Integrated Climate Data Center (ICDC) at the Cluster of Excellence at the University of Hamburg

M. Stockhause[1], H. Höck[2]
[1] University of Hamburg: martina.stockhause@zmaw.de, [2] Max Planck Institute for Meteorology, WDC Climate: heinke.hoeck@zmaw.de

Abstract
At the KlimaCampus (http://www.klimacampus.de), Cluster of Excellence at the University of Hamburg, an Integrated Climate Data Center (ICDC: http://www.icdc.zmaw.de) is established, suitable for data during the scientific project phase as well as storing long-term archive data. ICDC aims to make data out of different internal and external archives easily accessible for the daily work of the KlimaCampus scientists. It extends the existing services by the announcement of data during the scientific project phase, a data portal and collaboration services. Therein ICDC utilizes the available infrastructure at the WDC Climate by using it for metadata storage and as long-term archive. The concept of ICDC, its functionality, its implementation status and future perspectives are presented.

Keywords: earth system science, data center, climate research, data portal, metadata, klimacampus, icdc

1 Introduction

For the KlimaCampus (http://www.klimacampus.de), the newly funded Cluster of Excellence in climate change research at the University of Hamburg, a data center is set up for the participating institutions (18 institutes of the University of Hamburg involving natural sciences, economics and social sciences; Max Planck Institute for Meteorologie - MPI-M; Institute for Coastal Research - GKSS) with the main focus on supporting project data work and transfering relevant data to the long-term archive of WDC Climate (WDCC): the Integrated Climate Data Center (ICDC: http://www.icdc.zmaw.de). ICDC hosts:

- Geodata:
  - Numerical model
  - Observations, Radar
  - Remote sensing (e.g. Satellite)
  - Research campaign
  - Data Products
- Non-Geodata:
  - Reports
  - Studies: social sciences, economics, security and peace research
  - Graphics
  - Animations
  - Analysis Tools

Scientific data of the KlimaCampus is saved in the file archives of the German Climate Computing Center (DKRZ: http://www.dkrz.de) and of the participating institutes during the project phase, generally without sufficient metadata. Furthermore, for scientific research data out of different sources, i.e. from different data originators, is needed, which is stored in established specialized data centers or in databases. Thus, ICDC has to bridge data heterogeneity and to combine data out of different internal and external archives to establish a data portal as support for Earth System Sciences (ESS) and climate impact research at the KlimaCampus. In this paper the concept is presented with the main focus on the functionality.

There are two trends observed in ESS: The data amount grows exponentially and the interest in the data becomes more interdisciplinary.

The DKRZ observes a growth in storage demand, which has crossed the 6 PByte in mid of 2007 (LAUTENSCHLAGER AND STAHL, 2007) and is predicted to reach nearly 60 PByte in 2013 (Figure 1; LAUTENSCHLAGER ET AL., 2008). Correspondingly, the main costs of the DKRZ shift from compute to storage. Therefore there are currently 10 PBytes per year available for long-term archiving. Only selected project data can be archived. The non-relevant data has to be removed after the end of the project. The ICDC supports scientists in this long-term archiving process.

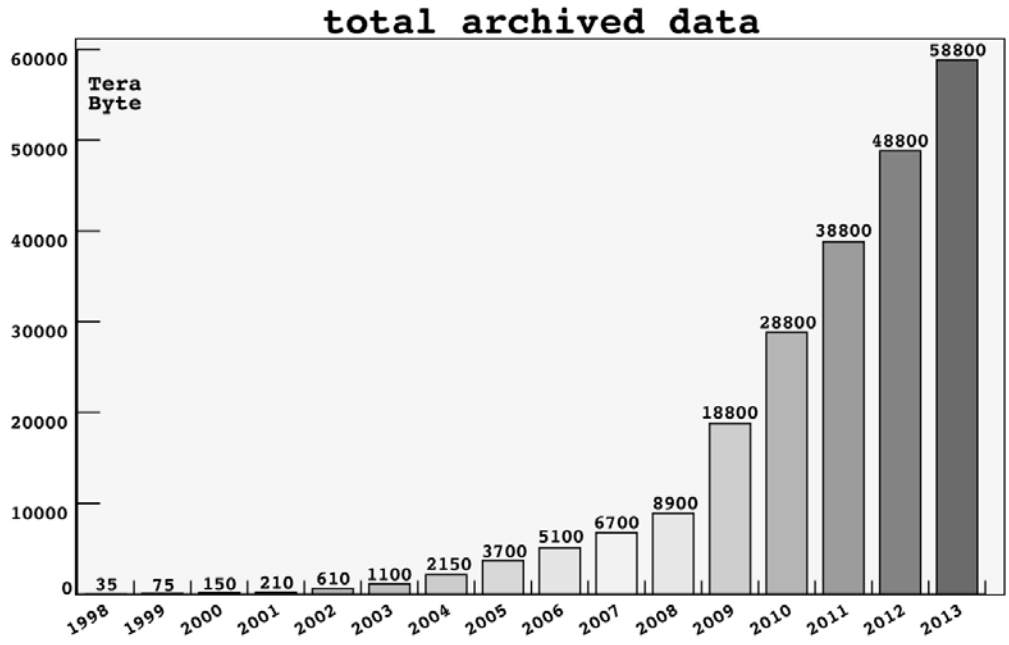

Figure 1: Development of the data archive at DKRZ (source: LAUTENSCHLAGER ET AL., 2008)

The user and provider communities for these data become more and more interdisciplinary and thus their topics of research and their data increasingly heterogeneous. E.g. at the KlimaCampus questions of economics and social sciences are examined as well as the more traditional aspects of natural sciences. Additionally, the data of central experiments is increasingly often accessed for scientific analyses and data (inter-)comparisons.

2 ICDC Concept

The Integrated Climate Data Center provides different services about scientific data as support for the scientists at the KlimaCampus (Figure 2). The focus lies on the support during the project phase. Data essential for the project is made searchable via the ICDC portal, including internal data sources (DKRZ and institute archives, CERA DB long-term archive) and data from external data providers (section 3.1). Furthermore ICDC provides new data products to enlarge the KlimaCampus database by data of high quality and of general scientific interest. The created data and data products are made available to others, - the scientific community or exclusively to project members or selected scientists -, with low efforts, quickly and transparently (section 3.2). For this a coarse metadata profile was developed for ICDC's project data, including the adaptation of the WDCC's metadata service. For further assistance for the collaboration between project members or members of research groups, different collaboration services are provided by ICDC (section 3.3). Finally, the long-term archiving process after the end of the project is attended by ICDC (section 3.4), in which selected project data is handed over to the long-term archive CERA DB.

3 ICDC Functionality – Data Lifecycle

To establish its integrated data services for the scientists, ICDC uses the existing infrastructure components at the KlimaCampus as much as possible: long-term archive CERA database, CERA2 metadata base with an individual ICDC profile, derived form the core CERA2 metadata schema (LAUTENSCHLAGER ET AL., 1998), and the individual local storage facilities of the KlimaCampus institutes (Figure 2). ICDC adds a data portal and collaboration services for the project phase to the infrastructure as well as in co-operation with the WDCC the integration of scientific data of external partners in the data search.

ICDC is to establish a data portal for these services, including data search, data provision, data-related information, and scientific collaboration. Since the current and future demands of the scientists cannot be fully known, adjustments and extensions of the ICDC portal are inevitable. This easy-adjustability to user demands was an important aspect for the decision for Zope/Plone (http://www.zope.org/, http://plone.org/) as web development framework and content management system as well as the active developer and user communities

of this open-source product. The ICDC portal is developed on Plone 3.2.3 with an interface to CERA2 metadata base using Oracle instantclient, SQLAlchemy for database connection and cx_Oracle for metadata access. Figure 2 gives an overview over the ICDC services, partly provided by WDCC. For the authentication of internal users for the collaboration services, for the metadata editing service of WDCC, and the access of internal project data the local LDAP servers are integrated.

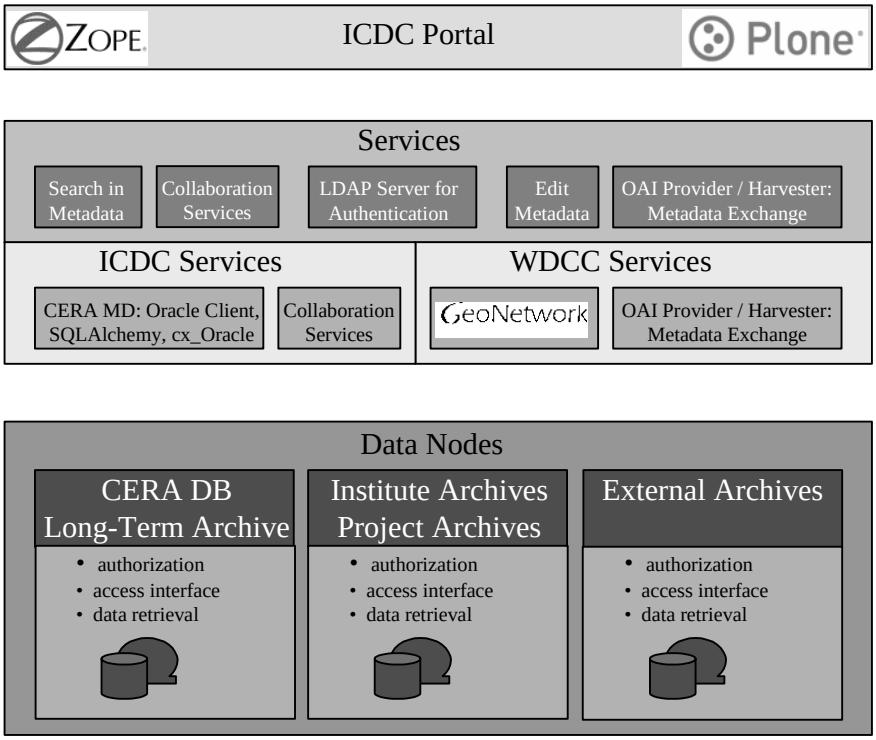

Figure 2: Technical overview of the data services of the ICDC portal

For the inclusion of scientific data of external partners in the ICDC catalogue, a co-operation with other climate data centers has to be established together with the WDCC. Generally, there exist two categories of data center co-operations:

- Bilateral, informal, and rather loosely co-operations between individual data centers with
  - Metadata exchange or
  - Data exchange,
- Multilateral, formal, and rather highly organized federations of multiple data centers in Data Grids.

The first is widely applied and easily set up because of its bilateral relation and the possibility to control the exchange. Metadata exchange has become more common than data exchange caused by the large data amounts. The data federations for Data Grids need organizational and technical structures. Usually, the metadata of partners is collected in a central metadata index or database. For user authorization during data access, an agreement about user roles and their accompanied data access rights has to be reached.

The authorization in data center federations is a sensitive security issue, which is not entirely solved, yet. All approaches include an automated user authentication by Globus GSI (Grid Secure Infrastructure) using short-living proxy certificates as user credentials and for delegation based on X509 Public Key cryptography. Some Data Grids like C3Grid (Collaborative Climate Community Data and Processing Grid, http://www.c3grid.de/portal) initially offer only open-source data. The ESG (Earth System Grid, http://www.earthsystemgrid.org) uses OpenID (SIEBENLIST ET AL., 2009) for authorization, which includes an additional attribute service for user role consideration, but it lacks the definition of administrative rules about building a federation of organizations and trusting attributes like Shibboleth (http://shibboleth.internet2.edu/).

Thus ICDC decided to avoid a complex Data Grid solution, unless the tool development of such a distributed authorization mechanism has reached a ready-to-use state. Instead a metadata exchange is established to enable a combined data search and leave the authorization at the different internal and external data providers. For the exchange of metadata with external data centers and databases usually an internationally well-known metadata model is applied, like Dublin Core (http://dublincore.org/) or in geosciences increasingly often the ISO 19115/19139 format (ISO 19115, 2003). The latter is part of the INSPIRE guideline of the EU (INSPIRE,

2007 AND 2009), which was established for homogenization of available geodata in Europe and therefore applied by ICDC. The ICDC metadata schema includes all information of ISO 19115/19139, the mapping is performed by WDCC together with the data of the long-term archive. Technically, the (experiment) metadata for exchange is provided on OAI servers (Figure 3). External metadata is harvested using the OAI-PMH protocol (LAGOZE ET AL., 2008), which is based on HTTP and XML, and then inserted in the ICDC metadata base. A prototypical metadata exchange with the DWD (CDC – Climate Data Center, http://cdc.dwd.de/catalogue/) is currently set up.

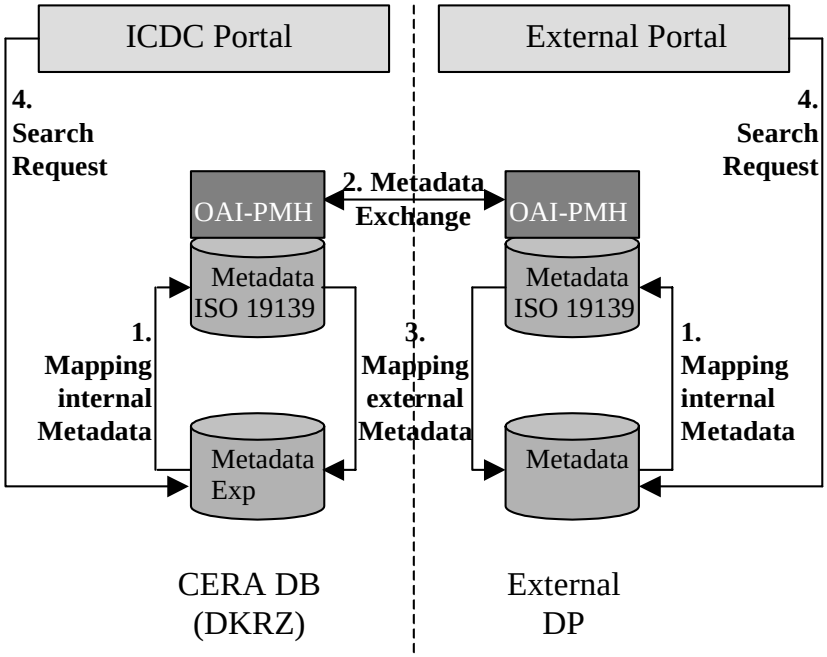

Figure 3: Exchange of metadata with external partners

In the following sections the lifecycle of scientific data is presented: Data is searched and retrieved via the ICDC portal (section 3.1). Within the research project new data products are created or derived, which are announced as project data for the research community (section 3.2). Scientific collaboration between the project members about results is performed in the private workspaces (section 3.3). Finally, at the end of the project, selected data is stored in the long-term archive CERA (section 3.4).

3.1 Search and Retrieve Data

For working on a scientific project initial data is needed as input data or for reference. This data is provided via the ICDC portal. It is stored internally at the KlimaCampus or externally at partner institutes (Figure 4) and can be distinguished in three types (Figure 2):

- **Internal Project Data**: Data extensively worked with in scientific projects and not fully analyzed, yet, which is of interest for a rather small group of specialized scientists. This is stored on local institute resources and partly on ICDC's server. Data descriptions (metadata) are of small detail and provided for a scientific experiment (aggregation of multiple datasets).
- **Internal Long-term Archive Data:** Data of interest for a larger group of scientists and non-scientists, which is to be provided for at least 10 years. Therefore the annotated metadata is of great detail to ensure a provider-independent data retrieval and usage, i.e. every dataset of a scientific experiment is described. For this data the existing CERA DB of the WDC Climate is used (http://cera.wdc-climate.de).
- **Data of External Partners:** Data stored at partner institutes is too large to be copied to ICDC or CERA DB. Therefore a metadata exchange in a standardized metadata model is performed: ISO 19115/19139. In this way, data of partners like the German Weather Service (DWD) is directly retrievable via the ICDC portal. Other associated partners are the Federal Maritime and Hydrographic Agency (BSH), GKSS, IFM-GEOMAR and WDC Mare.

The ICDC portal provides two entry points for data investigation: predefined catalogues and an advanced search over keyword/project (combined), content and spatial-temporal extent parameters (Figure 5 - left). The request is sent to the CERA2 metadata base (cf. section 3.2) and returns a list of numbered results descendingly ordered by

relevance (Figure 4, Figure 5 – right). To receive at least a few hits, the different search criteria are initially connected by a logical "or". For every hit short name and summary are displayed. Additional information is accessible by a couple of provided links.

The "data" link is a reference to the experiment directory or to the list of datasets (CERA DB GUI). The search request and the metadata are accessible without authentication. Data can be downloaded by the user directly in case of open-source data or after a previous authentication for restricted data.

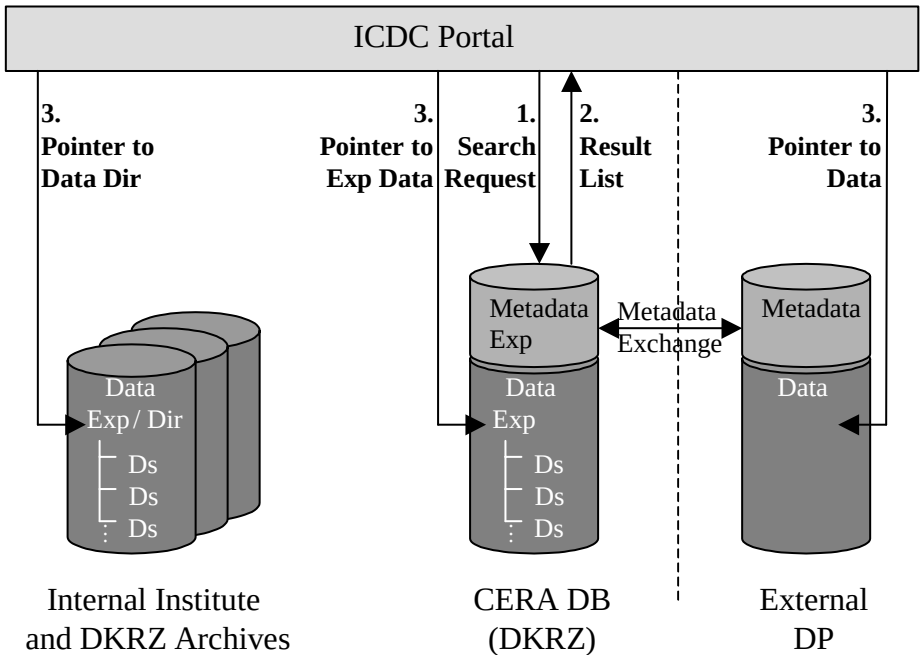

Figure 4: Data Search in the ICDC portal (Exp: Experiment, Ds: Dataset, Dir: Directory)

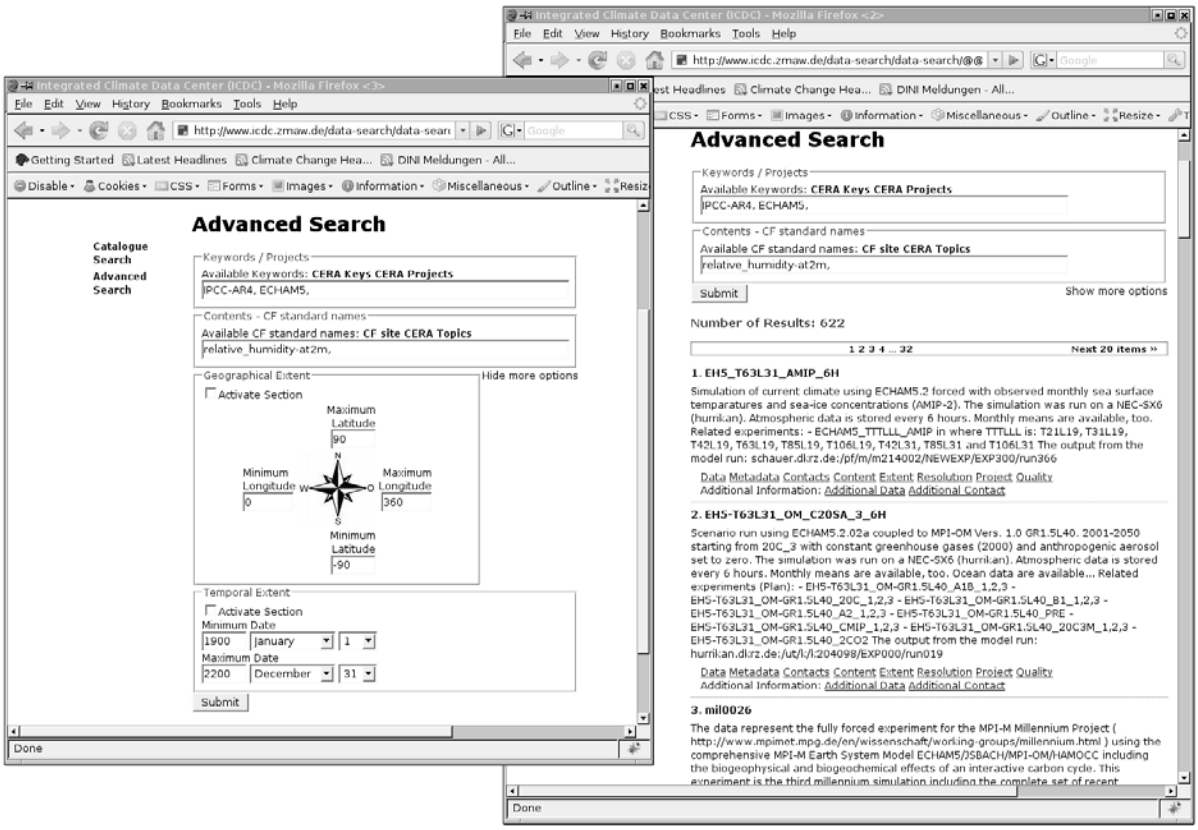

Figure 5: Advanced Search in ICDC Data: request form (left) and list of results (right)

Another task to be handled by the data providers is the authorization of users. ICDC can supply a link to the data location, but download permissions are provided by the institutes. With ICDC's aggregated data description approach, the heterogeneity of the data can be bridged, but on the expense of an information loss of the structure of the data in the archive directory, i.e. of the experiment's internal structure. The question "which part of the information is stored in which dataset" cannot be answered by the metadata. Therefore ICDC recommends the usage of the self-describing data format netCDF (REW ET AL., 2009; http://www.unidata.ucar.edu/software/netcdf/) for binary data the application of the ICDC guideline for structuring the data directory on the institute server. On the one hand, this authorization approach may cause some effort for a user to contact the data provider for download permission, on the other hand, it reduces the effort to be spent on convincing scientists to announce their project data because of the low effort for metadata preparation and the remaining authority over their data.

Internal users are identified on internal machines using LDAP authentication, external users need to login on KlimaCampus machines, as well as KlimaCampus users on external databases for data access. Data access mechanisms include CERA DB download (jblob), ftp server download, and login+scp. Other more comfortable access mechanisms like WebGIS or OPeNDAP servers are planned by some institutes or by ICDC, respectively.

3.2 Creation and Announcement of Project Data

During a scientific project new data or data products are created. To provide new data products of high quality for the scientists at the KlimaCampus is a second main topic of ICDC in order to improve and enlarge the available database. Exchange of project data with colleagues is essential as well as the protection of not-yet-published results to externals. ICDC provides the opportunity to announce the data without giving up the authorization over this data. Furthermore, only low effort has to be spent on the preparation of metadata. The scientist has to structure the data in an experiment directory on a central server and describe this structure in an accompanied README file in the uppermost directory level (Figure 6 – left).

Other additional recommendations for the data providers are:

- Use the self-describing netCDF format for binary data and well-documented ASCII for non-binary data and small data amounts,
- Make the data easy-available, favorable as open-source or develop a simple role concept,
- Offer download possibilities as simple FTP or more comfortable as OPeNDAP (SGOUROS, 2004**,** http://opendap.org/) or WebGIS GUI (web implementation of OGC interfaces, Open Geospatial Consortium, http://www.opengeospatial.org/) and avoid the use of unix logins unless the data access is restricted to internal users or an exclusive group of project members.

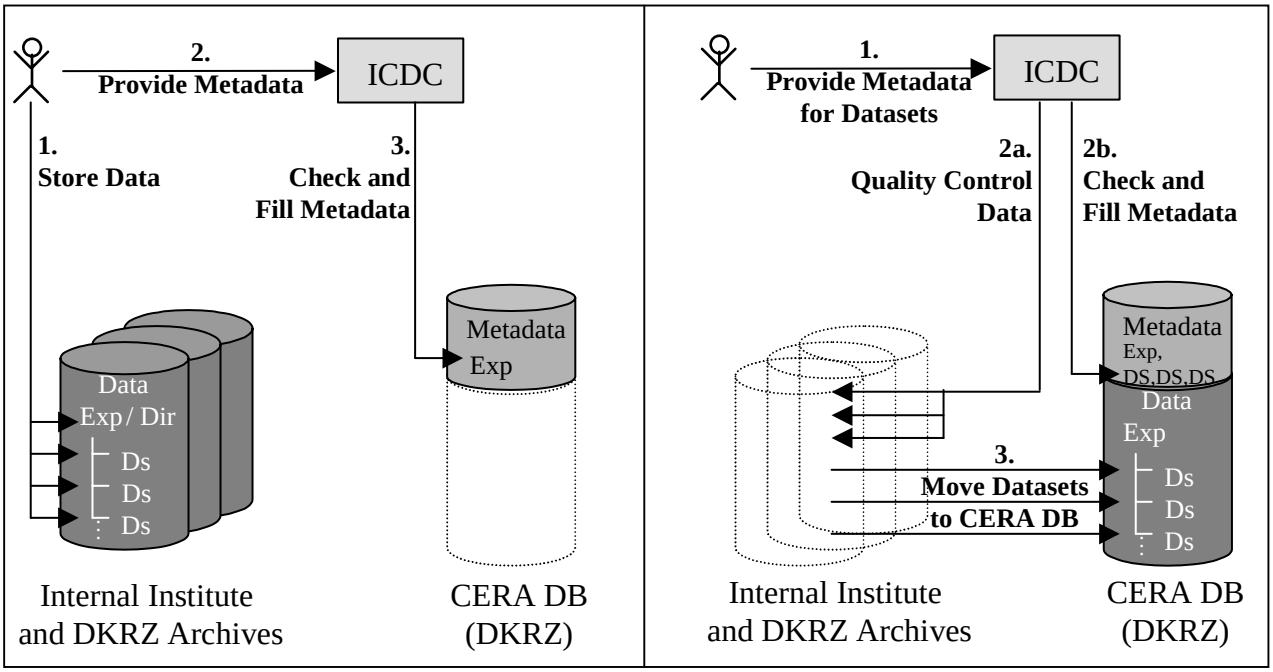

Figure 6: Announce Project Data (left) and Long-term Archive Project Data (right)

For the announcement of project data, metadata (data descriptions) has to be provided and filled into the metadata base CERA2. Then the metadata is created using the XML metadata template directly or by the help of CERA2's metadata editor (http://anticyclone.dkrz.de:8088/geonetwork), which is based on geonetwork (http://geonetwork-opensource.org/) and adapted to the ICDC metadata profile. This metadata is handed over to ICDC, where the metadata is checked and filled in the CERA2 metadata base. At that point the data is visible in the ICDC portal. Finally, the scientist is usually asked to check the metadata.

The ICDC metadata model is coarse and simple and therefore metadata is easy-to-provide for the scientists. It includes basic bibliographic information, quality information, summary, extent, content, and access (data location and access permissions). An extension for lineage or provenance or quality level information is currently developed by WDCC. For content description, the Climate and Forecast convention (CF standard names; EATON ET AL., 2009, http://cf-pcmdi.llnl.gov/documents/cf-standard-names/standard-name-table/current/cf-standard-name-table.html) is used, a well known standardization for ESS parameters. For the definition of spatial-temporal resolutions the CF standard is applied for model grid descriptions and the FDGC (Federal Geographic Data Committee; map projection; http://www.fgdc.gov/) standard for map projections.

The ICDC metadata schema has a hierarchical order from project (scientific topic under consideration) via the concrete scientific experiment, which produced the data, to the pieces of additional information belonging to the experiment (Figure 7, Appendix). In contrast to the CERA2 metadata model for the CERA long-term archive, the ICDC metadata includes only the aggregated information about the scientific experiment, but the individual datasets are not annotated by metadata. For data from economics and social sciences the metadata schema is reduced to bibliographic information, quality and summary with the possibility to specify additional information pieces like documentations (e.g. reports or campaign diaries) or materials (e.g. recorded TV programs or newspaper articles) or special analysis tools or graphics or animations. It was agreed with WDCC to prepare several predefined lists of values for CERA2 in order to improve data search requests by granting the establishment of a homogeneous metadata usage (e.g. for the spatial-temporal resolutions).

3.3 Scientific Collaboration

In a scientific project not only the data but information concerning the data is exchanged: its quality, its creation, its field of application, etc. At the moment, personal communication, mailing lists or phone conferences and wiki pages and used. Currently, ICDC offers the metadata as source of information about data. Additional services are planned for scientific collaboration:

- a forum for discussion about data-relevant themes and
- private workspaces for individual scientists or for research groups to introduce their data or to announce information about it, transparently.

In their private workspaces users are responsible for assigning access rights, which decides with whom to share this information. This collaboration area will be an area exclusively for authorized users using LDAP authentication. Its services will be initially offered for a period of time and then evaluated. Depending on the demand, a service will be either enlarged and advanced or not further extended. An example for such a collaborative data portal is the GEON Grid (http://portal.geongrid.org).

3.4 Archive Data in CERA DB

Within a year after the termination of the scientific project and before the end of the scientist's contract, the main results of the project are selected for long-term archiving for at least 10 years. For this the scientist has to add metadata for every dataset belonging to the already described experiment and hand it over to ICDC (Figure 6 – right). ICDC has to fulfill two tasks, before the metadata is filled in CERA2 and WDCC starts to fill the datasets into the CERA DB: quality control of the data and check of the metadata. Afterwards, a second thorough review process can be passed to achieve a higher quality assurance level, a STD-DOI publication (Scientific and Technical Data - Digital Object Identifier, http://www.std-doi.de). The process for awarding a STD-DOI was developed by WDCC and the Technical Information Library Hannover, TIB (LAUTENSCHLAGER, 2008).

4 Status and Future Directions

The Integrated Climate Data Center started in November 2008. Since then ICDC has developed a concept for the new data center, has purchased the required hardware, and has derived an appropriate metadata profile for project data from the CERA2 metadata schema. The ICDC portal went online in August 2009 initially focusing on data retrieval. The metadata editor of CERA2 was announced about the same time, which will ease the meta-data creation for scientists, decisively. The integration of external data as well as the collaboration services

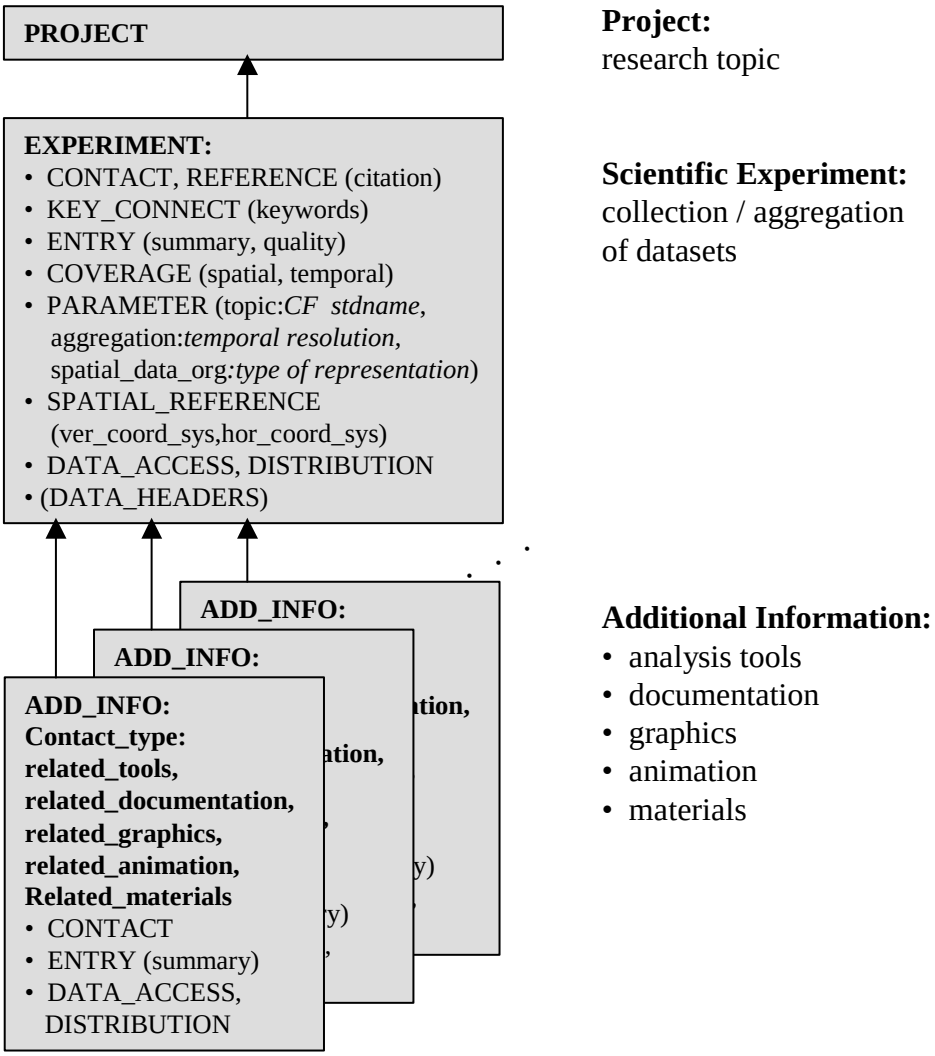

Figure 7: ICDC Metadata Profile

including the required LDAP authentication are under way. Apart from that, there exist several improvement and extension possibilities towards more user friendliness and for additional functionalities. In this early stage the data provider as well as the data user communities are still quite small. More documentation and a user workshop are planned, which are essential to interest scientists in this new data portal.

Therefore the next aims of ICDC are:

- User information to enlarge the provider and user communities and to recruit some power users for testing
- Improvement of the ICDC portal including new functionalities on user demand,
- Offer of additional communication possibilities in the ICDC portal like a forum or a help center or private workspaces for individuals or groups,
- Establishment of a user support infrastructure,
- Set-up of an OPeNDAP server for netCDF data access to gain experiences and to become able to help institutes installing their own OPeNDAP server,
- Improvement of the recommendations for the directory structure of the data servers of the institutes,
- Enlargement of the external data exchange with new partners and more experiments.

References

EATON, B., J. GREGORY, B. DRACH, K. TAYLOR, S. HANKIN (2009): NetCdf Climate and Forecast (CF) Metadata Conventions, Version 1.4, http://cf-pcmdi.llnl.gov/documents/cf-conventions/1.4/cf-conventions.pdf.

INSPIRE (2009): INSPIRE Metadata Implementing Rules: Technical Guidelines based on EN ISO 19115 and EN ISO 19119, Version V1.1, http://inspire.jrc.ec.europa.eu/reports/ImplementingRules/metadata/MD_IR_and_ISO_20090218.pdf.

INSPIRE (2007): Directive 2007/2/ec of the European Parliament and of the Council of 14 March 2007 establishing an infrastructure for spatial information in the European Community (INSPIRE), Official Journal of the European Union 50(108), http://eur-lex.europa.eu/JOHtml.do?uri=OJ:L:2007:108:SOM:EN:HTML.

ISO 19115 (2003): ISO/TC 211 - Geographic information / Geomatics.

LAGOZE, C., H. VAN DE SOMPEL, M. NELSON, S. WARNER (2008): The Open Archives Initiative Protocol for Metadata Harvesting, Version 2.0, http://www.openarchives.org/OAI/openarchivesprotocol.html.

LAUTENSCHLAGER, M. (2008): Preservation of Earth System Model Data, in: Digital Preservation Europe, Briefing Paper 30th June 2008, http://www.digitalpreservationeurope.eu/publications/briefs/preservation-of-earth-system-model-data.pdf

LAUTENSCHLAGER, M., W. STAHL, J. BIERCAMP (2008): Longterm Archiving of Climate Model Data at WDCC and DKRZ, presentation at NCAR on 2008-11-18; http://www.mad.zmaw.de/fileadmin/extern/lectures/NCAR-Visit-Lautenschlager-LongtermArchive-27-291008.pdf

LAUTENSCHLAGER, M., W. STAHL (2007): Long-Term Archiving of Climate Model Data at WDC Climate and DKRZ - In: E MIKUSCH (ed.): PV2007 - Ensuring the Long-Term Preservation and Value Adding to Scientific and Technical Data, Conference Proceedings. DLR, German Remote Sensing Data Center, Oberpfaffenhofen, 2007; http://www.mad.zmaw.de/fileadmin/static/IPCC_DDC/html/Lautenschlager_LongTermArchivingClimateModelData.pdf

LAUTENSCHLAGER, M., F. TOUSSAINT, H. THIEMANN, M. REINKE (1998): The CERA-2 Data Model, DKRZ series No. 15, doi: 10.2312/WDCC/DKRZ_Report_No15 http://www.mad.zmaw.de/fileadmin/extern/documents/reports/ReportNo.15.pdf.

REW, R., G. DAVIS, S. EMMERSON, H. DAVIES, E. HARTNETT (2009): The NetCDF Users Guide, NetCDF Version 4.0.1, http://www.unidata.ucar.edu/software/netcdf/docs/netcdf.pdf.

SGOUROS, T. (2004): OPeNDAP User Guide, Version 1.14, http://www.opendap.org/pdf/guide.pdf

SIEBENLIST, F., R. ANANTHAKRISHNAN, D.E. BERNHOLDT, L. CINQUINI, I.T. FOSTER, D.E. MIDDLETON, N. MILLER, D.N. WILLIAMS (2009): Earth System Grid Authentication Infrastructure: Integrating Local Authentication, OpenID and PKI, TeraGrid'09, 22.-25.06.2009, Virginia, USA; http://www.teragrid.org/tg09/files/tg09_submission_79.pdf.

Contact of author
University of Hamburg / KlimaCampus
Integrated Climate Data Center
Grindelberg 5
D-20144 Hamburg
martina.stockhause@zmaw.de
+49-(0)40 42838 – 7780
www.icdc.zmaw.de

Acknowledgements
The main author thanks her colleagues at the WDCC Michael Lautenschlager, Heinke Höck, Hannes Thiemann, Hans-Hermann Winter, and Hans Ramthun as well as Stephan Kindermann of the DKRZ for their continuous invaluable support.

Appendix: ICDC Metadata Profile