International Journal of Geo-Information

MDPI

Article

The HD(CP)^2 Data Archive for Atmospheric Measurement Data

Erasmia Stamnas [1,*], Andrea Lammert [2], Volker Winkelmann [1] and Ulrich Lang [1]

[1] Regional Computing Centre (RRZK), University of Cologne, Weyertal 121, 50931 Köln, Germany; winkelmann@uni-koeln.de (V.W.); lang@uni-koeln.de (U.L.)

[2] Meteorological Institute, University of Hamburg, 20146 Hamburg, Germany; andrea.lammert@uni-hamburg.de

* Correspondence: estamnas@uni-koeln.de; Tel.: +49-221-47089623

Academic Editors: Constanze Curdt, Christian Willmes, Georg Bareth, Josef Strobl and Wolfgang Kainz
Received: 28 January 2016; Accepted: 8 July 2016; Published: 19 July 2016

Abstract: The archiving of scientific data is a sophisticated mission in nearly all research projects. In this paper, we introduce a new online archive of atmospheric measurement data from the "High definition clouds and precipitation for advancing climate prediction" (HD(CP)^2) research initiative. The project data archive is quality managed, easy to use, and is now open for other atmospheric research data. The archive's creation was already taken into account during the HD(CP)^2 project planning phase and the necessary resources were granted. The funding enabled the HD(CP)^2 project to build a sound archive structure, which guarantees that the collected data are accessible for all researchers in the project and beyond.

Keywords: data archive; data standard; metadata; data quality management; research data management; interdisciplinary; meteorology; atmospheric physics; clouds

1. Introduction

To develop climate change strategies, it is essential to understand how the earth's climate system works, to notice climate changes as well as their causes. The understanding progress is based on observed and simulated climate data, whereby the data from climate simulation models is one of the fastest growing segments in the data world [1]. The process of cloud formation and regional precipitation is critical for simulating atmospheric dynamics, and therefore for climate prediction. Both clouds and precipitation are important elements of the climate system [2]. Almost every physical process in the earth system, from soil-moisture feedback to sea-ice interaction to biogeochemical interaction, depends critically on clouds and the hydrologic cycle. The project "High definition clouds and precipitation for advancing climate prediction" is a German-wide research initiative to improve our understanding of cloud and precipitation processes and their implication for climate prediction. The project is funded by the German Federal Ministry of Education and Research (BMBF) and started in October 2012.

In the first project phase, the observation and simulation domain of HD(CP)^2 was concentrated on Germany and border regions. In the second phase, which has recently started, the domain has been enlarged by a tropical zone (Barbados). Whereas the model developers designed a climate model capable of high-resolution simulations, the observation component organizes ground, in situ and satellite-based observations in order to evaluate the model results. The measurements are concentrated on spatial structure and cloud particle composition, using existing measuring facilities all over Germany and the Netherlands.

Along with the increasing volume and importance of climate data, the responsibility of data producers and data publishers has also grown [3]. Besides the establishing of reliable data archives,

it is important to make them more easy to use. In HD(CP)^2 the demand for easy usability was part of the project proposal. A brief data management plan was included, describing the requirements on data storage, data documentation, data policy and standard formatting for data producers, data publishers and data users. For example, all observation component partners were requested to deliver their collected data to the archive. The willingness to share the data with the scientific community was a prerequisite for participation in the project. The mandatory data archiving policy should remove barriers to effective data sharing. In general, data sharing allows other scientists to verify the data, and use it for reproducing model results, validating interpretations, and building upon the work of the project research [4].

A data archive is more than a collection of datasets; besides the data, the archive should include documentation, a physical store of the information and meta-information, and a user interface [5]. Within the project, the implementation of such a data management system (DMS) is mainly performed at the Regional Computing Centre of the University of Cologne (RRZK). The established DMS takes into consideration the practical recommendations given in the "Handbuch Forschungsdatenmanagement" which was designed to help face the general challenges in data management [6]. The RRZK is one project partner in HD(CP)^2 responsible for technical infrastructure and support. A computing center is specialized to provide data management services and support storage infrastructure in order to ensure a high level of quality and a sustainable basis for further development in the post-project phase. Most scientists have neither the resources, with regard to time and cost, nor the technical support to maintain an archive. In fact, the lack of technical support and structural funding are the most often-named threats to common digital data preservation [7].

An important reason for the establishment of a new data archive was the lack of a common infrastructure for the detailed supersite observations in Germany. There is already a wide spectrum of archives for atmospheric data in central Europe. Some archives focus on specific instruments or products, such as ACTRIS [8], which is concentrated mainly on LIDAR (light detection and ranging) data, WDC-RSAT [9], which specializes in satellite data, or Cloudnet [10], which offers data products resulting out of instrument combination for cloud specification. Other archives are more general such as PANGAEA [11], which is for georeferenced data or CERA [12], which is mainly for climate model output data; both are well established. These data archives were designed for specific instruments/instrument combinations or model output data. The wide range of instruments in HD(CP)^2 is not sufficiently covered in any of these existing archives. Furthermore, an intensive measurement campaign, planned for HD(CP)^2, with additional and partly new developed instruments and products, has indicated a need for a new data archive. Finally, the new archive contains data standardized to a high extent. The data standard used in the project was designed especially for the archive and supports an easy and fast data exchange between project scientists.

In the following we will describe the genesis of the archive in more detail. Section 2 shows an overview of the different kinds of observation data in the HD(CP)^2 archive. In Section 3, a short introduction into the binding conventions for each dataset is given. Section 4 describes the technical infrastructure the archive is built on, followed by the description of value-added services in Section 5 and some future perspectives in Section 6.

2. Standardized Data for the HD(CP)^2 Archive

In the beginning there were lively discussions within the HD(CP)^2 observation community about which data should be reasonably integrated, how to create sensible products out of it, and how to make the data reusable for other scientists, as simply and easily as possible, without technical barriers.

Further discussions took place about the term 'easy-to-use'. What does an 'easy-to-use' data archive mean, not only for users from the observation community who intend to share their data, but also for project partners without experience in special instrument types or processing observation data? Every data producer has her/his own preferred data format. Starting from these needs, the archive would have to serve many different data formats. A user might want to retrieve a specific

variable, ideally an output variable of a specific model with a specific resolution. In this case the archive would have to serve an endless number of variedly parameterized variables. Neither of these kinds of archiving observation data are either practicable or desirable. Therefore, we had to find a balance between different positions, considering the wide range of instruments, variables, and levels, as well as the temporal and spatial scales of the measurements.

The discussion results were summarized in the HD(CP)^2 observation data product standard (HOPS) [13]. This document now serves as a guide for all data producers, describing in detail the binding conventions for the datasets and their associated metadata (see Section 3).

Currently the HD(CP)^2 data archive provides standardized data of typical atmospheric observations and specific cloud parameters, measured and derived at different observation instruments. The spatial resolution of the data varies from regular and irregular geographical grids, such as satellite or ceilometer network data, to local observations (supersites) and four-dimensional data of scanning instruments such as the cloud radar.

The data cover long-term measurement cycles as well as intensive short-term observation campaigns (see Figure 1). The long-term observations are based on full-domain observations and local observations. One will find datasets of different satellite-based instruments (such as SEVIRI on the MeteoSat second-generation (MSG) satellites), as well as datasets of ground-based instrument networks (such as the C-band rain radar network of the German Weather Service, RADOLAN), as well as datasets from the so-called supersites.

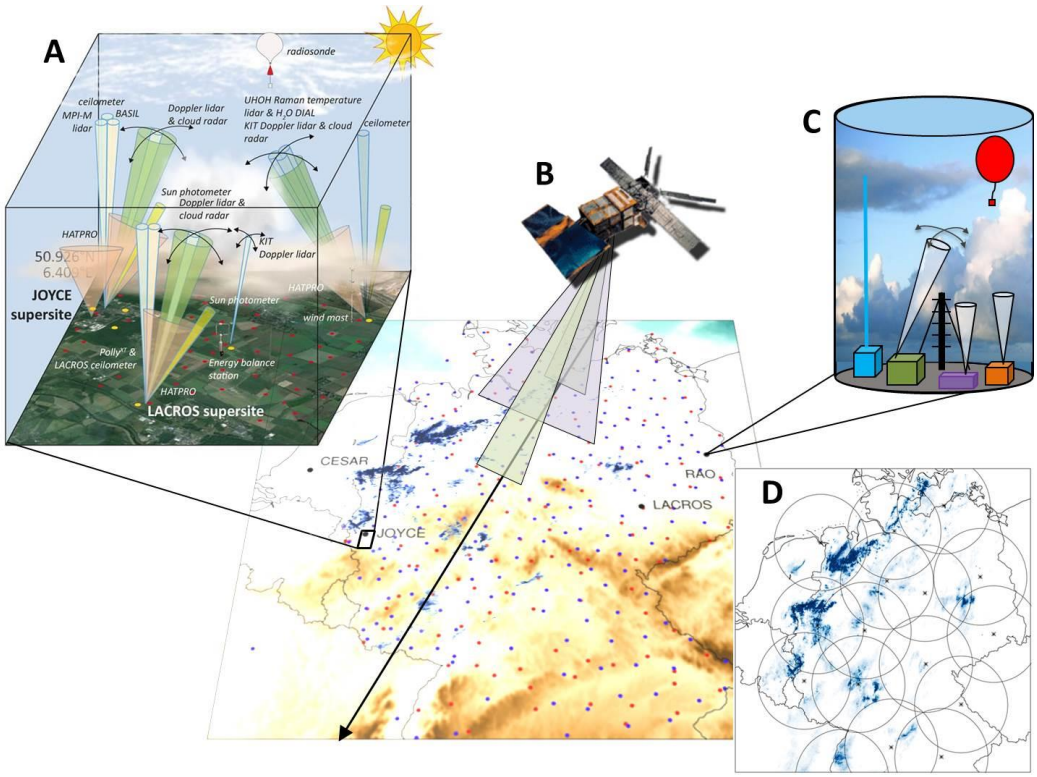

Figure 1. Overview of observations in HD(CP)^2. The underlying map of Germany shows the locations of the supersites and the GNSS and Ceilometer network stations. Example of short-term observations: Instrumentation during the HOPE campaign in Juelich (**A**); Of long-term observations: satellite observations (**B**); schematic instrumentation of Supersites (**C**); precipitation from C-Band rain radar network of the German Weather Service (**D**).

Currently there is no common definition of the term "meteorological supersite"; the meaning depends on the project and the instrument focus. For the HD(CP)^2 data archive, a supersite means an

observatory for long-term measurements including at least a cloud radar, a microwave radiometer and a LIDAR system. Each additional instrument, such as a meteorological tower, is a bonus.

At the moment, datasets of four supersites in central Europe are collected and processed consistently: Cabauw Experimental Site for Atmospheric Research (CESAR), Juelich ObservatorY for Cloud Evolution (JOYCE), Leipzig Aerosol and Cloud Remote Observations System (LACROS), and Richard-Aßmann Observatory in Lindenberg (RAO).

The data archive also provides short-term observations such as the data derived out of the HD(CP)^2 Observational Prototype Experiment (HOPE). The HOPE campaign took place in April and May 2013 in the region around the Juelich Research Center, which includes the supersite JOYCE. It was designed to provide a critical model evaluation at the scale of the model simulation (in the order of 100 m), and further to provide information on sub-grid variability and microphysical properties.

The variety of observation products in the archive is the basis for comparing instruments, as well as for validating climate models. As an example, the results of HOPE were used for the comparison of the measured integrated water vapor, derived from MODIS (moderate-resolution imaging spectroradiometer) and ground-based instruments such as GPS (global positioning system), MWR (microwave radiometer) and the sun photometer, and the validation of the models COSMO-DE and ICON-DE (numerical weather prediction models of the German Weather Service) [14].

3. The HD(CP)^2 Observation Data Product Standard

As a final outcome of the discussions, the HD(CP)^2 observation data product standard (HOPS) was developed for the archive, including a metadata scheme. Data files are required to be created in the NetCDF format, which is a self-describing, machine-independent data format that supports the specification of array-oriented scientific data. Programming libraries allow creation, access, and modification of NetCDF-formatted data. NetCDF [15] is a well-known format and is often used in the climate community, which has adopted it as a primary standard. For the data contained in the NetCDF files, the principles given in the Climate and Forecast (CF) metadata conventions have to be followed as far as possible [16]. In particular the HD(CP)^2 definition of the coordinate and data arrays is based on these conventions. The next-to-last NetCDF version 3 and the current version 4 (also with the compression option) are supported.

In consideration of the "easy-to-use" approach, data are stored on a daily file basis, with only a few exceptions, e.g., for airborne measurements or data of satellite overpasses, due to the large size of the respective daily data. An HD(CP)^2 dataset refers to a whole set of (daily) data files from one measurement series, sharing the same metadata. One set might contain many thousands of files, depending on the time period of the respective (continuing) measurement. The datasets are version-controlled by assigning a unique version number to different stages. Each change in the data, e.g., a fixed bug in the processing or a new position of the instrument, implies a new version number.

In general, each daily data file should contain only one variable and its estimated error, if available. There are exceptions for level 1 data, which may also contain more than one variable per file, and measurements for which scalar or single variables do not make sense, e.g., for the variable group wind (consisting of the three components plus horizontal wind speed and direction).

Each variable is described by a variable name, and in the case of a CF variable, by a standard_name attribute, and its unit. However, since the CF metadata conventions were developed for model data, not every kind of observation data is covered. Therefore, we cannot provide a CF standard_name attribute for all variables obtained by HD(CP)^2-related instruments. Variables, which do not have a CF standard name yet have been provided with a long_name attribute. The variables' units comply with the International System of Units (SI Units).

An important variable for all datasets is the time variable of the measurement. The convention, therefore, is to specify time in seconds since 1 January 1970 00:00:00, as commonly used in UNIX-like operating systems. Time must be related to the Coordinated Universal Time (UTC, Universal Time, Coordinated).

Particularly, a dataset file must have a unique name, which reflects the categorization system, consisting of seven parts as follows:

< kkk > _ < sss > _ < instnn > _ < ln > _ < var > _ < vnn > _YYYYMMDDhhmmss.nc.

The description of the respective parts is given in Table 1.

Table 1. Composition of the HD(CP)^2 data file names and description of each file name part.

Part	Description
<kkk>	kind of measurement type (supersites, full domain observation, campaign, etc.)
<sss>	abbreviation of supersite or owner institute of the instrument or distributor of data
<instnn>	instrument or synergy product, retrieval algorithm, + numbering (starting with 00)
<ln>	HD(CP)^2 level of data post-processing (starting with 1)
<var>	observation variable name
<vnn>	version of dataset (starting with 00)
YYYYMMDD	date of measurement (UTC): year (YYYY), month (MM), day (DD), and starting time of
hhmmss	measurement (UTC): hour (hh), minutes (mm), seconds (ss)

The construction of the file name is described in detail in the HOPS document.

Using standardized datasets means that no adaptation to an individual instrument or a publisher-specific variable name is necessary. As an example, the air temperature has the variable name ta, with the standard name 'air_temperature' and the unit *K*, as given in the CF conventions. According to the HD(CP)^2 taxonomy, the air temperature (ta) in a level 2 product (l2) of the first microwave radiometer (mwr00) from the supersite (sups) JOYCE (joy), first version (v00) from 1 January 2014, should be named as sups_joy_mwr00_l2_ta_v00_20140101000000.nc.

Well-documented data should enable other researchers to understand, use and reuse the data correctly. Good quality metadata are an asset [17]. For this reason, every dataset which is published by an HD(CP)^2 data server must be associated with metadata in an appropriate form (see Section 5.2). The benefit of semantically annotating data according to some well-established vocabulary is obvious. The controlled vocabulary will aid in searching and finding data, making it more 'shareable' with other researchers. Most common standards such as some ISO standards might seem to be ideal for our needs, but unfortunately for atmospheric measurement data there is no standard that fits well. The standard ISO 191xx family for geographical information systems (mainly ISO 19115-1:2014 [18]) with its sophisticated structure and large number of entities is far too complex for the small set of elements that suits the project needs and is almost a contradiction to our 'easy-to-use' approach. The same can be said of the INSPIRE Directive (2007) [19]. On the other hand, the Dublin Core (NISOZ3985) [20] and the DataCite Metadata Scheme (3.1) [21] are too general. Although the latter provides metadata fields to specify the geographical region where the data were gathered (i.e., geoLocation), we needed to define the exact position and orientation, especially the height, of every single measuring instrument, or combinations of instruments. That is why we decided to create our own metadata scheme.

We designed about 30 metadata descriptive elements, most of them with controlled vocabulary, adapting some well-established attributes in the climate community from the NetCDF header. The data type–specific elements reflecting the needs of our climate project and the global attributes are compliant with current metadata standards such as the Dublin Core, the DataCite Metadata Scheme, the ISO19115-1 Standard and the INSPIRE Directive. The HD(CP)^2 Metadata Scheme provides mandatory, optional and automatically generated attributes. For instance, HD(CP)^2 metadata must include at least the following global attributes: title of the data set, institution and name of the data producer, the name and e-mail address of a contact person, the instrument source, the information about the conventions used, the processing date and the HD(CP)^2 license policy. The fields for comments and the history of the data are optional.

The HOPS document specifies the metadata standard to describe and index the archived data in a consistent way. The metadata files should be encoded in the eXtensible Markup Language [22]. XML provides a common way of describing particular types of a document structure, which is why it was chosen as the metadata specification language. Data producers have to generate a separated file with detailed information about measurement and instruments, guided by our established vocabulary for metadata. For example, they have to specify keywords by the HD(CP)^2 taxonomy for later search and accurate retrieval results. Additional metadata, such as the description of the instrument location, references, data history and constraints to methods and/or data, should also be characterized.

HOPS is constantly progressing, mainly because it is based on dynamic datasets. A dynamic dataset is continuously growing, fed by ongoing observations, whose conditions and methods will change now and then. For example, a new data processing, a significant new instrument software or a new instrument combination means that the standard has to be updated to keep it fit for the purpose.

Consequently, the HD(CP)^2 observation data product standard could be adapted to further supersites, measurement campaigns, and instruments/variables and, beyond HD(CP)^2, could also be applied to other atmospheric research data to which the archive is open now.

4. Infrastructure and Data Management

The HD(CP)^2 data archive provides observation data of very different kinds and makes it accessible on a long-term basis. Therefore, cost-efficient data management is a challenge: on the one hand, the data should be centralized in order to ensure the data's consistency and easy accessibility for users; on the other hand, all infrastructural resources should be utilized, no matter where they may be available. The latter is a very important aspect due to the fact that the amount of data is steadily growing. Even though the data repository was launched in autumn 2013, it is filling remarkably fast. That is why the archive must be scalable and grow flexibly, respecting the requirements concerning availability and performance. At present—June 2016—the archive already contains more than 150 datasets, starting from the year 2007 on. The datasets include about 55,000 data files, using around 1 TB storage space.

Our scalable technical solution consists of several distributed data servers at different project partner sites. All servers are running a common infrastructure with one standardized design but miscellaneous operating services. The services are interconnected to form a single virtual archive with a common web portal as a central entry point for all users. Each data server is hosted by a university computing center or a university institute with a computing center in the background, guaranteeing long-term availability of the data. For each data server all files are stored in a standard hierarchical file system, any extensible file system will do, as long as the internal data file organization is based on the file name taxonomy specified in HOPS according to measuring instruments, variables, data processing state and others (see Section 3).

Due to the different local computing and data storage resources, the physical implementation differs between the single sites. For example, in Cologne the data services as well as the data storage are implemented redundantly using the virtual high availability infrastructure of the regional computing center (RRZK) based on VMWare ESX. The virtual machine (VM) runs on top of a High Availability Cluster (HA Cluster) and is expandable regarding processors, memory and storage. This does not only allow a flexible administration of the HD(CP)^2 data service, but also guarantees minimal timeout periods in case of hardware maintenance or hardware failures of the underlying virtualized infrastructure. In addition, all services on the data server are supervised by a central NAGIOS server notifying the data archive administrators if any of the services do not work well or are down. Concerning the safety of HD(CP)^2 data, every night all data are incrementally backed up on a central tape robot system on at least two different tapes at two different locations, thus providing four copies. The University of Leipzig uses a different concept for redundancy. For the data services, two identical virtual servers have been set up at different places at the university and are synchronized via a heartbeat. The Leipzig part of HD(CP)^2 data are stored on a central data server and replicated via

GlusterFS to different virtual RAID disks, which are mounted by the virtual data servers mentioned above. Therefore, a high availability of the data services is guaranteed, too.

The backbone of the HD(CP)^2 technical platform is the THREDDS Data Server (TDS), an open-source product developed by Unidata [23]. The function of the Thematic Real-time Environmental Distributed Data Services (THREDDS) is to give researchers access to a large collection of diversified and archived datasets at a number of distributed server sites. It provides a common interface for geoscientific data formats such as HDF5, GRIB, and NetCDF, and serves them through OPeNDAP, Web Coverage Service (WCS), NetCDF subset, and HTTP file transfer services. The first three server protocols allow a user to obtain subsets of the data, which is efficient for direct interaction, e.g., for data visualization. The TDS contains metadata in publishable inventories and catalogs. Based on XML, these inventories and catalogs can be created individually by data publishers. The inventories and catalogs can be harvested and indexed into digital libraries worldwide. THREDDS catalogs list all data resources located at a specific server. Remote catalogs from other servers can be referenced via URLs. Users searching for data browse through linked catalogs of distributed servers to explore metadata, not noticing where the data are originally located [24].

Up to now, HD(CP)^2 has established a network of three distributed THREDDS data servers (Cologne, Berlin, and Leipzig) which might be extended by more servers in the future. In the second phase of the project, its one goal to provide a preconfigured virtual machine to be delivered to other potential data publishers. Preconfigured means that an operating system, the THREDDS data server and a data harvester are already installed. This is a further step towards sustainability, and it supports the "easy-to-use" approach for ambitious scientists as well, who may want to set up their own data server and connect it with the archive.

5. Value-Added Services

One of our objectives is to simplify and automate the everyday actions and processes in data handling. Maintaining the data requires many resources: the data have to be quality checked, described in detail, organized in folders, indexed and published, access controlled, and found easily. The technical direction was to build the data archive using existing open-source components, and to supplement these with self-generated tools to fit the archive's demands if necessary.

5.1. Processing and Integrating Data

Unfortunately, the THREDDS data server does not include software tools for automated data processing and integration. Therefore, we had to develop our own software, which is now able to automate most steps from data delivery to data integration to data analysis. The tools created provide several applications and modules for miscellaneous tasks such as data delivery, compliance checking plus evaluation of data and associated metadata, data integration and supervising the metadata database. In particular, an especially designed online editor allows the data producers to generate the mandatory metadata in a convenient way (see Section 5.2).

The data publishing workflow (shown in Figure 2) is described as follows: a registered data producer has to upload a sample dataset for evaluation first. A HD(CP)^2 sample set contains a data file in NetCDF format version 3 or 4, its checksum file, and the XML encoded metadata file corresponding to the HD(CP)^2 scheme. If the evaluation (see Section 5.3) is successful, the data publisher is allowed to upload the entire dataset and its metadata. If the data evaluation fails, the data publisher receives correction suggestions. When the corrections are accepted, the sample set can be uploaded for evaluation again. Of course, a data publisher can make suggestions too, e.g., the introduction of a new variable. In case the dataset represents an ongoing measurement and there is a major change, such as a new position of the instrument, the evaluation procedure has to be repeated again and the dataset's version number is increased by one. Once the data files are uploaded, a software tool executes a brief test, e.g., whether the dataset has been approved, and if the test is positive the data files are integrated into the file system. The TDS catalog generation has to be done manually, but

will be automated in the next project phase. The newly added metadata files are harvested and the data path is linked to the web portal (see Section 5.4).

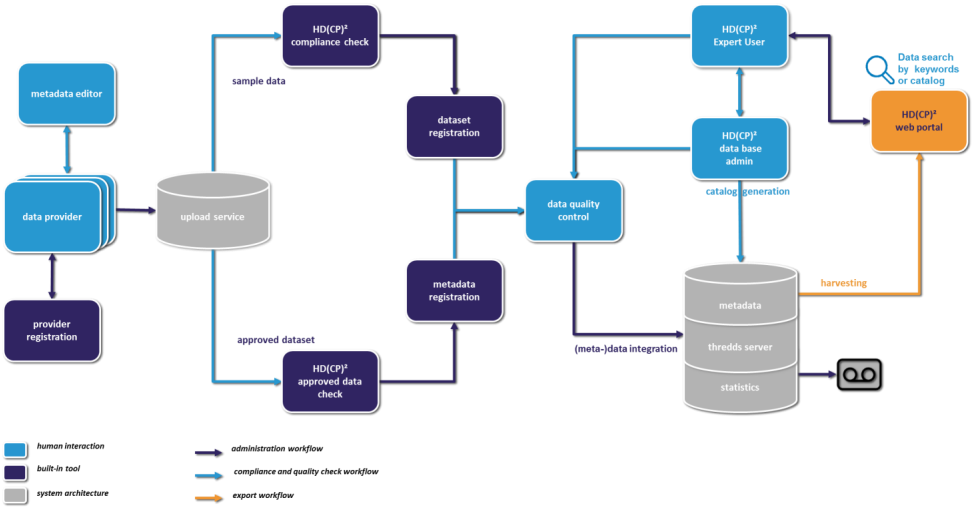

Figure 2. The HD(CP)^2 data publishing workflow.

In addition, we have developed a basic statistics package to answer some basic questions, such as how many files does the archive contain and how big is the monthly increase in data? Some technical details such as the download speed, the download count and the geographical location of the user client are also collected and stored into a relational database, which can be accessed at any time using a specifically created administrator interface. This particular data help a data archive administrator to measure the archive's performance and growth, optimizing calculation of future data volume and resources needed.

5.2. The Metadata Editor

To guarantee compliance with the defined standard, the metadata files have to be created via the HD(CP)^2 Metadata Editor (see Figure 3). It enables researchers, who are unfamiliar with XML or the HD(CP)^2 metadata scheme, to create their own metadata files by answering several questions in a specifically designed web form. The form mostly presents lists of selectable items, displayed within drop-down menus. Some fields allow free text in order to cover the full range of possibilities. As an example, the HD(CP)^2-approved variables can be picked from a list of more than 100 items. All of them were proposed and accepted by the observation community. In contrast, the product description field allows free text, helping to add information which is not covered by the specified vocabulary and might be important. At the end of the questionnaire the underlying application will transform the filled-in information into a valid XML document and provide it with a digital signature (hash key). Only signed metadata files are accepted, ensuring they have been edited by the metadata editor, for the data integration process.

The HD(CP)^2 Metadata Editor is a web server-supported application and does not require any plugins to view and/or edit metadata files. It can also be used for addition, modification, and removal of metadata file elements.

5.3. Data Quality Management

We define the data quality as the degree to which the data fulfills our requirements [25]. The data are carefully analyzed to make sure that they meet our standard in order to archive high quality data. High quality in this context means that the data are reliable and suitable for climate simulations and can be used without investing unreasonable time or cost, which, in essence, means that the data are easily accessible and understandable.

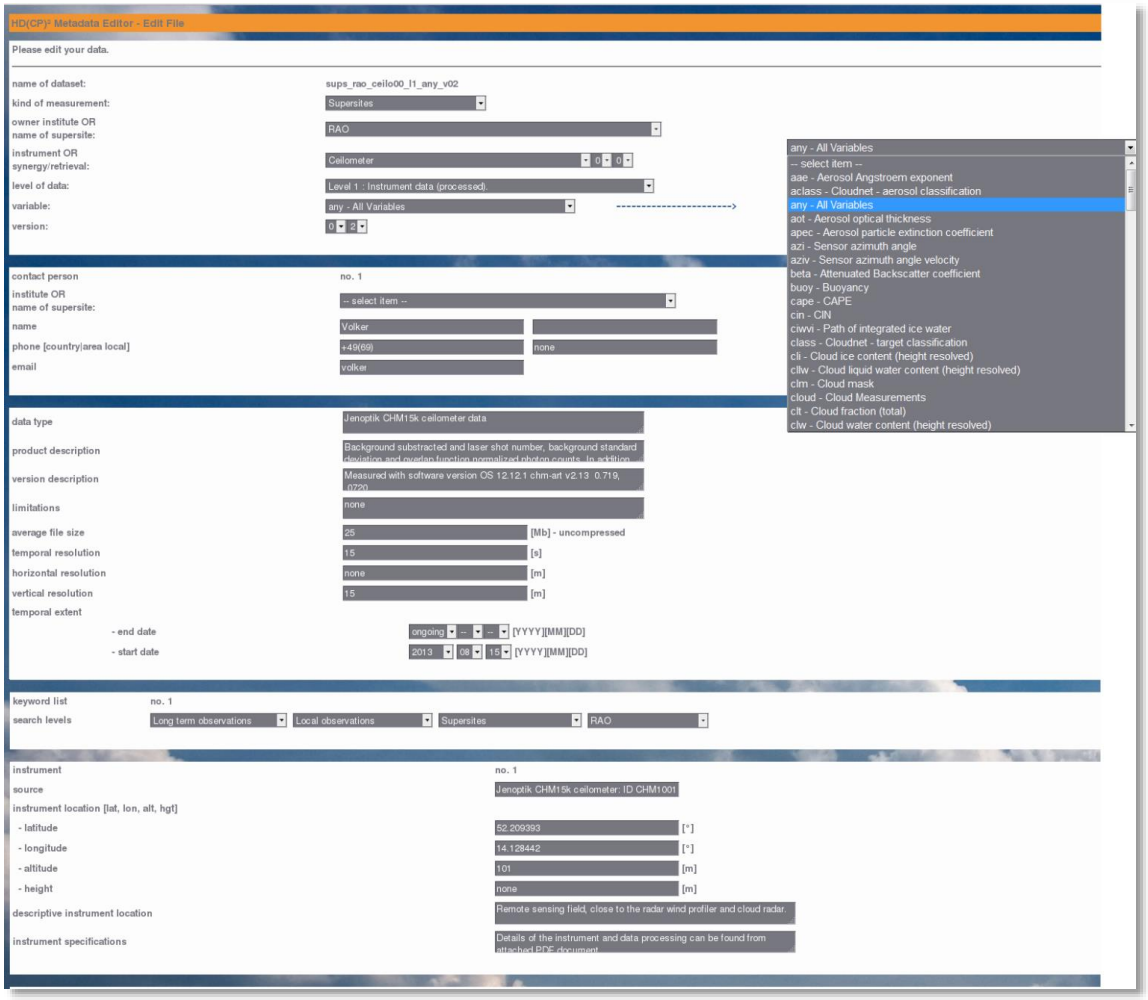

Figure 3. The metadata editor.

The reliability of data is related to their sources, the acquisition methods and the evaluation and storage procedures (see Section 4). With regard to the data acquisition, researchers who want to be accepted as data producers can register at the HD(CP)^2 data archive administration. After verification of their identity, they are required to upload a sample dataset which will be evaluated (see Section 5.1).

For the evaluation of the observation data and their associated metadata, a software-based control system was established. The software ensures strict compliance with the HD(CP)^2 standard and cross-checks the content of the metadata file against the global information stored in the NetCDF file header. Due to the large amount of different instruments and variables, human interaction cannot be omitted yet. We have created a role called the Observation Expert User (OEU), who is responsible for the final acceptance. The OEU checks the data and metadata in the NetCDF file and the editor-generated XML formatted metadata files. Even a formally correct dataset could include incorrect units, false standard names or corrupt data. For example, the decisions of whether a measurement makes sense for a specific instrument (there might be unreliable structures resulting from programming errors) or even whether the units in the dataset itself are chosen properly (there might be incorrect units like Pa/hPa for air pressure, or K/°C for temperature) still need a human factor. Therefore, a visual check of each sample set is necessary; the NetCDF file is examined with a generally agreed-upon software such as "Ncview" and "Panoply" to have a look into the time series itself. This guarantees that the data are of high quality and meet our standards.

The procedure still involves coordination and harmonization, especially when a new variable is introduced, but it is very helpful in supporting researchers to prepare their data for reuse.

Once a sample set is accepted, the whole dataset can be uploaded. The approved dataset will be transferred to the TDS, cataloged, and can be served almost immediately to a user.

5.4. The HD(CP)^2 Portal

The HD(CP)^2 web portal, located at the Integrated Climate Data Center (ICDC) of the University of Hamburg [26], was created as the central entry point to the data archive. The main task of the portal is the clearly arranged presentation of all available datasets, the respective metadata, and the respective links to the archive. For each dataset, the web portal provides a standardized data sheet, shown in Figure 4, automatically generated from its metadata. Included is information about the measuring instrument, the main measurement variables, and some global information such as the start and end of measurements and the version number. The data sheet also contains a link to the dataset and the associated metadata. The metadata files are periodically harvested from the distributed data servers using the OAI metadata harvesting protocol [27]. The user download of a dataset itself is performed by the distributed data servers, whose infrastructure is based on virtual hardware allowing an easy extension of hardware resources, e.g., the number of processors, memory or disk and backup storage (see Section 4). The extendible architecture may also face an increasing number of users and a growing volume of datasets in the future. All users have free online access to the metadata, but during the first project phase they have to authenticate themselves as project members to download the data themselves. With the start of the second phase, all data shall be freely accessible for the whole scientific climate community.

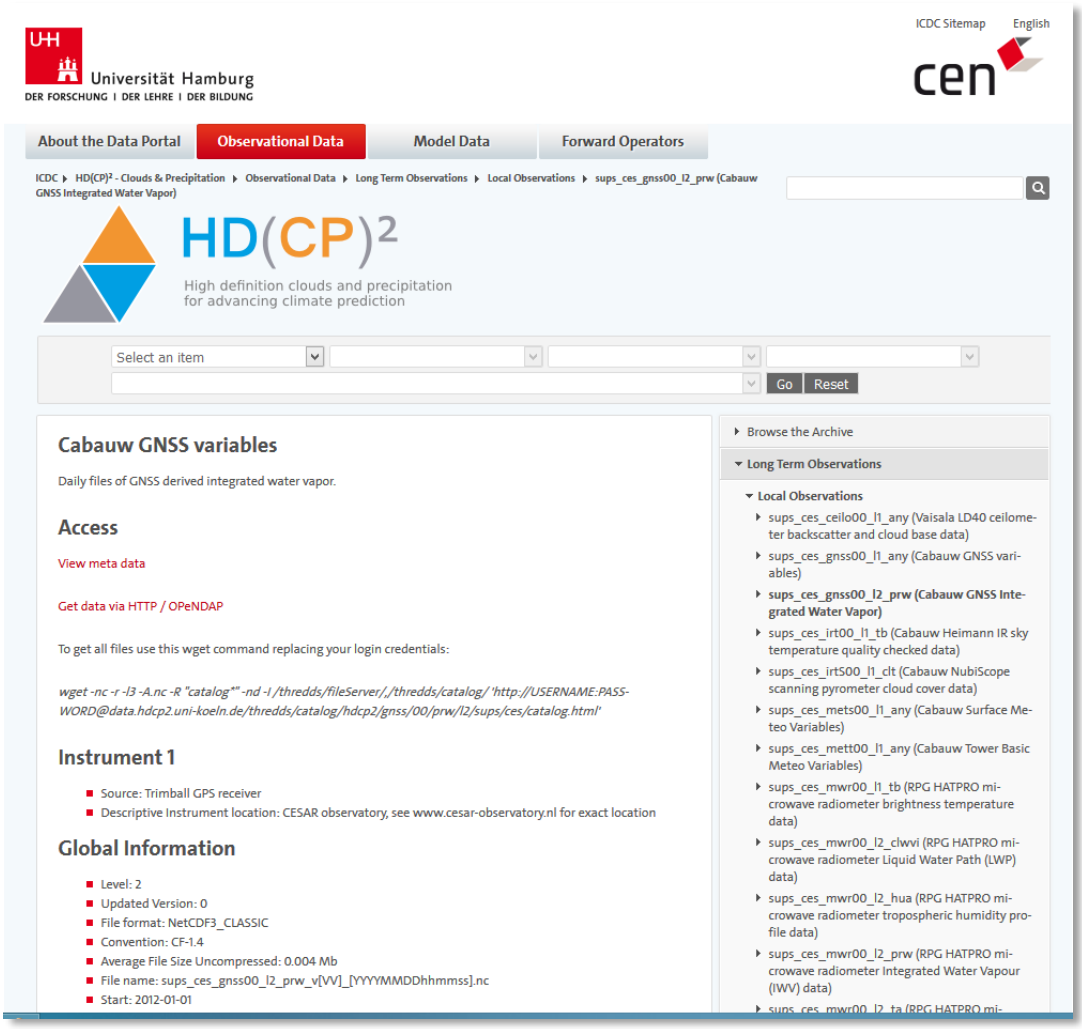

Figure 4. Example of a data sheet provided by the HD(CP)^2 web portal. The drop-down selection menu, based on the search tree (see Figure 5), appears directly below the project logo.

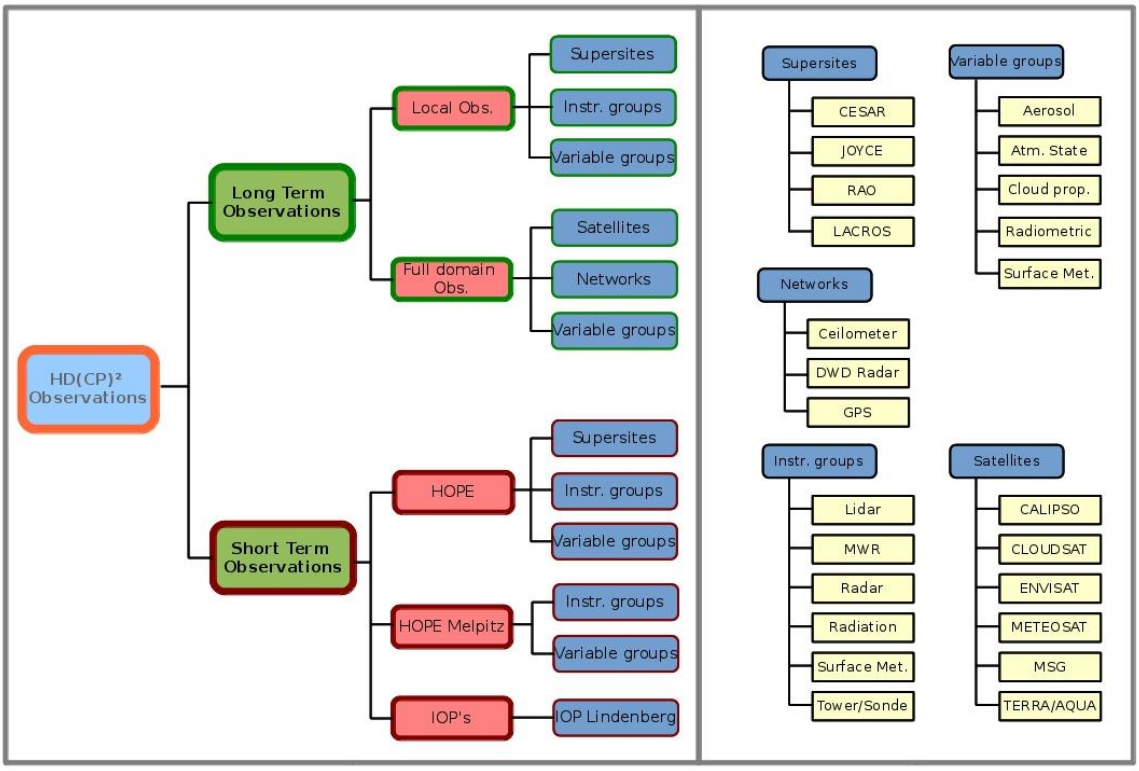

Figure 5. Search tree, following the metadata taxonomy for search keywords.

Beyond that, the portal offers general information about the HD(CP)^2 project, descriptions of the different measurement classifications, the measurement campaigns, and the HOPS document. Furthermore, the portal presents an overview of all forward operators used in HD(CP)^2. Forward operators translate the state of the atmosphere within the model into virtual measurements, which can be compared directly to real measurements. The inventory of all forward operators allows a comparative assessment among themselves, due to the standardized metadata for each operator.

Following the "easy-to-use" approach, the web portal offers the possibility to search for specific datasets by using the selection menu based on the HD(CP)^2 search taxonomy. The search tree, shown in Figure 5, is designed to facilitate the user's search for a specific dataset. The datasets are classified according to the measurement type in long- or short-term observations. The long-term observations are divided into local observations, equivalent to the measurement at supersites, and full domain observations, which include satellite and network data. The short-term observations are sorted according to the different campaigns, e.g., the HOPE campaign or Intensive Observation Periods (IOPs). At the next level, datasets are divided into specific groups, such as instrument groups, variable groups or satellites.

A user without any experience in meteorological observations might look for a specific variable, while an observation expert might look for an instrument group at a certain location (supersite).

For this reason, the data producers have the opportunity to create more than one keyword list for one dataset, for example ceilometer datasets from the supersite JOYCE should be classified at the first level as Long-Term Observations and at the second level as Local Observations. At the third level, there are more possibilities: the dataset could be associated with Supersites → JOYCE, Instrument groups → Lidar, and Variable groups → Aerosol. Consequently, one dataset is sorted into three different categories, which increases the probability of being found and used.

6. Future Perspectives

The collaboration between data publishers in a digitally networked world is of increasing importance, particularly in terms of sustainability for measurement data, because measurement data are not reproducible [28]. In the future we will cooperate with CERA [12], located at the German Climate Computing Center (DKRZ; Deutsches KlimaRechenZentrum), which is a partner in HD(CP)^2. In order to integrate our data into the CERA database, we will define finalized datasets and aggregate them in new data groups (e.g., one dataset for all measurements and products of a special supersite). Finalized datasets might be campaign data or long-term measurements for closed years. The concept of a parallel storage system is a solution to guarantee sustainability. The first step is a mapping of the metadata terms used in HD(CP)^2 and CERA. The integration process includes the allocation of a Digital Object Identifier (DOI) for referencing purposes. For the more 'experimental data' in the HD(CP)^2 data archive, we will establish a persistent identifier (PID).

In autumn 2016, the data archive will open up to the climate community for data sharing, and also to other digital libraries and archives for metadata harvesting to support open availability. Currently we are working on a mapping relation for EUDAT [29]. Wherever the HD(CP)^2 data are stored, they will be freely available for the research community and thus for non-commercial use.

Finally, though the CF conventions are mainly defined for model output, we will contribute to the CF conventions in the near future due to the large amount of often used but not defined measurement variables.

Acknowledgments: This work was funded by the Federal Ministry of Education and Research in Germany (BMBF) under the research program "HD(CP)^2: High Definition Clouds and Precipitation for Climate Prediction". We want to thank the ICDC (especially Annika Jahnke-Bornemann and Remon Sadikni) for the friendly and helpful assistance. Many thanks to Verena Grützun for her profound contribution to the observation data product standard. Furthermore, we thank all the data producers for their cooperation. Finally we thank the RRZK for providing the necessary hardware resources so far.

Author Contributions: The functional concept of the HD(CP)^2 data archive was evolved by Erasmia Stamnas and Volker Winkelmann. Erasmia Stamnas developed the archive applications and also the client applications; Andrea Lammert developed the HD(CP)^2 observation data product standard and is responsible for the data quality management; Volker Winkelmann and Ulrich Lang are responsible for the technical infrastructure; Erasmia Stamnas and Andrea Lammert wrote this manuscript together with Volker Winkelmann.

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

References

1. Overpeck, J.T.; Meehl, G.A.; Bony, S.; Easterling, D.R. Climate data challenges in the 21st century. *Science* **2011**, *331*, 700–702. [CrossRef] [PubMed]
2. Bony, S.; Stevens, B.; Frierson, D.M.W.; Jakob, C.; Kageyama, M.; Pincus, R.; Shepherd, T.G.; Sherwood, S.C.; Siebesma, A.P.; Sobel, A.H.; et al. Clouds, circulation and climate sensitivity. *Nat. Geosci.* **2015**, *8*. [CrossRef]
3. Procter, R.; Halfpenny, P.; Voss, A. Research data management: opportunities and challenges for HEIs. In *Managing Research Data*; Graham, P., Ed.; Facet Publishing: London, UK, 2012; pp. 135–150.
4. Tenopir, C.; Allard, S.; Douglass, K.; Aydinoglu, A.U.; Wu, L.; Read, E.; Manoff, M.; Frame, M. Data sharing by scientists: Practices and perceptions. *PLoS ONE* **2011**, *6*, e21101. [CrossRef] [PubMed]
5. Mückschel, C.; Nieschulze, J.; Weist, C.; Sloboda, B.; Köhler, W. *Challenges, Problems and Solutions in Data Management of Collaborative Research Centers*; eZAI (elektronische Zeitschrift für Agrarinformatik): Freising, Germany, 2007.
6. Büttner, S.; Hobohm, H.-S.; Müller, L. *Handbuch Forschungsdatenmanagement*; Bock + Herchen: Bad Honnef, Germany, 2011.
7. Kuiper, T.; van der Hoeven, J. Insight into Digital Preservation of Research Output in Europe; PARSE Insight Survey Report, 2009. Available online: http://www.parse-insight.eu/downloads/PARSE-Insight_D3--4_SurveyReport_final_hq.pdf (accessed on 8 December 2015).
8. Aerosol, Clouds, and Trace gases Research Infrastructure (ACTRIS). Available online: http://www.actris.eu (accessed on 20 May 2016).

9. World Data Center for Remote Sensing of the Atmosphere (WDC-RSAT). Available online: http://wdc.dlr.de (accessed on 20 May 2016).
10. Cloudnet. Available online: http://www.cloud-net.org (accessed on 20 May 2016).
11. Data Publisher for Earth & Environmental Science: PANGAEA. Available online: http://www.pangaea.de (accessed on 20 May 2016).
12. Climate and Environmental Retrieval and Archive (CERA). Available online: http://cera-www.dkrz.de/WDCC/ui (accessed on 8 December 2015).
13. HD(CP)^2 Observation Data Product Standard (HOPS). Available online: http://www.hdcp2.eu/Community-Data-Format.2810.0.html (accessed on 8 December 2015).
14. Steinke, S.; Eikenberg, S.; Löhnert, U.; Dick, G.; Klocke, D.; Di Girolamo, P.; Crewell, S. Assessment of small-scale integrated water vapor variability during HOPE. *Atmos. Chem. Phys.* **2015**, *15*, 2675–2692. [CrossRef]
15. Network Common Data Form (NetCDF). Available online: http://www.unidata.ucar.edu/software/netcdf/ (accessed on 8 December 2015).
16. CF Conventions and Metadata. Available online: http://cfconventions.org/ (accessed on 8 December 2015).
17. Greenberg, J.; Swauger, S.; Feinstein, E.M. Metadata Capital in a Data Repository. In Proceedings of the International Conference on Dublin Core and Metadata Applications 2013, Lisbon, Portugal, 2–6 September 2013; pp. 140–150.
18. ISO/TC 19115-1:2014. Geographic Information—Metadata—Part 1: Fundamentals; International Organization for Standardization: Geneva, Switzerland. Available online: https://www.iso.org/obp/ui/#iso:std:iso:19115:-1:ed-1:v1:en (accessed on 29 April 2015).
19. INSPIRE Directive, May 2007. Available online: http://inspire.ec.europa.eu/ (accessed on 29 April 2015).
20. Dublin CORE Metadata Initiative. Available online: http://dublincore.org/metadata-basics/ (accessed on 29 April 2015).
21. DataCite Metadata Schema Version 3.1, June 2015. Available online: https://schema.datacite.org/meta/kernel-3/doc/DataCite-MetadataKernel_v3.1.pdf (accessed on 29 April 2015).
22. W3C-Extensible Markup Language (XML). Available online: http://www.w3.org/XML/ (accessed on 8 December 2015).
23. THREDDS Data Server. Available online: http://www.unidata.ucar.edu/software/thredds/current/tds/TDS.html (accessed on 8 December 2015).
24. Domenico, B.; Caron, J.; Davis, E.; Kambic, R.; Nativi, S. Thematic Real-Time Environmental Distributed Data Services (THREDDS): Incorporating interactive analysis tools into NSDL. *J. Digital Inform.* **2006**, *2*, 4. Available online: https://journals.tdl.org/jodi/index.php/jodi/article/view/51/54 (accessed on 13 July 2016).
25. Fürber, C. *Data Quality Management with Semantic Technologies*; Springer Gabler: Wiesbaden, Germany, 2016.
26. The Integrated Climate Data Center (ICDC). Available online: http://icdc.zmaw.de/1/projekte/hdcp2.html (accessed on 30 March 2016).
27. The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). Available online: http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm (accessed on 30 March 2016).
28. Fritsch, B. Klimaforschung. In *Langzeitarchivierung von Forschungsdaten*; Neuroth, H., Strathmann, S., Oßwald, A., Scheffel, R., Klump, J., Ludwig, J., Eds.; Verlag Werner Hülsbusch: Boizenburg, Germany, 2012; pp. 195–212.
29. EUDAT B2find. Available online: http://b2find.eudat.eu (accessed on 27 January 2016).