Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Summary of the kick-off workshop
(Barcelona, 11th February 2016)

7th RDA Plenary Meeting, Tokyo, 1-3 March 2016

Pierre-Antoine Bretonnière

Summary of the kick-off workshop

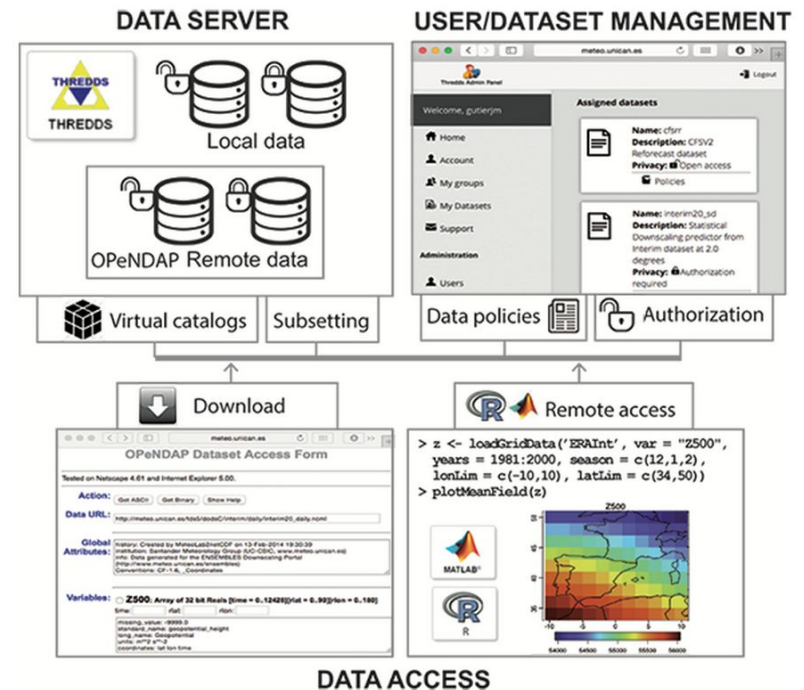Barcelona Supercomputing Center, 11th February 2016

- Kick-off meeting to launch the RDA Interest Group on weather, climate and air quality.

- 25 people from different research institutes, private and public companies and governmental organisations gathered to discuss the importance and need for such an interest group

- Speakers from BSC, JRC, ENEA and

Universidad de Cantabría

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

End user gateway for climate services and data initiatives (Antonio Cofiño, Universidad de Cantabria):

- UDG portal provides homogeneous access end point to collections of relevant variables. Deals with the different terms of use in a single data service without worrying about the complexities of the datasets, the file names or organizations.

- Access to remote datasets and links to the Earth System Grid Federation (ESGF), as well as a local downscaling capability and the link to external services provided by the climate4impact portal.

Main points from the presentations

Meteo-climate data at JRC-Mars (Andrea Toreti, JRC):

- Typical workflow at JRC includes: **data retrieval**, **QC**, inclusion in a **database** used to run crop models, and **visualization** of the data through an internal tool as well as a data portal that provides NetCDF files with different variables and the mars explorer.

- Need to find the equilibrium between being flexible enough to **handle the variety** of data and formats to be retrieved and processed daily and the **reaction in real time** to the complexity of the user needs, dealing with different spatio-temporal scales.

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA
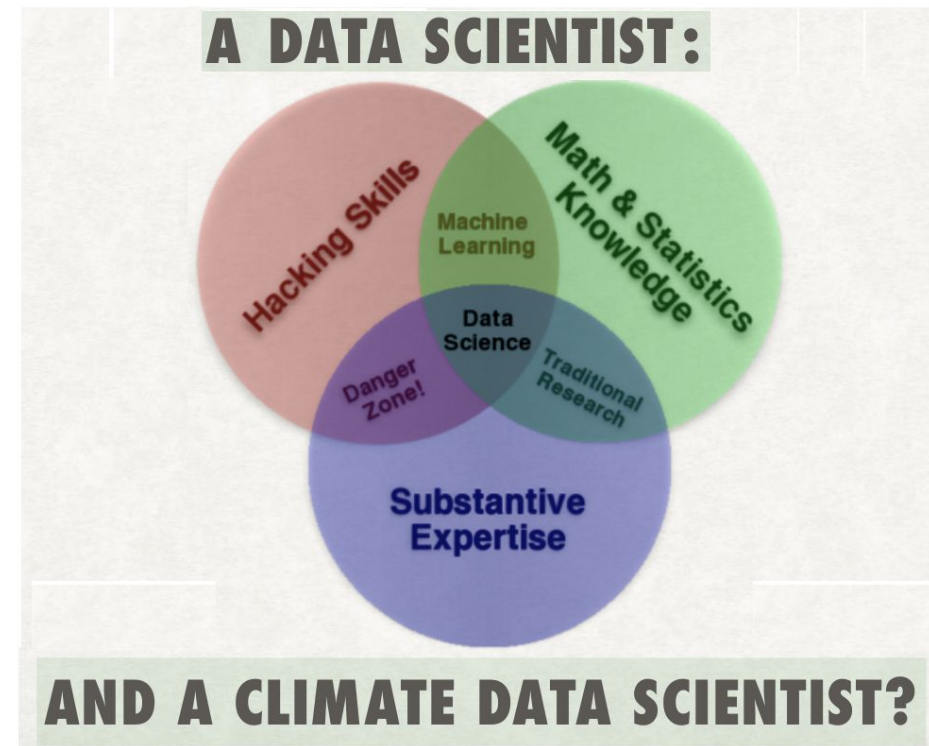
If all you have is a hammer… (Matteo De Felice, ENEA):

- Current data challenges should be **driven by the scientific and technical needs** and not by the tools we want to use to solve them.

- All Big Data problems don't require the Big Data machinery but can be split in multiple parallel small problems.

- Importance of **reproducibility** and tools to do so.

- Issue of open-access **META**data.

- Need for **interdisciplinary profiles** for data scientists.

A DATA SCIENTIST:

Hacking Skills — Math & Statistics Knowledge

Machine Learning

Data Science

Danger Zone!

Traditional Research

Substantive Expertise

AND A CLIMATE DATA SCIENTIST?

RDA — RESEARCH DATA ALLIANCE — Research Data Sharing without barriers

Summary of the discussion

What are the data-related problems specific to our community?

– Volume growth of the data is an issue but the storage is not the main issue, **data movement is more problematic**. Be careful not to store huge volumes of data only because **storage is cheap** and focus on the real usefulness of what we produce.

– Data strategy should be an important criteria in the attribution of the european union funded projects. The participation of data scientists in the evaluation committees should be considered.

– Think at which stage we want to **integrate the metadata**: directly when the data is produced or only after publication?

– Clear need for **automated tools** that can **generate the metadata** but also for an education effort to raise awareness about the importance of metadata.

What are the data-related problems specific to our community?

‒ At which **level** should the **metadata standards** defined? At the community level? At a higher level generating only some simple minimal fields for all the communities? If the solution is a decision at the community level, then this IG could be helpful.

‒ The use of Persistent IDentifiers (**PIDs, DOIs**) for scientific datasets should be encouraged but there is an issue with dynamic or real time datasets.

‒ **Bring the compute to the data** (data close to HPC) is an important point but it also raises the issue to decide who has to pay for the computing and emphasizes, in the case of when public institutions host the HPC, the lack of a clear addressee for the money.

‒ In the context of open data, there are still some issues: **embargo** on data (lag between production and publication), **privacy** for sensible data.

Summary of the discussion

What experiences exist to engage with the most recent developments from computer sciences?

–   If we want to develop a real data strategy for our community, we need to, at least, **formulate clear questions** to be addressed to computer scientists or external companies.

–   Who we should talk to to help us solve our issues. Earth sciences specialists or computer scientists? This IG could be the right place to formulate a **common list of requisites** that could be sent afterwards to different sources of funding.

–   Shouldn't (or couldn't) the **software development be part of the evaluation** criterias when submitting a paper o when asking for funding?

www.bsc.es

Thank you!

For further information please contact
pierre-antoine.bretonniere@bsc.es