ADOPTING RDA OUTPUTS FOR ... CLIMATE DATA

RESEARCH DATA ALLIANCE

DKRZ adopts 6 RDA outputs for climate data modelling

DKRZ is integrating persistent identifiers for use cases supporting precise data tracking, automated replication and versioning, custom and early data citation into the Earth System Grid Federation data infrastructure which supports WCRP CMIP6 data provisioning. This requires elemental PID information to be interoperable across multiple services and tools and formulating community-specific PID profiles. Furthermore, future automated processing workflows could leverage such information as well if bound to specific data types and brokered through a dedicated service.

To give structure to possibly huge numbers of objects and their identifiers, the services and tools involved can also benefit from a possible RDA recommendation on research data collections.

The Challenge

"Current data management practices still rely largely on managing files and directories in file systems. Factors such as the relative increase of data volumes compared to available network bandwidth and the easy availability of remote and on-demand computing resources are drivers behind bringing processing and data closer together. National and international policy changes in Earth Science funding may also cause a shift in the skills and expectations of archetypical data service users. "

Says Tobias Weigel, a Computer scientist at the adopting organisation, Deutsches Klimarechenzentrum (DKRZ)

Together, these factors lead to scenarios where it will be increasingly difficult to manage data on a per-file or per-directory basis and deal with data transfer, replication and life cycle management at a comparatively low level of automation. Future tools may intentionally hide the location and structure of scientific data objects from the user, requiring more intelligence from back-end services. Services that provide easy data preparation and processing and make data provenance transparent may be particularly valuable for interdisciplinary users unfamiliar with established community practices.

Weigel continues "Without solutions that increase automation, costs of maintaining services will increase, which would have a deteriorating effect on service quality or lessen resources available for developing new services required for future user demands. Past experience has shown that many tasks such as data transfers or replication suffer from manual intervention required as long as no comprehensible data tracking solution is in place. Such tasks may take up even more resources given that the data volumes and number of objects to manage increases exponentially.

RDA RECOMMENDATIONS ADOPTED

Data Foundation and Terminology: simplify understanding and communication about basic concepts such as digital object and persistent identifiers.

PID Information Types: describe PID record profiles and benefit from types possibly defined across communities.

Data Fabric: gain a better understanding of high-level modularity and make decisions concerning operations and strategic development.

Data Type Registries: register PID information types and eventually bind types with processing services.

Dynamic Data Citation: clarify data citation principles and evaluate implementation options for potentially unstable data.

ANSWERING COMMUNITY NEEDS

The community that benefits from middleware improvements in climate data management spans across multiple organizations internationally.

Of particular mention within the European context is the **European Network for Earth System Modelling (ENES)**. The breadth of scientific end-users however reaches beyond the core climate modelling community and can involve other disciplines such as climate impact research, adaptation and mitigation policies, public services, agriculture and so on, as climate change is a global phenomenon and challenge and touches on a huge number of areas.

WHY RDA

RDA put itself forward as a venue where experts from different disciplines and shaped by different community practices can combine their collective knowledge to build roughly workable solutions within manageable time spans. The RDA outputs to move to pilot adoption were selected by projecting future application scenarios and long-term challenges and finding good matches with manageable gaps with the existing solutions at DKRZ.

Find out more

Visit RDA @ rd-alliance.org

Email: enquiries@rd-alliance.org

The Adoption

The adoption of the mechanisms of the PID Information Types recommendation and the conceptual basis provided by the Foundation and Terminology group have shaped the implementation of a PID and object management service for climate modelling data that has been going on since late 2014. The implementation process is roughly two-fold, with policy and technical aspects.

A major effort at DKRZ for the coming years is the data management for CMIP6, which is technically done through an internationally hosted data infrastructure called ESGF, collaboratively built and operated by a community of climate data centres around the globe. DKRZ participates in ESGF technical development, governance and operations. The starting phase of the CMIP6 effort has led to a maturing in governance and development practices for ESGF, and as part of this, DKRZ has put forward the concept of a PID service that can support key data management processes within ESGF. The policy agreements that led to such a service being supported for implementation took place over the course of 2015. The RDA outputs have been useful in this regard to clearly communicate the benefits and requirements for such a solution. Technical development also started in 2015 and underwent pilot and prototype stages, building upon the wide expertise gathered through participation in RDA events, and at its technical core on the PID typing concepts. Data citation plays a major role in end-user services of ESGF, and the RDA recommendations and interactions with other organizations such as Force11 and DSA have contributed to better understanding of citation policies for ESGF.

DKRZ is also participating in the Horizon 2020 EUDAT project and has subsequently deployed concepts of PID typing for the cross-disciplinary EUDAT services. The high-level concepts proposed by the Data Fabric group have contributed to this as well, by fostering a better understanding of modularity through dedicated components and processes or policies to operate them. Regarding further use of data types, DKRZ has included ideas to build a pilot around processing services, service brokering and in a proposal revolving around the combination of climate modelling data and observational data. If approved, this may start in 2017.

To develop and efficiently operate a large-scale PID solution as part of a critical infrastructure such as ESGF requires a clear understanding of the requirements, user benefits, possible trade-offs and state of the technical solutions available.

The outputs of RDA have informed the necessary decisions and development plans in a significant way and while naturally not all future concerns can be addressed only by looking at the combined RDA outputs, there is yet a better understanding of possible roadblocks and areas for further investigation.

The development of PID services for ESGF CMIP6 involves multiple people at DKRZ, including Katharina Berger, Merret Buurman, Andrej Fast, Stephan Kindermann, Michael Lautenschlager, Martina Stockhause and Tobias Weigel

Lesson Learnt

Adoption takes time, patience and a working social community infrastructure. Without adequate governance processes that steer development and bolster it by reliable decision-making, the adoption efforts might have gotten stalled or received too little resources to be pushed over the initial pilot project hump.

The challenges and requirements inherent in other communities and differences in legacies and resources are probably too broad to recommend a single adoption strategy. In the case of climate data infrastructures, adoption followed a schema of building and evaluating pilots, securing agreement and performing development, the latter with a significant amount of resources. This very much resembles common software engineering practices and is not atypical, yet the details will likely differ for other communities, particularly concerning governance and legacy.

DKRZ
DEUTSCHES KLIMARECHENZENTRUM

Dr. Tobias Weigel
Computer scientist Deutsches
Klimarechenzentrum (DKRZ)
https://www.dkrz.de

DKRZ German Climate Computing Center

DKRZ (German Climate Computing Center) is a national German facility, providing state-of-the-art super-computing, data and other associated services to the German and also the international scientific community to conduct top of the line Earth System and Climate Modelling. DKRZ operates a fully scalable supercomputing system designed for and dedicated to earth system modelling including mass storage system to a capacity of at least 400 PByte.

DKRZ is partner in ENES (European Network for Earth System Modelling) and is one the representatives of the Earth system research communities in the EUDAT project. DKRZ is operating the ICSU World Data Centre Climate (WDCC), a community specific long-term data archive. Linked to WDCC, DKRZ provides best practice examples in scientific data life cycle management for the Earth system research community (federated data infrastructures, long-term archiving service, grid-based data processing workflows).