CODATA-RDA Research Data Science Summer Schools

A Collaborative Foundational Research Data Capacity Building Initiative for 21st Century Researchers

Introduction

Contemporary research – particularly when addressing the most significant, transdisciplinary research challenges – cannot effectively be done without a range of skills relating to data. This includes the principles and practice of Open Science and research data management, curation and annotation, the use of a range of data platforms and infrastructures, analysis, modelling and statistical techniques to make sense of the data, visualisation and software development. The ensemble of these skills, grouped into the strands of Research Data Management, Open Science and Data Science, we define as 'Research Data Science'.

The CODATA-RDA Research Data Science schools aim to address the critical need for this set of skills by pulling together existing and emerging initiatives and resources into a sustainably run, globally applicable foundational programme in Research Data Science, particularly for Low and Middle Income Countries (LMICs).

The First CODATA-RDA Research Data Science Summer School

The first summer school was held at ICTP, Trieste in August 2016. Comprehensive information about the programme, participants, funders, and more can be found at http://indico.ictp.it/event/7658/. The event was labelled in social media as #DataTrieste. Photos and other media is available at https://www.flickr.com/photos/145909074@N04/albums/with/72157671368407182.

Organising Committee

- Andrew Harrison (University of Essex/RDA)
- Simon Hodson (CODATA ICSU Committee on Data for Science and Technology)
- Hugh Shanahan (Royal Holloway, University of London/RDA)
- Romain Murenzi (UNESCO)
- Anelda van der Walt (Talarify and North-West University, South Africa)
- Ivan Girotto (ICTP)
- Clement Onime (ICTP)

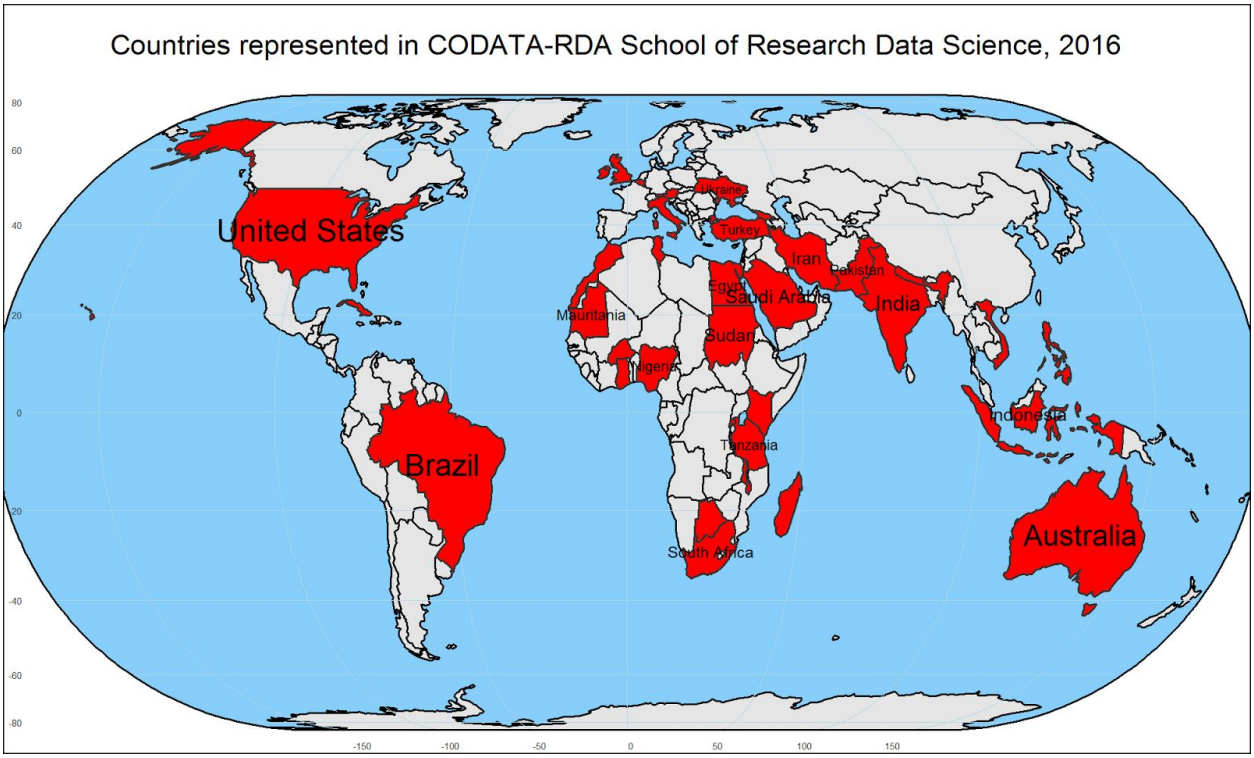
Funding

The event was funded by a number of international sponsors including

- TWAS
- GEO
- ICTP
- RDA-US
- CODATA
- ACU
- GODAN
- RDA-Europe

Participants

Over 320 applications were received from countries around the world. Over 70 learners from 35 countries were selected to participate in the foundational school. Researchers and postgraduate students attended from diverse disciplines such as genomics, social sciences, proteomics, astronomy, ecology, computer science and more. Participants represented all career stages from early postgraduate students and early career researchers to a small number of established academics. Established researchers would later play a role in replicating the school and related events in their home countries. Participants and instructors were selected to maintain as even a gender balance as possible.



Source: http://codata.org/blog/wp-content/uploads/2016/08/Countries_Represented.png

Instructors

A total of 16 instructors from 12 institutions in the UK, the USA, Italy, and the Netherlands were supported by three helpers during the two-week course.

The Programme

The Research Data Science Summer School curriculum aimed to provide participants an introduction to research computing and relevant skills upon which discipline specific courses can build. Limited prior exposure to computing was assumed and to this effect the school started with an introduction to Linux and the command line. Subsequent modules used skills learned at earlier stages during the school to ensure new knowledge were reinforced.

The curriculum was composed of the following:

- Open Science - challenges and opportunities
- Research Data Management
- Author Carpentry - best practices for 21st century scientific publishing
- Automation of repetitive tasks in Linux Bash
- Data analysis in R
- Principles of visual data analysis
- Data visualisation in R
- Databases and SQL
- Version control and collaboration with git and GitHub
- Machine learning and recommender systems in R
- Neural networks
- Research computing

The largest part of the school were run as hands-on computer sessions to allow participants plenty of opportunity to practice new skills. Learners were able to either use their own laptops or lab computers with software pre-installed. A collaborative culture was encouraged in the computer labs. Participants supported each other in terms of keeping up with the exercises and peer learning played an important role in understanding foundational concepts.

Peer learning plays an important role in the comprehension and consolidation of important concepts and skills.

A supportive learning environment is created by instructors and helpers who actively ensures that participants are not left behind during practical coding exercises.

Lively discussions took place throughout the school. Participants enjoyed learning about challenges and opportunities faced by colleagues in different regions and countries.

Several of the sessions involved in-depth discussions around emerging key issues around Open Science and collaboration. Participants were encouraged to discuss challenges and opportunities in their immediate environments but also those pertaining to their countries, broader regions and global research.

Feedback from participants

Feedback was sought often throughout the two-week school to ensure speedy remediation of inefficiencies that could be solved in the short term. Participants also reviewed each module separately and had the opportunity to provide longer feedback about the full school.

In general some of the highlights pointed out by learners were:

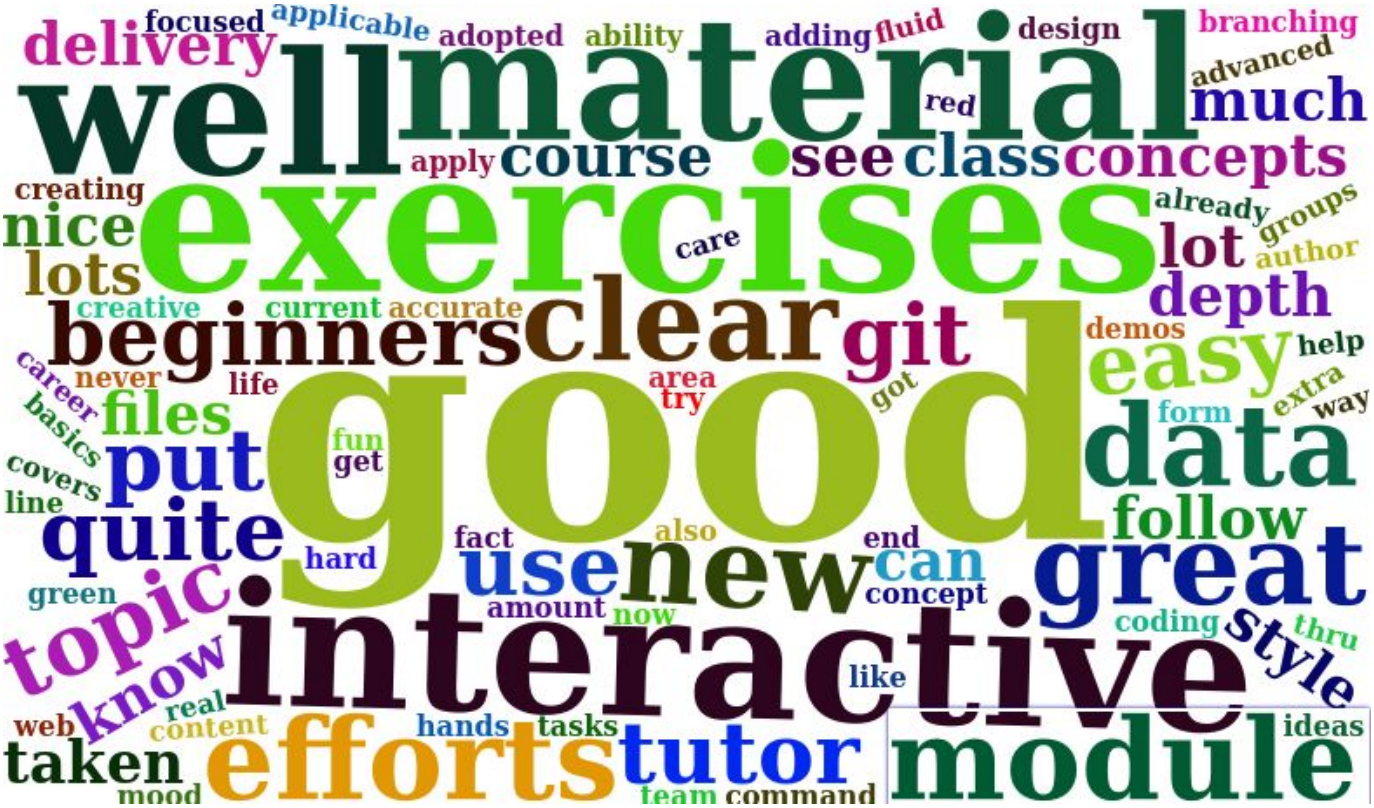
- the curriculum design really addressed the critical issues;
- instructors were very helpful and well-prepared;
- organisation was done very well;
- meeting such a diverse group of people and having lots of opportunities to engage throughout the two weeks was very beneficial;

"*It was an amazing work by all the organizers and I really fortunate that I got an opportunity to be a part. Regarding the lectures, it was highly informative although I t would definitely take time to implement all those and it's a fact that we would never become the expert in few hours only. But the best part was that I was able to cross the threshold so that in near future whenever I will encounter all the new things that I had been taught here would not appear as stranger for me as well I can for sure take help from the faculties as well. the effort made by all to put us together under an umbrella is incomparable.*"

"*Very good lecturer and material. Useful exercises and examples.*"

"*Learning skills that are very useful for researchers in order to present the data effectively. I like so much the examples he made and the references to computer coding, softwares, and internet web pages useful for visual data design.*"

"*I think that a good balance between amount of information and rhythm for the people that never use R before and for the people that already use it.*"



A wordcloud summarising feedback from learners given on all modules taught during the inaugaral school.

Outcomes since August 2016

Several participants and colleagues have reported back on their experiences during #DataTrieste since the event (this is an incomplete list so far!).

- https://sciencenode.org/feature/the-open-science-grid-goes-to-school.php
- http://codata.org/blog/2016/10/12/report-to-ictp-trieste-codata-rda-research-data-science-summer-school-1st-12th-august-2016/
- https://www.opensciencegrid.org/osgs-rob-quick-and-kyle-gross-share-expertise-in-rwanda-and-italy/
- http://codata.org/blog/2016/08/21/map-showing-distribution-of-codata-workshop/
- https://thepursuitofweirdness.blogspot.co.za/2016/08/codata-rda-school-of-research-data.html
- https://itconnections.iu.edu/2016-september/quick-africa.php

A session on building research data science capacity was held at SciDataCon 2016 where several of the instructors and organisers presented and participated (http://www.scidatacon.org/2016/sessions/57/).

New collaborations have been established since the school:

- A collaboration between University of Botswana and North-West University who will be running a smaller workshop together in Botswana in 2017;
- A regional climate modeling project has been entered into by ICTP researchers and faculty of Indiana University, this collaboration will couple oceanographic, atmospheric, and earth systems models to more accurately predict ocean conditions. Proposals for ongoing funding are being sought by both key institutions through national and international funding agencies.
- One of the instructors from #DataTrieste was subsequently invited to teach an R workshop in South Africa (https://arreyves.github.io/2016-10-31-univenda/)

A Task Group has been founded by CODATA to support these activities and to extend the programme.

Forthcoming activities: 2017

A CODATA Task Group was established to further develop the school curriculum and sustainability. The task group will work closely with the RDA and other organisations that were previously involved in the joint working group that brought the schools to live.

The CODATA-RDA Research Data Science Summer School will run again from 10-21 July 2017 at ICTP, Trieste as well as in Sao Paulo in December 2017. Discussions are also underway to run a similar school in South Africa towards the end of 2017 and in the USA and in Canada.

Next Steps: Building a Scalable Network of Data Schools

The first school in Trieste was a great success and has already had considerable influence. It has demonstrated the benefits of foundational training in core data skills. Our vision is to achieve great impact by taking steps to scale this approach. We will use open approaches and a network effect to achieve this.

Based on the success of the first RDS Summer School in 2016, the confirmed schools for 2017, and the large number of requests from participants and others to replicate the school at their own institutions or in their own countries, the project is entering the next phase where longer term funding is sought to further develop the schools into a transferable, scalable and sustainably run initiative. We intend to continue the Trieste school as a hub, run on an annual basis, to showcase the approach and to engage future instructors and organisers. Additionally, we will work with partners to better enable schools taking this approach, reusing the materials etc. to be run in a variety of locations globally.

Organisers

Partners

Sponsors