Workshop on the creation of a Research Data Alliance Interest Group on weather, climate and air quality

BSC - 11/02/2016

The Earth Sciences community, and especially the weather, climate and air quality sectors, are producing enormous amounts of data at a rate that makes the science behind this data more and more difficult. Several problems arise with this growth: how to handle the volume of data itself, but also how to disseminate, process and document (metadata) the data.
Some of these issues are common to many scientific communities but some of them are very domain-specific and need a very specific answer.
To address these issues, the Earth Sciences Department of the Barcelona Supercomputing Center (BSC-CNS[1]) decided to take advantage of its involvement in the Research Data Alliance (RDA[2]), an organization that tries to solve data sharing issues from a general point of view and for many communities, using the RDA infrastructure and fundings to create an interest group on weather, climate and air quality to gather different actors from these communities and try to address together the common issues that have been identified.

This document summarizes the workshop held at the BSC to launch this interest group.

http://www.bsc.es/about-bsc/press/bsc-in-the-media/bsc-organises-rda-earth-sciences-interest-group-workshop

Presentations by the invited speakers (10h-12h30)

Introduction to RDA (Fabrizio Gagliardi - BSC)

RDA aims at sharing, preserving and distributing scientific data. It was created in 2013 with a plenary event in Goteborg by a US, EU, AUS initiative and now has participants from all over the world and especially in the European Union, Australia, Canada, Japan and South Africa. The European part (RDA Europe) is funded by DG-Connect as a Coordination and Support

[1] https://www.bsc.es/earth-sciences
[2] https://rd-alliance.org/

Action (CSA). Most of the participants come from research institutes (64%), followed by governmental organizations (23%) and IT consultancies and SMEs (5%).
RDA has a complex structure, but important elements are the set of interest and working groups. While the working groups have funding, precise work plans and list of deliverables, the interest groups are unfunded and are mainly discussion fora. An interest group is the first step before the creation of a working group.
When trying to set up one of those structures, the steps to follow are the creation of a network of people interested in the activity to gain momentum and create a partnership with the industry, following a bottom-up approach: from local SMEs to big international companies.
It is important to keep in mind, as objective, that the 9th RDA plenary meeting will be held at the BSC in 2017.

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

End users find difficult accessing climate data, ECOMS[3] is an example of a group having to address this issue.
Based on the THREDDS technology, the User Data Gateway (UDG[4]) provides a homogeneous access end point to collections of relevant variables. It 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. It currently provides access to 23TB of data with a single access point.
The infrastructure also includes 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[5].
UDG comes along with capacity building, training and documentation. The future of the portal is to become fully integrated in ESGF and become more a service provider than a simple data catalogue.

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

JRC-MARS performs on crop growth monitoring and yield forecasting for Europe and beyond for the European Commission. Its members also work on climate-change impacts (including those from extremes) on agriculture and crop production.
The observational meteorological dataset is based on data retrieved in near real-time from around 4000 observations that are interpolated to a regular 25 km grid, whereas the climate extreme studies use ERA-Interim and EURO-CORDEX data. The data are stored locally and both reports and data are available publicly in their data portal.
The typical workflow at JRC includes: the data retrieval, their quality check, their inclusion in a database that is used to run the crop models, and the visualization of the data through an internal tool as well as a data portal that provides netcdf files with different variables[6] and the

[3] http://www.eu-ecoms.eu/
[4] http://www.meteo.unican.es/es/node/73213

[5] https://climate4impact.eu/impactportal/general/index.jsp
[6] http://agri4cast.jrc.ec.europa.eu/DataPortal/Index.aspx?o=d

mars explorer[7]. Other data, like remote sensing, agro-phenology, crop yield, etc. are not that well organised and clearly lack the necessary metadata, which is a recurrent problem in some of the communities.
One of the current issues of the system is the 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.

If all you have is a hammer… (Matteo De Felice, ENEA)
It is important to keep in mind that the current challenges in our communities should be driven by the scientific and technical needs and not by the tools we want to use to solve them. We are often talking about big data problems, but not all require the big data machinery and most of the time could be split in multiple parallel small problems.
Better tools mean less time needed to answer questions, but it implies a larger investment to produce them, which means that the relevance of the question should be assessed first.
Reproducibility is an important issue: all the information in undocumented figures is lost and useless after a short period of time. Possible solutions are github-like tools, plotly[8]-like tools and self-contained scripts (tools that in addition to the final plot, automatically save the metadata, some of the data and the scripts used to generate them).
In any case open access to data is not enough because it doesn't guarantee persistence and proper documentation, so it requires open metadata.
We need to have more multidisciplinary profiles, including data science, IT and scientists in a same profile.

Discussion (14h-17h)

A summary of the main points raised during the discussion follow.

- What are the data-related problems specific to our community?
    - The data management plan and strategy should be an important criteria in the attribution of the european union funded projects. For that, the participation of data scientists in the evaluation committees should be considered.
    - Of course the volume growth of the data is an issue but the storage is not the main issue, it is more the movement of data and their efficient use that is problematic. We should also be careful not to store huge volumes of data only because storage is cheap and focus on the real usefulness of what we produce. Thinking in terms of ratio between cost of reproduction of the data over storage and dissemination must be an idea to keep in mind when defining the data management plans.
    - The metadata is important but we have to think at which stage we want to integrate them: do we need them directly when the data is produced or only after publication? Then, how should we deal with dynamic datasets? There is a clear need for automated tools that can generate the metadata but also for an education effort to raise awareness about the importance of metadata.

[7] http://agri4cast.jrc.ec.europa.eu/mars-explorer/
[8] https://plot.ly/

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.
- What experiences exist to engage with the most recent developments from computer sciences?
    - This is a really important question. 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. We have here the issue of 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. This leads again to one of the previous questions about how to evaluate the projects and the scientific work in general: should or could the software development be part of the evaluation criterias when submitting a paper o when asking for funding?
- What sources of funding could be explored?
    - Even if it can be difficult for a sector that doesn't directly generate money to get funding, we mustn't underestimate the momentum, both at scientific, societal and political level that our communities can generate (example of the last COP21 in Paris). Therefore, being conscious ourselves of this "power" should help us convincing the eventual sources of funding to support us.

Conclusions

There seems to be a clear interest from the participants to this workshop in creating an Interest Group in the framework of the Research Data Alliance. We can also involve in the process a broader audience that, even if couldn't be physically present at the workshop, showed a strong interest when we first contacted them.
To be able to engage with industries and public funding, we should define very precise objectives for this Interest Group.

Some first solutions envisaged as targets for the Interest Group are:

- the formulation of a precise list of requirement from the scientific communities in our domains of weather, climate and air quality in terms of computational requirements and advanced data management solutions to be provided to external entities.
- raising awareness about the importance of metadata and how to define and create them in our particular community

- pushing for the integration of data management and software development in the evaluation of scientific projects and work.

Next steps/ongoing work

In all the outcomes from this first workshop and the work that will follow, we must keep as objective the RDA Plenary session organized in the Barcelona Supercomputing Center where most of the European and worldwide participants of RDA, both from industrial and academic sectors will be present. For this event, we should have a clear roadmap and list of requirements for this Interest Group.

The work of the different RDA Working and Interest Groups will have to be carefully analysed to avoid redundant work in our “Weather, Climate and Air Quality Interest Group”.

Meanwhile, a mailing list (earthsciences-rda-ig@bsc.es) have been created to discuss the different issues and follow on with the work.

We will also consider a consultation of the industry to get input about their needs and requirements.

Finally, a Birds of a Feather about the interest group will take place at the next RDA Plenary Meeting (Tokyo, Japan, 1-3 March 2016[9]). The conclusions will be circulated among the participants of this workshop.

[9] https://rd-alliance.org/bof-weather-climate-and-air-quality-ig.html