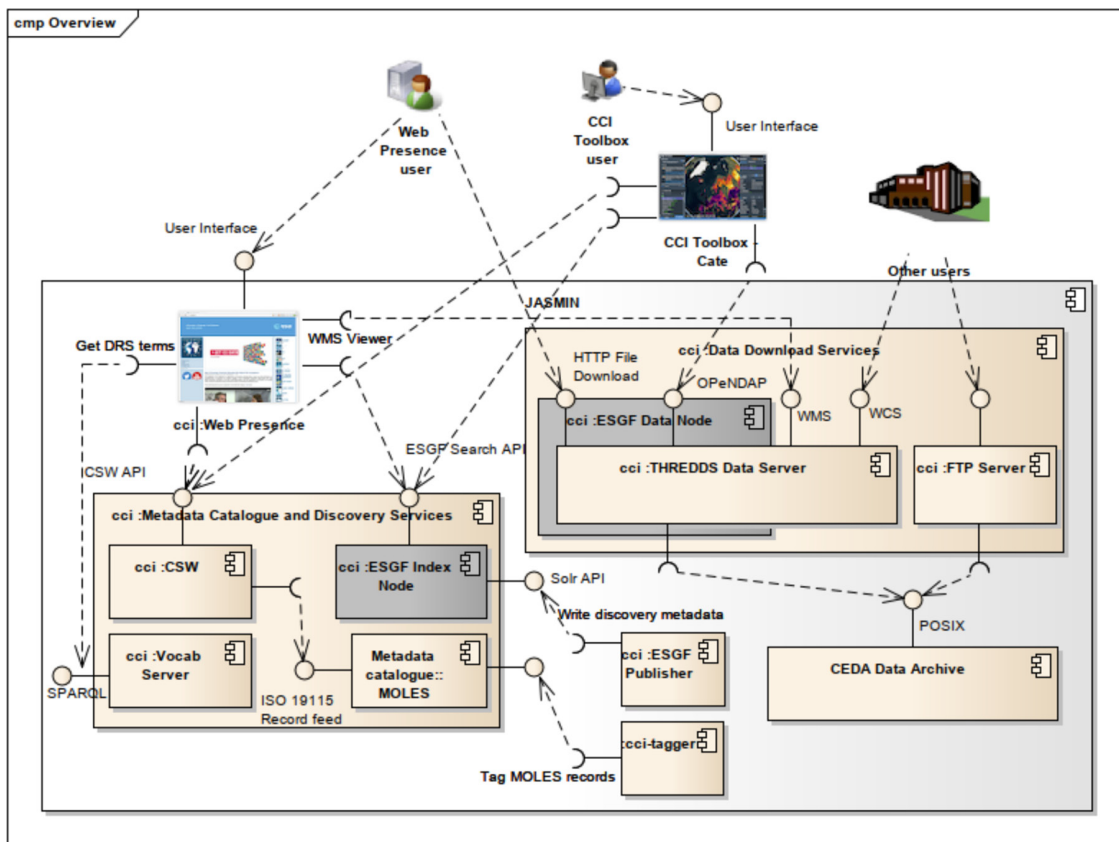
Figure 2: CCI Open Data Portal architecture showing key public and private interfaces, including the CCI Web Presence and Toolbox interface, data services, and other CEDA, ESG and OGC interfaces (all discussed in the text).

---

[4] THREDDS Data Server https://www.unidata.ucar.edu/software/thredds/current/tds/.

delivery services, the CEDA catalogue interface (all the data is ingested into the CEDA archive), and regular ESGF and GEOSS portal interfaces.

Alongside the primary consumers of the architecture (the Web Presence and the CCI Toolbox), the user facing data services fall into four categories:

1. Data discovery and dissemination (the OGC CSW – Catalogue Service for the Web – and ESGF search service[5]);
2. Data download (FTP, GridFTP[†], HTTP);
3. Server-side processing and subsetting (OPeNDAP, GrADS Data Server[†6], LAS[†7], OGC WCS, and OGC WCPS[†8]); and
4. Visualisation services (OGC WMS, LAS[†]).

As the project progressed it became clear that the range of data types precluded easy deployment of some services – those marked with a [†] above were deemed low priority and dropped.

The next section describes the data and metadata workflow necessary to support this architecture, and section 5 describes the services which it exposes.

4. Data and Metadata Workflow

The development of the metadata systems for supporting the workflow necessary to publish data into the portal and out to users formed a major component of the work. This architecture depends on several key underlying technologies: the ESGF Publisher which indexes key metadata from NetCDF source files into the Solr distributed search system;[9] both the ESGF and the Web Presence utilise faceted search based on the Data Reference Syntax (DRS, Petrie et al., 2020); and the metadata systems depend on the CEDA catalogue system (which uses the Metadata Objects for Linking Environmental Sciences, MOLES, Parton et al., 2015).

The publishing workflow involves five key steps: data acquisition, initial metadata creation, metadata customisation, and support for the ESGF and OGC/GEOSS systems. These are summarised in **Figure 3** alongside a depiction of the interaction of the Web Presence interacting with the metadata services. In the rest of this section we provide some detail of the publishing workflow, the service interactions is discussed section 5.
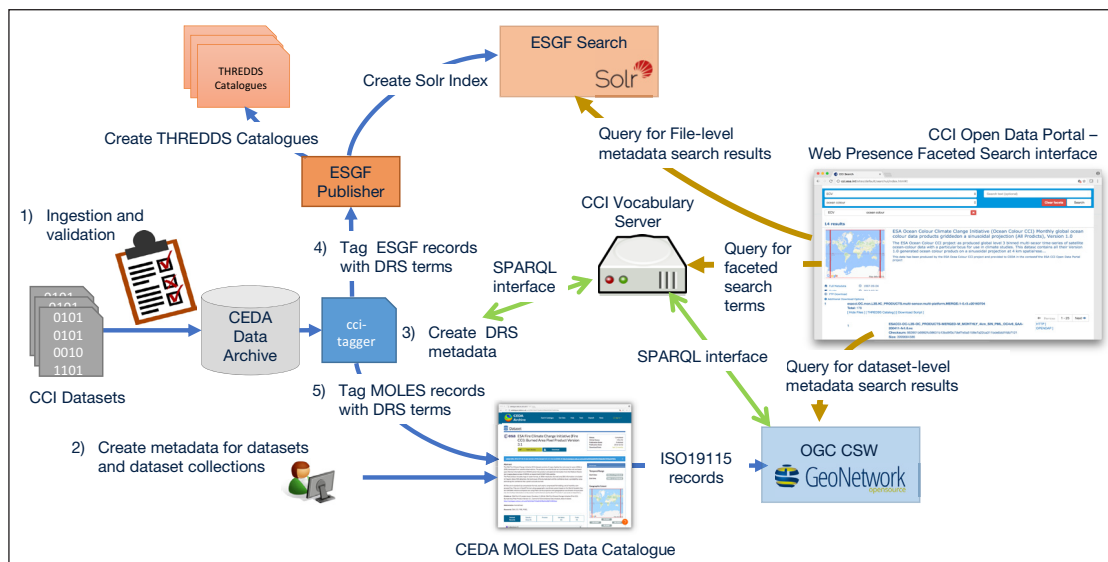
Figure 3: The CCI Open Data Portal data publishing and querying workflow. Publishing consists of five key steps: acquisition of data into the archive (ingestion and validation), archive metadata creation, creation of DRS metadata, then followed by ESGF ingestion, and OGC-CSW integration). Querying can exploit both the ESGF and OGC-CSW interfaces as well as a vocabulary service.

---

[5] ESGF Search API https://www.earthsystemcog.org/projects/cog/esgf_search_restful_api.
[6] GrADS Data Server, http://cola.gmu.edu/grads/gds.php.
[7] Live Access Server https://ferret.pmel.noaa.gov/LAS/.
[8] Web Coverage and Process Service http://www.opengeospatial.org/standards/wcps.
[9] Apache Solr: https://lucene.apache.org/solr.

4.1. Acquisition

Data is first pulled into a staging area for quality control checks, checking conformance with the CCI data standards (ESA Climate Office, 2015) which had been introduced to make data as consistent as possible. It is then formally incorporated into the CEDA archive, exploiting existing CEDA systems which conform to the OAIS archiving standard for managing migration, ensuring fixity of data etc (Corney et al., 2004). This involved storage allocation and organisation into a directory hierarchy, with each CCI dataset mapping to a directory in a POSIX hierarchy.

4.2. Creation of Metadata Records

Metadata is needed to support the CEDA, ESGF and CSW discovery services, as well as support management and data delivery services.

Dataset-level and dataset collection records are manually created in the CEDA data catalogue by data scientists at the granularity of whole datasets. Data products are organised into datasets divided up by the different characteristics (processing level, spatial resolution, instruments, algorithms or product type) of each ECV.

4.3. Unifying metadata approaches

The ESGF discovery and CSW discovery metadata do not share the same semantics, schema, or API, although they are complementary (see **Table 1**). The most important distinction arises from both limitations and advantages of the existing ESGF technology: it only supports publishing gridded CF-NetCDF data, but it supports discovery down to the granularity of individual files. By contrast, the CSW is designed for general data, but discovery is at the level of entire datasets and collections.

The limitation on datatypes in ESGF would be problematic for the CCI, were it not for the alternative CSW cataloguing, as the CCI includes different datatypes (e.g. glacier boundaries) and formats (Shapefiles). By supporting both interfaces, it is possible to both construct a complete catalogue of all CCI datasets for exposure using the CSW and ESGF search systems. The ESGF system complemented CSW by providing a more powerful faceted search interface for the gridded CF-NetCDF data. The two interfaces also support broad dissemination of the data via a) the ESGF distributed infrastructure and b) CSW harvesting of discovery records into, for example, the GEOSS Geoportal.

Much of the development effort was concerned with the evolution of a system to reconcile the two different approaches of the ESGF and the OGC-CSW. However, it was possible to exploit a data reference syntax, DRS, to provide a common logical organisation. A DRS consists of a set of controlled vocabularies which enable any given dataset to be uniquely identified and enables the development of faceted search capability. The first use of a DRS in this mode was for CMIP5 (Taylor et al., 2012), but a CCI specific version was developed, including terms such as the ECV name, the processing level, the instrument used for the retrieval and the sensor platform (Petrie et al., 2020).

The utility of DRS approaches is enhanced by exploiting SKOS[10] to link vocabulary terms and serving the vocabularies and linked data definitions from a vocabulary server. The use of a vocabulary service provides

Table 1: Comparison of ESGF Search and CEDA CSW metadata catalogue and search services.

Catalogue and Search Technology	Granularity of content	Supported data types	Use of controlled vocabularies	Dataset metadata	Links to data access services
OGC CSW/ISO 19115 metadata for CEDA MOLES Metadata Catalogue	Supports collections and Datasets only	Any – dataset and dataset collection information input manually	Limited use of unbound keywords	Full scope of ISO 19115: including abstract, responsible party, licensing and other constraints on use.	Links at the granularity of datasets only. Links provided using OnlineResources. Links have name and description but not classified by service type.
ESGF Search	Datasets and file level metadata	ESGF publisher works with gridded netCDF data only	Per project DRS controlled vocabularies supports faceted search	Limited dataset metadata such as dataset variables information	Links provided at dataset and file level. Links categorised by type e.g. OPeNDAP, HTTPServer, GridFTP

---

[10] Simple Knowledge Organisation System https://www.w3.org/2004/02/skos/.

a single canonical source for information for use by (multiple) search interfaces and any client applications consuming its content. These developments built on experience with the Climate Information Platform for Copernicus (CLIPC, Mihajlovski et al., 2016) and CHARMe (Blower et al., 2014), both of which had demonstrated the advantages of using independent systems to manage controlled vocabularies.

A data modelling exercise was conducted to generate the necessary content. Data scientists entered vocabulary terms and relationships into spreadsheets. Using a special Python script, the information from the sheet cells was converted into SKOS and OWL[11] representations using the Turtle[12] serialisation. The use of SKOS allowed the mapping of concept relationships between geophysical parameters (e.g. sea ice is *broader* than sea ice concentration, and conversely, sea ice concentration is *narrower* than sea ice). SKOS concept schemes were developed for the CCI DRS (**Figure 4**) and the full 50 ECVs defined by the Global Climate Observing System, GCOS.[13] Where possible pre-existing definitions already available from the NERC Vocabulary Server (Leadbetter, 2012) were used. Once prepared, vocabularies were uploaded to an internal vocabulary server implemented using the open source Apache Jena[14] Fuseki triple store.

4.4. The ESGF Publishing System

The ESGF has its own publishing system which is predicated on gridded CF-NetCDF compliant data providing information for the THREDDS catalogue and the Solr distributed database. After consistency checking, the files can be automatically parsed for metadata extraction; information extracted includes variable detail, the necessary DRS terms and details about spatio-temporal extent. Dataset and file-level information were recorded in both databases along with service endpoint information for access methods such as standard HTTP data download and OPeNDAP.

Given the constraint of gridded CF-NetCDF some datasets were not suitable for publishing through the ESGF system. Such datasets were those which did not conform to a coverage array describing values at a point in space and/or time, such as those describing the spatial extent of a feature (e.g. the Greenland and Antarctic Ice Sheets and Glaciers ECVs).

ESGF publishing was started when the first release of CCI datasets had already been made and the second release was underway. In many cases, datasets did not comply with the data standards. This impacted on ESGF publishing since specific metadata were expected in order to represent the different search facets set by the DRS. The ESGF publisher uses a system of map files to link information from the source files to these facets. In order to deal with the inconsistencies with the data, a special `cci-tagger` script was developed to extract vocabulary terms from the data and implement workarounds for problems such as inconsistent naming for a term, a term being absent completely or terms being defined in filenames, rather than file metadata. The `cci-tagger` script makes heavy use of the CCI vocabulary service.
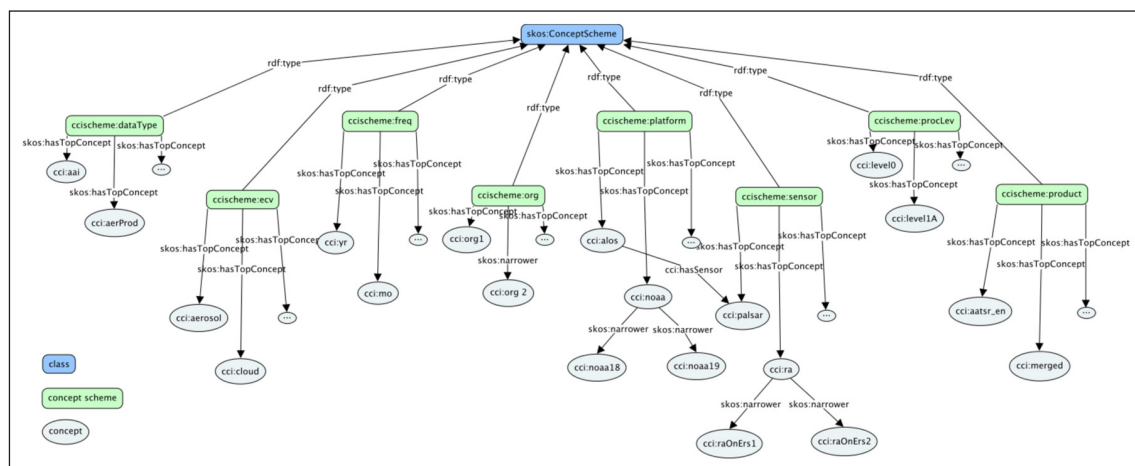
Figure 4: SKOS concept scheme for CCI DRS vocabularies (showing only downward concept mappings). Each is represented as a *Collection* (consistent with conventions adopted for the NVS) and a *ConceptScheme*, to enable bidirectional navigation between *ConceptScheme* and *Concept*.

---

[11] Web Ontology Language https://www.w3.org/OWL/.
[12] Turtle https://www.w3.org/TR/turtle/.
[13] Global Climate Observing System https://gcos.wmo.int/.
[14] Apache Jena https://jena.apache.org/.

4.5. The OGC-CSW Publishing System

The MOLES system underlying the CEDA catalogue supports the export of dataset records serialised as ISO 19115. This capability was used to populate a dedicated project specific OGC CSW interface with CCI meta-data and construct a search interface supporting free text search and results at the granularity of individual datasets. Although project specific, this same interface allowed other systems (e.g. GEOSS) to harvest CCI metadata.

The OGC-CSW records were populated with DRS terms from an extension to the `cci-tagger` script. The resulting ISO 19115 records for the CSW specified the relevant vocabulary SKOS concept in each case and the URI for the source SKOS collection. Where possible these took advantage of a direct mapping between ISO 19115 records returned from the CSW and the related ESGF dataset search results. In practice, there was not always a one-to-one mapping between ISO records and the ESGF dataset records in all cases, with some ISO records containing multiple DRS dataset identifiers.

5. Services

The data itself is hosted by CEDA using the core JASMIN storage, and exposed directly using existing CEDA services. Here we discuss the additional CCI specific services, hosted using a combination of JASMIN's cloud and virtualisation services. We begin by describing the initial data services deployed as the system was developed, then the dashboard development, before describing the Open Data Portal production system as of September 2019.

5.1. Data Access Services

The first development was to deploy data services. Standard FTP access to CCI data was directly available via the CEDA services, all that was necessary for the CCI was to ensure that FTP endpoints for the datasets were entered into the ISO 19115 dataset records served from the OGC-CSW. Each ISO record linked to the FTP service by including an online resource link mapped to the FTP path of the equivalent dataset directory in the data archive.

The THREDDS Data server provided HTTP and OPeNDAP out of the box (for gridded datasets), and WMS and WCS were configured as non-standard options. Although built on organising, finding and delivering files, THREDDS also supports aggregation and the attachment of services at both the file and aggregation level. All the files in a dataset were served individually via HTTP, and as an aggregation – a single contiguous set of data combined along the time axis – via OPeNDAP, WMS and WCS.

The necessary configuration for the aggregations was created by an additional Python script which augmented the THREDDS catalogue and Solr search indexes created by the ESGF publisher by adding the aggregation endpoints using the NetCDF markup language (NcML, Nativi et al., 2005). The NcML specifies the variables to be aggregated and includes the time steps for each file as a measure to optimise load performance. Repeated queries benefit from significant performance improvement by virtue of caching by the HTTP server.

ESA required open access, so no access control was configured (eventually the Web Presence provided the capability for users to register in order to provide a mechanism for them to be kept informed about changes to data and services).
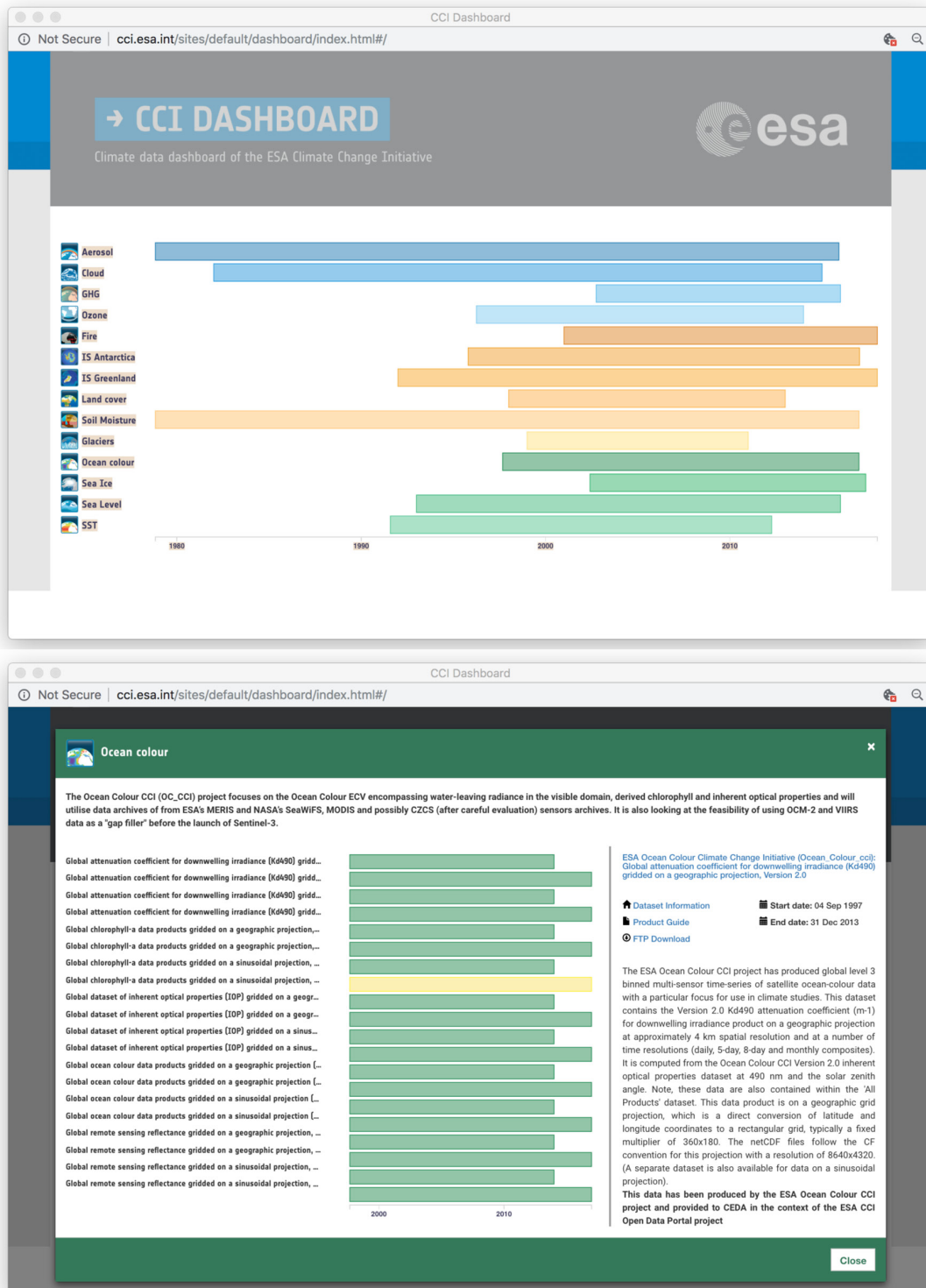
5.2. Portal Dashboard

An early "dashboard" search interface was developed to provide a novel graphical overview of all the ECV data products available and the respective temporal coverage for each (**Figure 5**). In this interface the CSW content was exposed in such a way that the user could select by ECV to see specific ISO 19115 records. This interface now forms part of the larger Web Presence.

5.3. Web Presence

Following the publication of the first data into ESGF services, the CCI Open Data Portal had dual search interfaces. This served to underline the disparity between the two: the CSW provided complete coverage of all datasets but an only rudimentary interface for file access and manipulation through FTP; ESGF Search provided a comprehensive faceted search interface linking to individual file download via HTTP and OPeN-DAP but was limited in its coverage since the ESGF Publisher is only capable of indexing NetCDF files into the search catalogue.

Taking advantage of the controlled vocabularies and vocabulary service however, it was possible to combine ESGF and OGC-CSW to provide a single search interface in the Web Presence which integrated search results from both. (**Figure 6**).

Figure 5: CCI Open Data Portal Dashboard showing the temporal extent of datasets (top panel) and discovery metadata and temporal extent for contents within a specific dataset (ocean colour, bottom panel).

Facets and the available terms for each facet are populated dynamically from the definitions of concepts and collections in the vocabulary service. It can display context-sensitive help text sourced from the 'preferred label' text of the vocabulary terms. When a query selection is made, these are translated into keyword items included in a given CSW query (e.g. **Figure 7**). Keyword items correspond to the URIs of concepts in the appropriate SKOS ConceptScheme. A drill down approach was implemented for the search logic in the user interface. This takes advantage of the fact that the ISO records returned from the CSW contain links to the corresponding ESGF search records:

Figure 6: The Web Presence faceted search interface – integrating content from the Vocabulary Server, CSW and ESGF Search services.

```
<?xml version="1.0" encoding="UTF-8"?>
<csw:GetRecords
 xmlns:csw="http://www.opengis.net/cat/csw/2.0.2"
 xmlns:ogc="http://www.opengis.net/ogc"
 xmlns:gml="http://www.opengis.net/gml/3.2"
 xmlns:dc="http://purl.org/dc/elements/1.1/"
 xmlns:dct="http://purl.org/dc/terms/"
 xmlns:gmd="http://www.isotc211.org/2005/gmd"
 xmlns:gco="http://www.isotc211.org/2005/gco"
 xmlns:geonet="http://www.fao.org/geonetwork"
 service="CSW" version="2.0.2" resultType="results_with_summary"
 outputSchema="http://www.isotc211.org/2005/gmd"
         startPosition="1" maxRecords="1">
<csw:Query typeNames="csw:Record">
  <csw:ElementSetName>full</csw:ElementSetName>
  <csw:Constraint version="1.1.0">
    <ogc:Filter>
      <ogc:And>
        <ogc:PropertyIsEqualTo>
          <ogc:PropertyName>AnyText</ogc:PropertyName>
          <ogc:Literal>%*%</ogc:Literal>
        </ogc:PropertyIsEqualTo>
        <ogc:PropertyIsEqualTo>
          <ogc:PropertyName>keywordUri</ogc:PropertyName>
          <ogc:Literal>http://vocab.ceda.ac.uk/collection/cci/procLev/proc_level4</ogc:Literal>
        </ogc:PropertyIsEqualTo>
        <ogc:PropertyIsEqualTo>
          <ogc:PropertyName>keywordUri</ogc:PropertyName>
          <ogc:Literal>http://vocab.ceda.ac.uk/collection/cci/freq/freq_day</ogc:Literal>
        </ogc:PropertyIsEqualTo>
      </ogc:And>
    </ogc:Filter>
  </csw:Constraint>
</csw:Query>
```

Figure 7: Example CSW query using Processing Level and daily frequency SKOS concepts for keyword searching. This will query for records corresponding to Level 4 processed data (SKOS vocabulary concept http://vocab.ceda.ac.uk/collection/cci/procLev/proc_level4) with a temporal frequency of daily (SKOS vcocabulary concept http://vocab.ceda.ac.uk/collection/cci/freq/freq_day).

1. An initial search query was invoked using the CSW. This returns ISO records containing information about the matching datasets.
2. The ISO records returned are checked for matching ESGF records – stored as DRS identifiers. If found, the ESGF search API is then invoked based on these matching DRS identifiers. The ESGF search query returns the list of files (if any) for that dataset.

5.4. Third party clients

The Open Data portal services can be consumed by third party clients. **Figure 8** shows two such examples: a WMS client developed by Plymouth Marine Laboratory which pulls data from the CCI Open Data Portal, and the CCI ToolBox desktop client – *Cate* – developed as part of a separate ESA project. The latter provides more comprehensive functionality for data analysis and visualisation, using the Open Data Portal search and data services. In particular, it takes advantage of OPeNDAP sub-setting to selectively retrieve portions of data from the portal data archive to the user client's host computer.
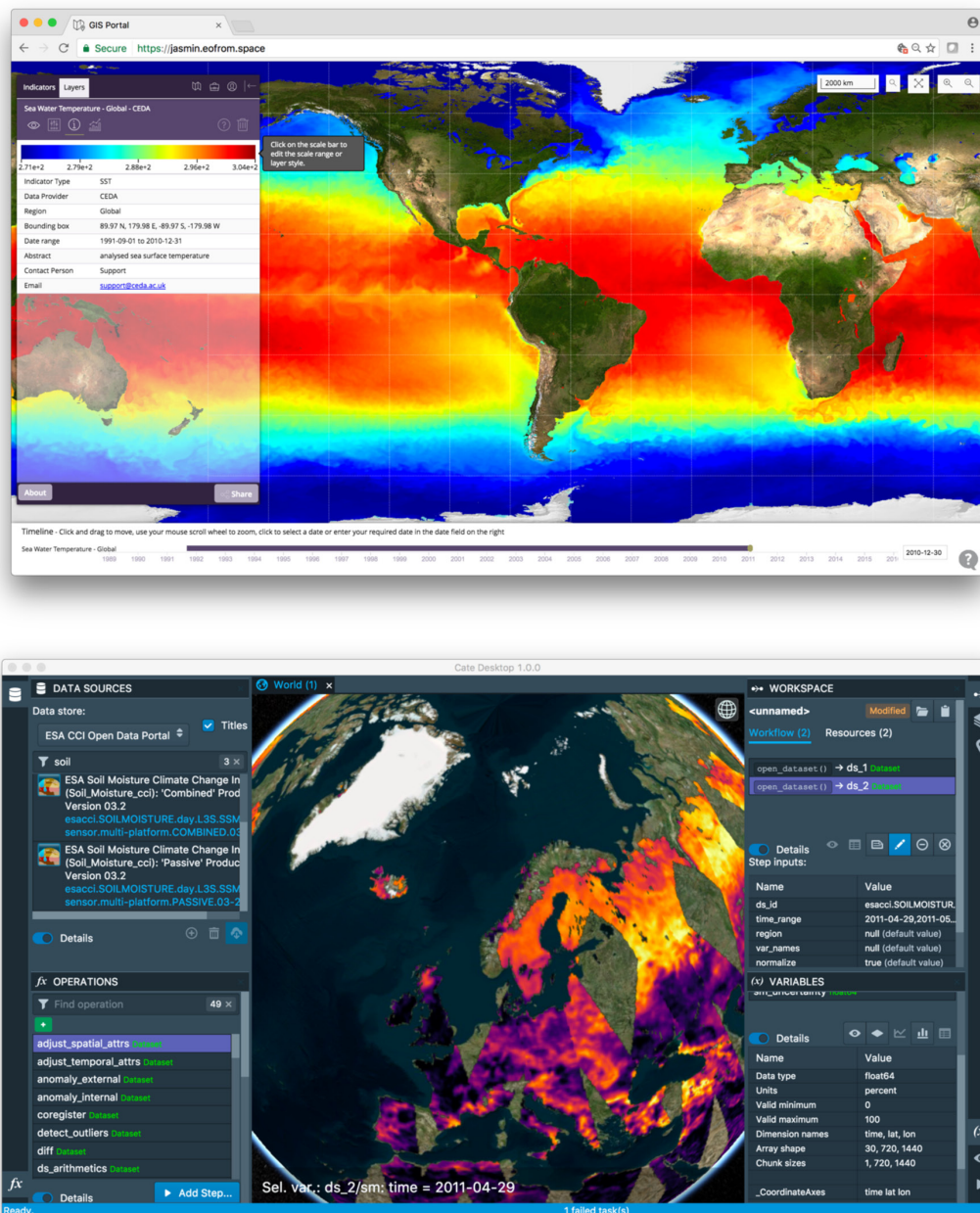
Figure 8: Two examples of third-party applications consuming CCI Open Data Portal services: A WMS client (top panel) developed by PML Applications Ltd showing Sea Surface Temperature and the CCI ToolBox (bottom panel) showing soil moisture data.

6. Discussion and Lessons Learned

The final system was the result of iterative development with capability added via a set of discrete developments through the project. Four key phases were involved:

1. Establishment of architectural baseline, addressing data access and metadata services.
2. Establishing an aggregation policy for files into datasets and using it to deliver both file and dataset delivery services.
3. Addressing inconsistencies in metadata representations which inhibited unified search by defining metadata vocabularies and a vocabulary service to provide a single authoritative source of information for service consumption within the system.
4. Unifying the two search interfaces which arise from the use of both ESGF and CSW in the architecture.

While the use of an iterative approach was partly to meet ESA requirements as they unfolded, it was also a result of the a priori recognition that trying to design upfront for all the data complexity issues would have been inefficient and impractical. The twofold challenges of a developing a single consistent set of services for a large and varied collection of datasets and managing a complex set of requirements for discovery and access services could have consumed endless design time with little practical outcome. Instead, by using existing technologies and a sequence of milestone deliveries, problems were addressed as they were encountered.

There were three main classes of problem: information mismatch between the tools, scale issues and content issues. We give three examples of each.

6.1. Information mismatch

The task of integrating a varied set of data into the metadata catalogue and archiving services has demonstrated the intimate relationship between the metadata content and the services, with services dependent on particular information provision. Unfortunately, different service requirements had led to different metadata standards, and reconciling alternative metadata technologies and strategies to deliver a range of services was non-trivial. The development of vocabulary services based on linked data and a common data reference syntax made this possible. Equally important was the quality and extent of the primary content – services and usability benefitted from attention to detail by data providers and extra metadata information added by data scientists during data ingestion.

6.2. Scale Issues

The sheer volume of files *per dataset* exposed performance limitations in the ESGF Publisher software. The Ocean Colour ECV for example, represents a large data collection which splits into 40 datasets along the lines of the geophysical parameter recorded, the co-ordinate system used and the temporal frequency. Even with this division, the number of files for individual datasets was large; for example, *daily* data involved in excess of 7000 files over the complete time range 1997–2016.

The ESGF Publisher was originally designed for managing CMIP5 data with a very different structure determined by the experiment design. The initial work-around in the case of Ocean Colour, involved splitting the data and reorganising it into smaller temporal chunks so that rather than publishing the dataset as a single 20-year time span, it was divided into smaller datasets. Similar hacks were initially applied to other ECVs. However, a patch was developed for the Publisher such that it was possible to create datasets spanning the full twenty years. In one publishing update in autumn 2018 over a quarter of a million data files were indexed organised into over 100 datasets.

6.3. Content Issues

For some datasets there were last-mile issues with the publication of the data: though ingestion, checks and metadata catalogue generation processes would succeed, minor formatting inconsistencies in NetCDF header file content (e.g. use of non-standards compliant names for the time dimension) would prevent services from operating correctly (e.g. THREDDS being unable to serve OPeNDAP and WMS queries).

In most cases it was possible to correct these by liaising with the respective ECV data producers to get new versions of the data, but these sorts of issues served to underline the need for rigorous checking and enforcement of data standards both by the data providers, and by the ingestion process.

7. Related and Future Work

The THREDDS Data Server (TDS) was selected early in the project to satisfy the requirements to provide HTTP, data download, OPeNDAP, WMS and WCS. This has the added benefit of delivering a powerful aggregation capability (via OPeNDAP), effectively creating a data cube across spatial and temporal dimensions important for analysis of climate datasets. However, the TDS has limited support for the OGC services and future developments may need to exploit other technology such as Geoserver.[15] In addition, whilst it is possible to create large aggregations, the large temporal range (20–30 years in some cases) is at the limit of what the technology can support – at all layers, from the TDS down through the web application server and underlying host capability. Work at CEDA as part of the ESiWACE[16] project is exploring alternatives for scalable systems for data access.

Like the CCI portal activity, the Copernicus Climate Change Service (C3S, Raoult et al., 2017) is a system to provide an authoritative source of quality-assured data to support adaptation and mitigation strategies for climate change. Datasets include climate models, re-analyses as well as ECVs. Rather than collating data together in a single repository, C3S uses a distributed architecture. There is a central web interface and toolbox for server-side analysis of data but the data itself is hosted at remote data providers. A broker and a system of adaptors for the different data sources and computational resources mediate access through to the toolbox and web front end. The use of agreed standards for the adaptor interfaces was important taking advantage of OGC web services, ESGF and OPeNDAP as used in the CCI Open Data Portal.

There is clearly a large overlap between the system developed for C3S and the Open Data Portal. The C3S use of an *adaptor* in its architecture has provided an abstraction between the data source and data consumer. Such a model could have provided some benefits in managing the heterogeneity between the different datasets in CCI. Equally however, CCI has evolved a comprehensive system for managing controlled vocabularies and integrating distinct services for data discovery and access. The decision to host the data in a single repository has benefits for the future evolution of the CCI system to support hosted processing and analysis since with all the data in one place, computing resources can be colocated eliminating the need to move and stage data between infrastructures.

The use of OpenSearch in the earth observation community (CEOS, 2015) shows promise as a means to combine content from multiple categories of metadata into a single search interface. The challenge of granularity of data is explicitly addressed with a drill-down approach to search from *Collection* (dataset or dataset collection) to *granule* (individual file/data product) level. This is achieved by establishing a hierarchy of search endpoints from Collection down to granule search. Dataset metadata can be attached at the collection level in the search hierarchy serialised for example as O & M (ISO 19156 and OGC 10025) observations or using ISO 19139. OpenSearch itself is a simple open standard for data discovery which defines a response based on syndication formats such as Atom and RSS, extending these with search metadata. Search endpoints include a description document which enables clients to introspect the given service and determine the semantics for search queries. Initiatives such as CWIC[17] and FedEO[18] are taking advantage of OpenSearch to create federated search infrastructures linking data repositories from multiple agencies (Miura, 2016).

For discovery, ESGF itself could migrate to the OpenSearch standard. This has been implemented for the CEDA archive ESA Sentinel data products and clearly it could address the problems encountered in the project with the differing semantics and content for OGC CSW and ESGF Search. A future system could provide an OpenSearch service that would serve dataset and file-level response content. Dataset metadata could be represented using the O & M observations or as ISO 19115 records. OpenSearch is being considered for adoption for ESGF in its technical roadmap.

8. Summary

We have built the CCI Open Data Portal ecosystem consisting of a Web Presence exploiting standard APIs for data discovery and access integrated together with standardised vocabulary services exposing new vocabularies developed for the project. The development methodology involved:

- Applying an iterative approach to the development of the software and data management processes, with frequent releases incrementally building capability;

---

[15] Geoserver http://geoserver.org/.
[16] ESiWACE https://www.esiwace.eu/.
[17] CWIC http://ceos.org/ourwork/workinggroups/wgiss/access/cwic/.
[18] FedEO https://eoportal.org/web/eoportal/fedeo-client.

- Prioritising close communication between actors: representatives from the CCI teams providing the data, data scientists ingesting and publishing data, members of the development teams building the sub-systems;
- Integrating conflicting technologies and where possible simplifying by reducing the number of dependencies.

The system depends on a workflow which takes data that conforms to the CCI data specification, ingests it into the CEDA Archive, and uses both manual and automatically generated metadata to support data discovery, browse and delivery services based on the deployment of both Earth System Grid Federation nodes and OGC-CSW metadata catalogues. An integral part of the system is a new vocabulary server, populated with CCI specific terms and relationships developed as part of an iterative dialogue between domain scientists and linked data specialists. Third party applications, including GEOSS and the CCI ToolBox are able to consume the CCI Open Data Portal services.

Table of Acronyms

CEDA	Centre for Environmental Data Analysis
CF-NetCDF	denotes data formatted using the Network Common Data Form (NetCDF) from Unidata which complies with the Climate and Forecast Conventions for describing and naming data variables.
CCI	The ESA Climate Change Initiative
CMIP5	5th Coupled Model Intercomparison Project
CSW	Catalogue Service for the Web – standard from the Open Geospatial Consortium defining a web service interface for discovery of catalogue records that describe geospatial data.
CWIC	CEOS (Committee on Earth Observation Satellites) WGISS (Working Group on Information Systems and Services) Integrated Catalogue
DRS	Data Reference Syntax – a system of controlled vocabularies for describing data defined for datasets hosted in the Earth System Grid Federation.
ECV	Essential Climate Variable (within the ESA CCI).
ESA	European Space Agency
ESGF	Earth System Grid Federation
GEOSS	Global Earth Observation System of Systems
HTTP	Hypertext Transfer Protocol
JASMIN	The UK Joint Analysis supercomputer
MOLES	Metadata Objects for Linking Environmental Sciences – a metadata information model developed for CEDA.
NCML	NetCDF Markup Language
NetCDF	Network Common Data Form – data format and software libraries from Unidata for array-based scientific data
NOSQL	Not-only SQL – Collective name for database technologies which do not follow the strict rules governing traditional relational databases.
NVS	NERC (Natural Environment Research Council) Vocabulary Service.
OBS4MIPS	The Project for Observations for Model Intercomparison Projects
OGC	Open Geospatial Consortium
OPeNDAP	Open-source Project for a Network Data Access Protocol – initiative focussed on the development of services for remote access of gridded data through the DAP (Data Access Protocol) specification.
OWL	Web Ontology Language
SKOS	Simple Knowledge Organisation System
THREDDS	Thematic Real-time Environmental Distributed Data Services – project from Unidata to develop middleware to bridge between data providers and users
WMS	Web Map Service – Open Geospatial Consortium standard defines a web service interface for accessing geo-referenced map images
WCS	Web Coverage Service – Open Geospatial Consortium standard defines a web service interface for service data coverages defined by geo-temporal query parameters.
WCPS	Web Coverage and Processing Service – Open Geospatial Consortium standard which defines a query language for filtering and processing multi-dimensional coverage data.

Acknowledgements

Competing Interests

The authors have no competing interests to declare.

Author Contributions

1. Philip Kershaw: technical architect for metadata catalogue and archive sub-systems. Wrote first draft of manuscript.
2. Bryan Lawrence: inspired the general technical direction, provided technical and scientific context text, led on finalising the manuscript.
3. Victoria Bennett: CCI Open Data Portal Science Leader, author of the data standards document for the CCI datasets.
4. Steve Donegan: data scientist with responsibility for deployment and configuration of the CSW service.
5. Kevin Halsall: CCI Open Data Portal Project Manager.
6. Alan Iwi: software developer, contributed to publishing to ESGF system, fixes to ESGF publishing software, scripting for publishing.
7. Martin Juckes: contributed to the data standards document, development of vocabularies for CCI and related projects CLIPC and CMIP5.
8. Eduardo Pechorro: ESA technical manager for CCI Open Data Portal and CCI Toolbox projects.
9. Ruth Petrie: data scientist, contributed to data modelling.
10. Joe Singleton: software developer, published data through ESGF system, developed configuration for dataset aggregations. Contributed to patches for ESGF publishing software.
11. Ag Stephens: contributed to the technical architecture and implementation and integration of enhancements to MOLES metadata catalogue system.
12. Alison Waterfall: data scientist, managed ingestion of data products and created MOLES system metadata catalogue records.
13. Antony Wilson: software developer, contributed to development of SKOS and OWL models and deployment of Vocabulary and CSW services.
14. Alexander Wood: software developer, responsible for the development of Search and Dashboard services for the Web Presence.

References

Allcock, W, Bresnahan, J, Kettimuthu, R, Link, M, Dumitrescu, C, Raicu, I and **Foster, I.** 2005. The Globus Striped GridFTP Framework and Server. In: *Proceedings of the 2005 ACM/IEEE Conference on Supercomputing. SC '05*, 54. Washington, DC, USA: IEEE Computer Society.

Bai, Y, Di, L, Nebert, DD, Chen, A, Wei, Y, Cheng, X, Shao, Y, Shen, D, Shrestha, R and **Wang, H.** Dec. 2012. GEOSS Component and Service Registry: Design, Implementation and Lessons Learned. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(6): 1678–1686. DOI: https://doi.org/10.1109/JSTARS.2012.2215914

Baumann, P. 2012. OGC Web Coverage Service 2.0.1. Standard 09-110r4. Open Geospatial Consortium.

Beaujardiere, JDL. 2006. OGC Web Map Service V1.3.0. Standard OGC 06-042. *Open Geospatial Consortium.*

Blower, JD, Alegre, R, Bennett, VL, Clifford, DJ, Kershaw, PJ, Lawrence, BN, Lewis, JP, Marsh, K, Nagni, M, O'Neill, A and **Phipps, RA.** Jan. 2014. Understanding Climate Data Through Commentary Metadata: The CHARMe Project. In: Bolikowski, Ł, Casarosa, V, Goodale, P, Houssos, N, Manghi, P, Schirrwagen, J (Eds.), *Theory and Practice of Digital Libraries – TPDL 2013 SelectedWorkshops. No. 416 in Communications in Computer and Information Science*, 28–39. Springer International Publishing. DOI: https://doi.org/10.1007/978-3-319-08425-1_4

CEOS. 2015. CEOS OpenSearch Best Practice Document. Tech. Rep. CEOS-OPENSEARCH-BP-V1.0.1, Committee on Earth Observation Satellites.

Corney, D, De Vere, M, Folks, T, Giaretta, D, Kleese van Dam, K, Lawrence, B, Pepler, S and **Strong, B.** 2004. Applying the OAIS standard to CCLRC's British Atmospheric Data Centre and the Atlas Petabyte Storage Service. In: *UK E-Science Programme All Hands Meeting*, 2004. Nottingham.

Cornillon, P, Gallagher, J and **Sgouros, T.** 2003. OPeNDAP: Accessing data in a distributed, heterogeneous environment. *Data Science Journal*, 2: 164–174. DOI: https://doi.org/10.2481/dsj.2.164

ESA Climate Office. Mar. 2015. Data Standards Requirements for CCI Data Producers (Issue 1, Revision 2). Howpublished: online.

Eyring, V, Gleckler, PJ, Heinze, C, Stouffer, RJ, Taylor, KE, Balaji, V, Guilyardi, E, Joussaume, S, Kindermann, S, Lawrence, BN, Meehl, GA, Righi, M and **Williams, DN.** Nov. 2016. Towards improved and more routine Earth system model evaluation in CMIP. *Earth Syst. Dynam.*, 7(4): 813–830. DOI: https://doi.org/10.5194/esd-7-813-2016

Goddard, L, Kumar, A, Solomon, A, Smith, D, Boer, G, Gonzalez, P, Kharin, V, Merryfield, W, Deser, C, Mason, SJ, Kirtman, BP, Msadek, R, Sutton, R, Hawkins, E, Fricker, T, Hegerl, G, Ferro, CAT, Stephenson, DB, Meehl, GA, Stockdale, T, Burgman, R, Greene, AM, Kushnir, Y, Newman, M, Carton, J, Fukumori, I and **Delworth, T.** Jan. 2013. A verification framework for interannual-to-decadal predictions experiments. *Climate Dynamics*, 40(1–2): 245–272. DOI: https://doi.org/10.1007/s00382-012-1481-2

Hankin, S, Callahan, J and **Sirott, J.** 2001. The Live Access Server and DODS: Web visualization and data fusion for distributed holdings. In: *17th Conference on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology, American Meteorological Society*. Albuquerque, NM.

Hollmann, R, Merchant, CJ, Saunders, R, Downy, C, Buchwitz, M, Cazenave, A, Chuvieco, E, Defourny, P, de Leeuw, G, Forsberg, R, Holzer-Popp, T, Paul, F, Sandven, S, Sathyendranath, S, van Roozendael, M and **Wagner, W.** Mar. 2013. The ESA Climate Change Initiative: Satellite Data Records for Essential Climate Variables. *Bulletin of the American Meteorological Society*, 94(10): 1541–1552. DOI: https://doi.org/10.1175/BAMS-D-11-00254.1

Lawrence, B, Bennett, V, Churchill, J, Juckes, M, Kershaw, P, Pascoe, S, Pepler, S, Pritchard, M and **Stephens, A.** Oct. 2013. Storing and manipulating environmental big data with JASMIN. In: *2013 IEEE International Conference on Big Data*, 68–75. DOI: https://doi.org/10.1109/BigData.2013.6691556

Lawrence, BN, Lowry, R, Miller, P, Snaith, H and **Woolf, A.** Mar. 2009. Information in environmental data grids. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1890): 1003–1014. DOI: https://doi.org/10.1098/rsta.2008.0237

Leadbetter, A. 2012. NERC Vocabulary Server version 2.0.

Mihajlovski, A, Plieger, M, de Cerff, WS and **Page, C.** 2016. Developing a Metadata Infrastructure to facilitate data driven science gateway and to provide Inspire/GEMINI compliance for CLIPC. In: *Geophysical Research Abstracts*, 18: EGU2016–4667–1. Vienna.

Miura, SH. Jul. 2016. Earth Observation data access interoperability implementation among space agencies. In: *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 3621–3623. DOI: https://doi.org/10.1109/IGARSS.2016.7729938

Nativi, S, Caron, J, Davis, E and **Domenico, B.** Nov. 2005. Design and implementation of netCDF markup language (NcML) and its GML-based extension (NcML-GML). *Computers & Geosciences*, 31(9): 1104–1118. DOI: https://doi.org/10.1016/j.cageo.2004.12.006

Nativi, S, Craglia, M and **Pearlman, J.** Jun. 2013. Earth Science Infrastructures Interoperability: The Brokering Approach. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6(3): 1118–1129. DOI: https://doi.org/10.1109/JSTARS.2013.2243113

Nebert, D, Voges, U and **Bigagli, L.** Jun. 2016. OGC Catalogue Services 3.0 – General Model. Standard 12-168r6, Open Geospatial Consortium.

Parton, GA, Donegan, S, Pascoe, S, Stephens, A, Ventouras, S and **Lawrence, BN.** Feb. 2015. MOLES3: Implementing an ISO standards driven data catalogue. *International Journal of Digital Curation*, 10(1). DOI: https://doi.org/10.2218/ijdc.v10i1.365

Pepler, S and **Callaghan, S.** Jun. 2015. Twenty Years of Data Management in the British Atmospheric Data Centre. *International Journal of Digital Curation*, 10(2): 23–32. DOI: https://doi.org/10.2218/ijdc.v10i2.379

Petrie, RE, Lawrence, BN, Juckes, M, Bennett, V, Kershaw, P, Stephens, A, Waterfall, A and **Watson, A.** 2020. Exploiting and Extending Vocabularies for Faceted Browse in Earth System Science. Patterns Manuscript in Preparation.

Raoult, B, Bergeron, C, López Alós, A, Thépaut, J-N and **Dee, D.** 2017. Climate service develops user-friendly data store. *ECMWF Newsletter*, 151.

Taylor, KE, Balaji, V, Hankin, S, Juckes, M, Lawrence, B and **Pascoe, S.** Jun. 2012. CMIP5 data reference syntax (DRS) and controlled vocabularies.

Teixeira, J, Waliser, D, Ferraro, R, Gleckler, P, Lee, T and **Potter, G.** Sep. 2014. Satellite Observations for CMIP5: The Genesis of Obs4MIPs. *Bulletin of the American Meteorological Society*, 95(9): 1329–1334. DOI: https://doi.org/10.1175/BAMS-D-12-00204.1

Williams, DN, Taylor, KE, Cinquini, L, Evans, B, Kawamiya, M, Lautenschlager, M, Lawrence, BN, Middleton, D and **ESGF contributors.** May 2011. The Earth System Grid Federation: Software Framework Supporting CMIP5 Data Analysis and Dissemination. *ClIVAR Exchanges*, 16(56(2)): 40–42.