JOAQUÍN BEDIA
PREDICTIA INTELLIGENT DATA SOLUTIONS

R&D MANAGER

WHAT IS THE ROLE OF RESEARCH DATA WITHIN YOUR COMPANY'S ACTIVITY?

Predictia is a spin-off from a research group that specializes in climate science, data mining and high performance computing, at the University of Cantabria, Spain. We develop user-tailored data management and data mining solutions for specific sectors. Initially, we worked with problems related to climate and health. We offer products such as seasonal and short-term weather forecasts, for example. Predictia has been involved in European projects like EUPORIAS and PRIMAVERA, other European initiatives like the VALUE Action Cost as well as in different Lots within the the Copernicus Climate Change Programme (C3S), like the currently on-going Project QA4Seas, aimed at the quality assurance of seasonal forecast products. As a result of our active involvement in climate data exploitation and production, we have developed a metadata framework (we have coined the brand METACLIP, METAdata for CLImate Products, to refer to this framework), building upon RDF standards. Our aim is to develop a semantic data provenance model that aligns with international standards to provide reproducible climate data products with provenance information that can be displayed in a user-friendly way by users with varying levels of expertise. We are using W3C standards and some existing RDF models, and extending other widely applied models to adapt to the specific needs of the climate data users community.

WHILE WORKING WITH THESE DIFFERENT STANDARDS AND MODELS, WHAT ARE THE MOST RECURRING PROBLEMS THAT YOU ENCOUNTER?

We work with several different datasets and, even though there are common standards and patterns that many of them apply, there are also differences. Data is scattered in many sources and not all of them are open and available for reuse. In Europe, for example, there are many datasets that are not open, while in the US there are more open datasets available (NASA, NOAA... for instance). We believe that standards must be global, not only European.

Moreover, data traceability is a very relevant requirement for our climate-services community, both in the public/academy and private sectors. It is indispensable to trust the data we are using. That is the reason why we are focusing on data provenance.

NOW THAT YOU MENTION DATA PROVENANCE, WHICH WAS THE SUBJECT OF THE RDA EUROPE WORKSHOP ORGANIZED BY THE WEATHER, CLIMATE AND AIR QUALITY INTEREST GROUP, DOES YOUR COMPANY MAKE USE OF RDA RECOMMENDATIONS?

We haven't been related to RDA before participating in this workshop and we were not familiar with the recommendations produced by them. It was useful for us to attend and learn what RDA does, especially the work that the Provenance Patterns Working Group is carrying out regarding the development of the Provenance Patterns Database (PPDB) that was presented during the workshop. It has been very rewarding to see how our idea of METACLIP is very well aligned with the RDA recommendations, and also to identify those aspects that would require modification or further work for a better integration with the international standards and good practices promoted by the RDA. It would be good if RDA could organize more workshops like this one. Companies like Predictia would be interested in participating in such technical meetings, getting together with data scientists and researchers, and sharing our experiences.

AND IN THIS LINE, WHAT OPPORTUNITIES FOR ACADEMIA AND THE INDUSTRY TO COLLABORATE IN THEIR MUTUAL BENEFIT CAN YOU IDENTIFY?

Technical workshops, like the one on Data Provenance approaches, would be very useful and, certainly, they would foster collaboration between academia and industry. There are many European projects in which partners come from both the public and the private sector, so we do have common problems to face and requirements to fulfil. It would be interesting if C3S (The Copernicus Climate Change Service) could be involved with these kind of RDA activities. This could attract more participants from both sectors.

We think that finding ways to "take" academic contents to the industry is key. There are many small companies working with research data, open data. And there are high expectations related to the development of business models that could build upon open research data. Keeping data closed is a drawback. In European projects, openness is starting to become increasingly important. It could end up being a condition for funding, for instance.

That is why it is so important to advertise recommendations, best practices and success stories with open data. There is a lot of work embedded in RDA recommendations and it would be so useful for companies to know about them. In my view, there is also room for making these recommendations more accessible to non-experts, helping to further spread the word on the need for traceable data and how to achieve it.

ARE THERE ANY SUCCESS STORIES THAT YOU WOULD LIKE HIGHLIGHT AS EXAMPLES?

I would like to mention the project climate4R developed by the Santander Meteorology Group, in which I have been involved in the last years. It is a bundle of R packages to access, process and visualize climate data. It is also compatible with other packages for impact studies. All components are distributed under a GNU general public license and it is still

available, with a variety of users, including private companies. Some applied case studies using this set of tools are already illustrated in several journal papers, whose outcomes (data and paper maps/plots) are fully reproducible. I think this kind of tools greatly contributes to open science and research reproducibility. Currently, one of our main goals is adapting the METACLIP approach to the climate4R framework in order to enable a full provenance description of the climate data products generated.