D4.3 Report on EC-US workshop

Grant Agreement number: 312118

Project acronym: COOPEUS

Project title: Strengthening the cooperation between the US and the EU in the field of environmental research infrastructures

Funding Scheme: FP7.INFRASTRUCTURES-2012-1

Name, title and organisation of the scientific representative of the project's coordinator:

Dr. Ketil Koop-Jakobsen

Marum – Bremen University

Loebnerstrasse

28359 Bremen

Tel: +49 421 218 - 65629

Fax:

E-mail: kjakobsen@marum.de

Project website address: http://www.coopeus.eu/

Contents

Strategic workshop on future harmonization of data sharing among Research Infrastructures 11.04.2013 09:00-12:00 at EGU2013, Vienna, Austria

Executive summary:

COOPEUS and EUDAT hosted a well-attended workshop on future harmonization of data sharing among Research Infrastructures (RIs) at the EGU2013-conference. The workshop aimed at identifying and prioritizing future challenges and impediments for data-sharing globally among environmental research infrastructures in a short-term as well as long-term perspective. The following topics, which had been selected based on an immediate need for the involved research infrastructure projects, were debated with an audience of 40 experts from EU and US representing more than 15 different research infrastructures and/or research infrastructure projects:

- Future challenges in the use of Persistent Identifiers (PIDs)
- Identifying synergies across research infrastructures in regard to workflow and data processing steps and identifying impediments for data-sharing
- The future data infrastructure landscape – how we learn from GEOSS.

The workshop identified key challenges for the use of PIDs and revealed an urgent need for further work in this area. A list of pressing challenges for data-sharing among Research infrastructures were produced, however, prioritization of the identified challenges based on participant's opinion did not reveal any trends. There was a general consensus that future projects on data and knowledge sharing among research infrastructures can learn from the experience that GEOSS has gain in past decades.

This workshop formed the foundation for a working group on PIDs, which subsequently held a 2-day workshop on PIDs in open time-series and will continue its work within the framework of the Research data alliance (RDA).

Participants list:

Name	First Name	Organization	Project Involvement	EU/US
Aguilar	Fernando	CSIC	COOPEUS	EU
Annoni	Allessandro	EC-JRC	INSPIRE/ICOS	EU
Arko	Bob	LDEO	RZR/ODIP	US
Asmi	Ari	Uni Helsinki	COOPEUS/ENVRI/Partner	EU
Bailo	Daniele	INGV	EPOS	EU
Beranzoli	Laura	INGV	COOPEUS/ENVRI	EU
Brus	Magdalena	Uni Helsinki	COOPEUS/ENVRI	EU
Cocco	Massimo	INGV	EPOS	EU
Darroch	Louise	BODC	Oceanographic Data Management/Seadatanet	EU
Euteneuer	Frieder	GFZ Potsdam	EPOS	EU
Fiameni	Guiseppe	CINECA	EUDAT/ICORD/EPOS	EU
Glaves	Helen	BGS	ODIP	EU
Hoffmann	Thomas	GFZ Potsdam	EPOS	EU
Huber	Robert	Uni Bremen	COOPEUS	EU
Koop-Jakobsen	Ketil	Uni Bremen	COOPEUS	EU
Leaobetter	Adam	BODC	Seadatanet/ODIP	EU
Loescher	Hank	NEON	COOPEUS	US
Marco	Jesus	CSIC	COOPEUS	EU
Meertens	Charles	UNAVCO	COOPEUS	US
Michelini	Alberto	INGV	EPOS/EUDAT/VERCE	EU
Nativi	Stefano	CNR	ODIP/Seadatanet/EC/GEOWOW	EU
Patten	Kim	AZGS	Earthcube	US
Pearlman	Francoise	JdFE	COOPEUS/Earthcube	US
Pearlman	Jay	JdFE	COOPEUS/Earthcube	US
Purss	Matthew	Geoscience Aust.	NIC	EU
Riedel	Morris	Juelich Supercu.	EUDAT	EU
Rowan	Linda	UNAVCO	UNAVCO	US
Schaap	Dick	MARIS	SDW Seadatanet	EU
Schentz	Herbert	UBA-A	EUDAT/ENVRI	EU
Schneider	Nadine	CEA/LSCE	COOPEUS/ENVRI/ICOS	EU
Schubert	Chris	JRC	INSPIRE	EU
Sorvari	Sanna	FMI	COOPEUS/ENVRI	EU
Spinuso	Allessandro	KNMI	VERCE/NERA/EPOS	EU
Torn	Margeret	U.C. Berkeley	the Climate and Carbon Sciences Program	US
van Eck	Torild	ORFEUS/EPOS	NERA/VERCE/COOPEUS/ENVRI	EU
Waldmann	Christoph	Uni Bremen	COOPEUS	EU
Wier	Stuart	UNAVCO	UNAVCO	US

Introduction.

The workshop was organized in a joint collaboration between the FP7-projects COOPEUS and EUDAT represented by Ketil Koop-Jakobsen (COOPEUS - UNI-Bremen), Robert Huber (COOPEUS, ENVRI - UNI-Bremen), Christoph Waldmann (COOPEUS - UNI-Bremen), Morris Riedel (EUDAT, Juelich Supercomputing Centre) and Peter Wittenburg (EUDAT, Max Planck Institute for Psycholinguistics). The workshop was conducted on 11.04.2013 09:00-12:00 at the EGU2013-conference, Vienna, Austria. (http://meetingorganizer.copernicus.org/EGU2013/session/13343). In order to assure high-level discussions, leading experts involved with data-sharing and research infrastructures development were carefully selected and invited by the organizers.

Discussion topic 1: PIDs (Chair: Robert Huber COOPEUS, ENVRI)

Synopsis: For the sake of transparency and reproducibility of research, it is crucial to be able to unambiguously identify and cite data that were used as the basis of a publication. Globally unique and resolvable, persistent digital identifiers (PID) for digital data sets is an important tool to achieve this goal. They enable unambiguous links between published research results and their underlying data (Data Citation) and publishing and connecting structured data on the Web (Linked Data). Several approaches towards the implementation of PID technologies as well as related workflows and policies are now running within e.g. COOPEUS, ENVRI and EUDAT. In this session, we discussed challenges and impediments in regard to assignment of PIDs across research infrastructures and scientific communities and debated the requirements for a joint implementation strategy. A particular issue is related to real-time data that have to be grouped to allocate PIDs and currently different concepts exists in regard to allocating PIDS to predefined data clusters.

Discussion Outcome Summary:

Identified challenges regarding PIDs:

Each scientific community currently has different schemes for the use and assignment of PIDs, which is an impediment for the interdisciplinary projects using data from multiple RIs. Hence to improve the interoperability between RIs, it will be advantageous to find common standardization guidelines; for example, by having a common matrix for PID assignment. Specific challenges relates to the following topics:

- PID assignment to larger non-permanent, dynamic data (ex. open time series)
- Timing of PID assignment (ex in relation to publications)
- Level of PID assignment, - for raw data – for metadata - for model data
- Reproducibility: what to do when PID-assigned datasets are changed? (ex. retrospective recalibration)

Challenges of a more general character relate to:

- Credits: Who get credit for the data?.
- Cost of DOI - a challenge in the Global perspective, for 3rd world countries
- Assuring persistency of data-archives (Data-centers)
- Capacity building (training data managers, recognizing small players too)

In order to find the optimal solution to the abovementioned challenges, there was a general understanding that it is of outmost importance to also involve research communities in the decision making, in particular the end-users. In this regards, it may be relevant to identify more stakeholders.

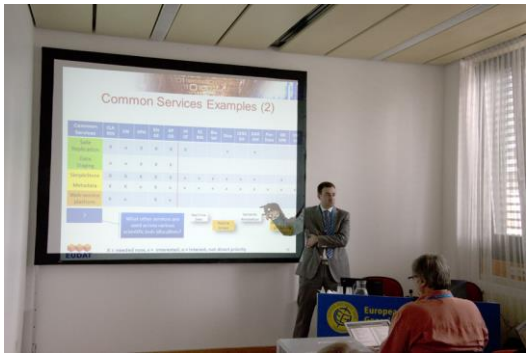
Discussion topic 2: Synergies (Chair Morris Riedel, EUDAT)

Synopsis: The steadily increasing amount of scientific data is a challenge in science and tackled in parallel by various research infrastructures and projects as well as other initiatives (e.g. Research Data Alliance). The goal of the discussion is to bring together researchers and practitioners working in the areas of scientific data management to exchange and share their experiences and new ideas around potential synergies such as commonly required services across different scientific (sub-) disciplines. The discussion includes an exercise in which we will identify the important topics that holds challenges or impediments to data management and exchange and to prioritize these topics in order to get an overview of the interests from the different scientific field represented at the work shop

Discussion Outcome Summary:

A wide variety of topics ranging from governance of RIs to inclusion of data from economics and social sciences, were identified as being of significant importance and as holding challenges or impediments to data management and exchange. The participants were asked to indicate, by show of hands, if they found the topic in question to be holding imminent challenges that should have priority. Even though this prioritization process revealed that

many topics could be identified as relevant to the different scientific fields represented at the workshop, there was no common denominator, which could be identified as the most important topic. Hence, an obvious starting point for work on a future harmonization strategy for RIs and datacenters could not be identified in this way.



Morris Riedel (EUDAT) presenting at COOPEUS/EUDAT workshop

Discussion topic 3: The future data infrastructure landscape (Chair: Christoph Waldmann)

Synopsis: The complexity of the landscape of available data and information produced by research infrastructures in the environmental field prevents a ubiquitous use. Although a number of projects and initiatives are currently around, like EarthCube, COOPEUS, EUDAT, iCORDI, making significant contributions to ease international and interdisciplinary data exchange, there is no easy solution to that issue. For more than a decade, The Group on Earth Observations (GEO) has been coordinating efforts to build a Global Earth Observation System of Systems (GEOSS) and has thereby gained a lot of experience in the field of connecting research infrastructures. "- With the many FP7-projects recently initiated with a similar scope, it is natural to ask – What can we learn from GEOSS?

Discussion Outcome Summary:

During the creation of the Geo portal, GEO learned a lot of lessons over the years including the following topics.

- The importance of documenting and reporting not only successful initiatives, but also things that did not work. For example; in the GEO context; the initial top-down approach failed, but there are indications that a bottom-up approach might work better.
- It is not enough to build the infrastructure. Assuring usage by training users and including users in the development process is important.

- The social aspects of Data-sharing are more problematic than the technical issues.
- Creating Interoperability has a long-term perspective and takes time.
- There are three major topics which must be clarified when connecting research infrastructures: 1) Vision 2) Tools 3) Resources

There was a general consensus at the workshop that future projects on data and knowledge sharing among research infrastructures should learn from the experience that GEO(SS) has gain in past decades in order to avoid repeating past mistakes.

Impact and perspective

In particular the debate of PIDs revealed and urgent need for work on standardization of PID assignment and consequently a forum debating the progression on the use of PIDs was initiated at the workshop. This was followed up by a 2-day workshop designated to the topic "PID for open time series" in Bremen June 2013 with more than 25 participants from various project within EC and US such as COOPEUS, EUDAT, ENVRI, KOMFOR, DataCite, RDA and EPIC. A First white paper on the assignment of PIDs in open time-series is in the process of being drafted, and the work on PIDs will continue in the data citation working group under the framework of the Research data alliance (RDA).