Working Group proposal :-

Summer schools in Data Science in low and middle income countries

Working Group Charter

The goal of the WG is to set up the framework to run a series of Summer Schools in Data Science and data sharing in low and middle income countries (LMICs). The series will run at a variety of locations but will have a base in a specific LMIC and the UK. This series would be endorsed by both the RDA and CODATA and would make use of the wide variety of expertise within these organisations.

In the 18 month period of the WG it will

* Determine the best curriculum for the school in collaboration with others.
* Organise partnerships with LMIC institutions.
* Arrange how the materials can also be delivered online.

Value Proposition

Research in LMICs is hampered by a variety of infrastructural issues. Access to research data in the public domain and the computational resources to analyse them, would give researchers in LMICs the chance to do world-class research that is relevant to their society. In order for this to happen, researchers need to understand the techniques of Data Science and the importance of the principles of data sharing.

Students for the school will be either postgraduate students or lecturing staff at universities and will be selected by the hosting institutions. Such students will be in an ideal position to implement their own research programmes and will disseminate what they’ve learnt to a larger cohort. In this way over a comparatively short period of time the level of understanding of this field can be built up and impact research. The creation of a cadre of individuals who can analyse, maintain and curate data sets will be an important skill which will also positively impact the local society as enterprises based on data analysis will expand.

Engagement with existing work in the area:

There exist a number of initiatives that provide training in these fields; for example there are a number of Summer Schools in Europe on cloud computing (see attached document of links). Microsoft Research provide training sessions on their cloud Azure. There are now two MOOCs in Data Science and the number of master’s programmes in this area is expanding rapidly. Analysis platforms such as R have an ever-expanding set of teaching resources. The Software Carpentry movement, which has many of the correct pedagogical approaches for this school has events across the developed world (and has had sessions in South Africa). AidData have provided training in the analysis of social data in the developing world. The Wellcome Trust has provided funding for a successful series of schools in Bioinformatics in the Developing World. The Accelerated Data Programme of the International Household Panel Survey, and the IHPS in general provide tools and training to many developing country institutions, to increase the availability and quality of survey data.

CODATA has longstanding activities to build capacity in countries with developing and emerging economies. The most recent include:

* The PASTD Task Group (Preservation of and Access to Scientific and Technical Data in/for/with Developing Countries) has run a series of workshops in developing countries. [The last was in Nairobi, Kenya in August 2014.](http://www.codata.org/news/9/62/Kenya-Cabinet-Secretary-Supports-the-International-Workshop-on-Open-Data-for-Science-and-Sustainability-in-Developing-Countries) Target countries for similar workshops over the next two years are likely to be Malaysia or Indonesia and Brazil.
* CODATA China, with support from the Chinese Academy of Science, ran a training workshop in [Big Data for Science for Researchers from Countries with Developing and Emerging Economies](http://codata.org/blog/2014/03/26/codata-international-training-workshop-in-big-data-for-science-for-researchers-from-emerging-and-developing-countries/).
* [SciDataCon 2014, International Conference on Data Sharing and Integration for Global Sustainability](http://www.scidatacon2014.org/), is taking place in New Delhi, India, in November 2014 and features a number of sessions and workshops on related issues.

We have already made contacts with a variety of the above different organisations. CODATA is very happy to participate and to be a co-sponsor of the group. AidData has expressed an interest in the proposal. Within the RDA there is the new IG in Education and training on handling of research data.. This new IG can provide guidance on what the curriculum should be. The Wellcome Trust has expressed a willingness to provide guidance based on their experience.

Adoption Plan:

Final Deliverables :-

1. Agreements from relevant partners to provide technical support for online materials. This will involve a breakdown of what the online materials will provide, the level of support that needs to be provided to develop and maintain these resources.
2. Agreements from institutions willing to run the school. Each agreement will be an MoU that the institution will host and provide the logistical resources to make the school happen. As outlined in the timeline below, this will be completed between the 5th and 6th RDA plenaries. On the basis of the first two agreements the RDA council and CODATA will be called upon to endorse the proposal.
3. The outline for a typical Summer School. This will include a breakdown of the curriculum, methods of assessment, requirements in terms of technical resources and the organization of lectures and labs. The IG on Education and Training on handling of research data and other parties listed above, with whom this WG will liaise with, will be able to provide a detailed breakdown of what that curriculum should be. As it is, each school will be based on the distinct themes that have been observed in Data Science, e.g. computing resources, analysis and data sharing which themselves are composed of different sub-themes; for example the data sharing theme would have data storage and annotation sub-theme and an accessibility (making own data available and accessing outside data) sub-theme; the analysis theme would have a statistical/machine learning sub theme and a visualization sub theme and so on. Each school would focus on a specific theme but also give an introduction to the other themes. Each school would, in addition, also have a specific domain where this is applied to give a context to the topic and hence demonstrate to the students how this could be applied, for example describing appropriate data sets for the discipline and how to access them. As a result, graduates of the school would have an understanding of the bigger picture that Data Science provides while being able to apply them readily to their area of research. In addition to this, other options will be considered to improve the sustainability of the initiative, in particular joint PhD programmes for students who have taken part in the school and the possibility of accreditation for individual who specialises in the field on a longer term basis.

The timeline for this is as follows :-



Intermediate documents.

The documents for this WG are as follows:

1 Initial document to RDA TAB on organization and running of WG

2 Template agreement with online delivery partners.

3 Template MoU with hosting institution.

4 Draft endorsements from RDA and CODATA.

5 Draft proposal.

6 Curriculum for schools, including learning outcomes and forms of assessment.

7 Detailed summary of requirements from hosting institution.

Operation

An initial document on the WG will be sent to the RDA TAB on the WG’s organization and basic running. An email list of the WG members will be set up and all emails will be based on this. WG meetings will be based on monthly conference calls, using the RDA service provided, which will review progress using the above Gantt diagram. Significant disagreements between members will be dealt with by the co-chairs. The co-chairs will meet and discuss on a more regular (fortnightly) basis. In order to achieve consensus or resolve conflicts, the WG will consult with the RDA TAB and the IG on Education and Training on handling of research data.

The WG will reach out to the parties mentioned above and liaise with a number of discipline specific IG’s and WG’s within the RDA including the Wheat Data Interoperability WG, Agricultural Data Interoperability IG, Big Data Analytics IG, Biodiversity Data Integration IG, Digital Practices in History and Ethnography IG, Geospatial IG, Marine Data Harmonisation IG, RDA/CODATA Materials IG, Infrastructure and Interoperability IG, Research data needs of the Photon and Neutron Community IG, Structural Biology IG, Data for development IG and Toxicogenomics Interoperability IG by contacting directly their chairs and asking for the opinion and feedback on the WG’s goals.

As the school goes forward the organizers will organize a panel to advise on how to update the teaching material on an annual basis on this rapidly evolving topic.

Initial Membership: A specific list of initial members of the WG and a description of initial leadership of the WG.

Hugh Shanahan, Royal Holloway, University of London (co-Chair)

Andrew Harrison, University of Essex (co-Chair)

Simon Hodson, Executive Director, CODATA (co-Chair)

[Amye Kenall](https://rd-alliance.org/user/1131), Biomed Central

Arturo Martinez, University of Queensland

[Avgoustinos Constantinides](https://rd-alliance.org/user/2270), IBSAC Intelligent Business Solutions

Cheng Liu, University of Notre Dame

Christine Kirkpatrick, UC San Diego

Esther Dzale Yeumo, INRA

[Godwin Yeboah](https://rd-alliance.org/user/3222), Aberdeen University

Hervé L'Hours, UK Data Archive

Ingvill Mochman, GESIS-Leibniz Institute for the Social Sciences and Cologne Business School

Jerome Pansanel, CNRS, Strasbourg

Jonathan Tedds, University of Leicester

Miguel-Angel Sicilia, University of Alcalá

[Rainer Stotzka](https://rd-alliance.org/user/378), Karlsruhe Institute of Technology

[Siddeswara Guru](https://rd-alliance.org/user/1292), University of Queensland

[Suchith Anand](https://rd-alliance.org/user/2444), University of Nottingham

Vincent Breton, CNRS, Clermont-Ferrand

Yuri Demechenko, University of Amsterdam

[Yuri Malitsky](https://rd-alliance.org/user/2440), University College Cork

Zhijie Zhang, Fudan University

Ziad El Bitar, National Centre for Scientific Research, Strasbourg