A curriculum for foundational Research Data Science skills for Early Career Researchers

DOI: 10.15497/rda00038

Authors: Hugh Shanahan, Rob Quick, Marcela Alfaro Córdoba, Gail Clement, Louise Bezuidenhout, Venkat Shanmugasundaram, Sarah Jones, Kevin Ashley, Stephen Diggs, Colin Gillespie, Sara El Jadid, Maria Sorokina, Roger Barlow, Ekpe Okorafor, Gergely Sipos, Alessandro Constantini, Hannah Short.

Published: October 10th, 2019

Abstract: This recommendation describes the curriculum and example materials to give Early Career Researchers (ECR's) the foundational skills in Data Science to work with their data. This curriculum combines technical skills, such as Software Carpentry with responsible research practices such as Open and Responsible Research. This curriculum is composed of

- .) a set of curriculum specifications for the modules run in this curriculum,
- .) an example timetable for a 10 day intensive training event,
- .) a diagram to show how these modules are connected,
- .) a spreadsheet of links to example materials that implements this,
- .) metadata for the submission,
- .) an impact statement,
- .) a document discussing the maintenance plan of the materials.

The purpose of this curriculum is to be deliberately broad and shallow to be delivered over approximately 70 contact hours. It could also form the basis for a much deeper programme which will not be explored further.

In 2016 we ran one school in Trieste, Italy. In 2017 we ran two, Trieste and São Paulo, Brazil. In 2018 we ran three, Trieste, São Paulo and Kigali, Rwanda. In 2019 we will run four schools (Addis Ababa, Ethiopia, Trieste, Abuja, Nigeria and San José in Costa Rica). At the end of this year approximately 400 students will have been taught on four continents using the curriculum developed here.

Note: The submission package is available at http://doi.org/10.5281/zenodo.3478590.

Keywords: RDA Recommendation; CODATA; Data Science; Early Career Researchers.

Language: English

Citation and Download: Shanahan, H., Quick, R., Córdoba, M. A., Clement, G., Bezuidenhout, L., Shanmugasundaram, V. Short, H. (2019). A curriculum for foundational Research Data Science skills for Early Career Researchers (Version 1). DOI: https://doi.org/10.15497/rda00038

CODATA-RDA Research Data Science Summer Schools

Curriculum Specifications

Open and Responsible Research

Aims:

To consider the importance of open and responsible research

Learning Outcomes:

At the end of this course a student will

- Understand responsible conduct of research as it pertains to data science
- Have a broad understanding of the Open Science movement
- have reflected on the impact Open Science on their own research and future career.

Course content:

- Introduction to “open and responsible (data) science citizenship”
    - Responsible conduct of research intro
    - Open Science intro
- Being open and responsible at home
    - Challenges to open and responsible science group work and discussion
- Understanding open and responsible research in the “big picture”
    - Introduction to societal impact of data
    - Introduction to “infraethics”

How long: 3 hours

Research Data Management

Aims:

To have an understanding of the principles of research data management (RDM) and the impact of Openness and Sharing in Research

Learning Outcomes:

At the end of this course a student will

- Understand the data curation lifecycle
- Appreciate the practical advantages of good RDM and open research/science
- How to add value and longevity to your data
- Understand the principles and importance of standardisation
- How to publish data

Course content:

- Incentives for curation
- The data curation life-cycle
- FAIR principles
- Open vs FAIR
- File formats
- Metadata
- Ontologies
- Licenses
- Repositories
- Persistent identifiers (PIDs)
- Data management plans (DMPs)

How long: 4 ½ hours

Software Carpentry

Aims:

To have an introductory understanding of programming and software engineering skills to manipulate data and analyse data in reproducible fashion.

Learning Outcomes:

At the end of this course a student will

- have an introductory understanding of the Unix shell,
- be able to execute simple commands in R,
- be able to use Git.

Course content:

- Introduction to the Unix shell.
- File concepts in Unix.
- Combining Unix commands, pipes and filters.
- Shell scripts.
- Functions in R.
- Conditionals in R.
- Command line R programs.
- Best practices in R.
- Setting up Git.
- Tracking changes in Git.
- Collaboration and Open Science with Git.

How long: 2 ½ days

Analysis

Aims:

To have an understanding of the principles necessary to analyse data in terms of being able to make decisions from large amounts of data and applying machine learning techniques.

Learning Outcomes:

At the end of this course a student will

- understand the basic principles of machine learning,
- apply pipelines to build recommender systems,
- understand how to use Artificial Neural Networks, with hands-on experience,
- understand the principles of Boosted Decision Trees and SUpport Vector Machines.

Course content:

- Machine learning concepts,
- Recommender systems,
- Artificial Neural Networks,
- Other machine learning methods.

How long: 2 ½ days

Visualisation

Aims:

To have an understanding of the principles of visualising data.

Learning Outcomes:

At the end of this course a student will

- understand how to use R to perform visualisation,
- be able to perform a critical assessment of effective visualisation techniques.

Course content:

- Data wrangling.
- Visualisation packages in R (such as ggplot2).
- [Optional] Visualisation in Python.
- Workshop based approaches to critical assessment of visualisation.

How long: 2 days

Computational Infrastructures

Aims:

To introduce students to open computational infrastructures available to them when analysis tasks outgrow their local computational resources.

Learning Outcomes:

At the end of this course a student will

- understand the basic concepts of HTC, HPC and Cloud computing,
- be able to execute a distributed computing job
- be able to use more advanced features such as batch schedulers or containers.
- Be able to interact with Cloud services

Course content:

- Introduction to cloud computing concepts such as IaaS and PaaS and SaaS and their aspects.
- Secure authentication mechanisms
- Deploying scripts.
- Interacting with mass storage repositories.
- Use of batch schedulers of containers.
- Adopt cloud-based environment and services

How long: 2 days

Author Carpentry

Aims:

To have an understanding of authorship in the 21st century.

Learning Outcomes:

At the end of this course a student will have

- Created an ORCiD for themselves,
- Understood the concept of reproducible reporting
- Generating a DOI for a report and depositing it into a repository
- Understood Copyright and Data Licensing

Course content:

- Introduction to ORCiD’s
- Reproducible reporting using (for example) Rstudio
- Generating DOI’s and depositing reports
- Copyright and Data Licencing

How long: 4 ½ hours

Information Security

Aims:

To have an understanding of the importance of Information Security in an Open era .

Learning Outcomes:

At the end of this course a student will have

- Understood that their online activity, and any systems they create or use for Data Science, will be subject to online attack

- Understood security design principles that they can apply to their work
- Understood the basics of cryptography and encryption, and their importance

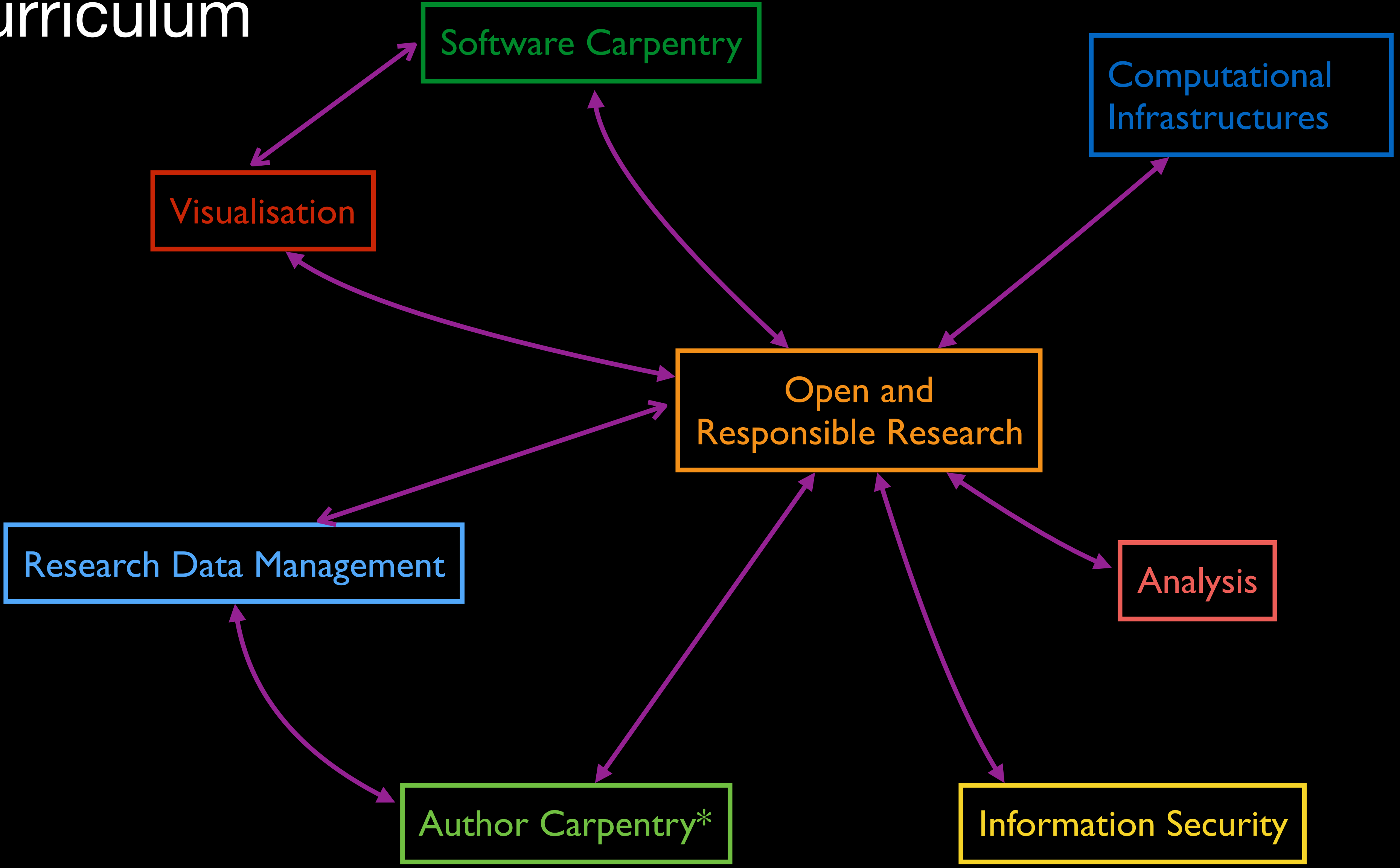
Course content:

- Introduction to Computer Security
- Practical Security and Cryptography
- Evening session; cracking ciphers OR ethical discussion

How long: 3 hours

For further questions on the curriculum, please contact
Dr. Hugh Shanahan,
Department of Computer Science,
Royal Holloway, University of London
Hugh.Shanahan@rhul.ac.uk

Curriculum



* Not usually run in regional schools

Summary	Start : 08:30	11:00; 16:00-	13:00-14:00
Open and Responsible Research		Day 1 am	8:30 - 10:30
Author Carpentry		Day 1 am	10:30-13:00
Software Carpentry (Unix Command Line)		Day 1 pm	14:00-17:45
Software Carpentry - Ethics exercise		Day 1 pm	17:45-18:00
Software Carpentry (Git)		Day 2 am	8:30 - 12:45
Software Carpentry - Ethics exercise		Day 2 am	12:45 - 13:00
Software Carpentry (R)		Day 2 pm	14:00 - 18:00
Software Carpentry (R)		Day 3 am	8:30 - 13:00
Software Carpentry (R)		Day 3 pm	14:00 - 17:45
Software Carpentry - Ethics exercise		Day 3 pm	17:45-18:00
Research Data Management		Day 4 am	08:30-13:30
Author Carpentry		Day 4 pm	14h30 - 18:00
Author Carpentry		Day 5 am	8:30 - 10:00
Open and Responsible Research		Day 5 am	10:30 - 12:00
Research Data Management		Day 5 pm	12:00 - 16:00
Days off		Day 6,7 am/pm	NA
Visualisation		Day 8 am/pm	8:30 - 16:00
Visualisation - Ethics Exercise			15:45 - 16:00
Information Security		Day 8 pm	16:30 - 18:15
Information Security - Ethics Exercise		Day 8 pm	18:15-18:30
Analysis		Day 9 am	8:30 - 18:00
Analysis		Day 10 am/pm	8:30 - 17:45
Analysis - Ethics Exercise			17:45 - 18:00
Analysis		Day 11 am	8:30 - 11:00
Computational Infrastructures		Day 11 am/pm	11:00 - 18:00
Computational Infrastructures		Day 12 am	8:30 - 12:45
Computational Infrastructures - Ethics Exercise			12:45 - 13:00

Class	Links
Open and Responsible Research	
Theory	https://doi.org/10.5281/zenodo.3579092
Ethics feedback and exercises	https://doi.org/10.5281/zenodo.3579099
Research Data Management	
RDM	https://doi.org/10.5281/zenodo.3579130
Software Carpentry	
Shell	https://doi.org/10.5281/zenodo.3266823
git	https://doi.org/10.5281/zenodo.57467
R	https://doi.org/10.5281/zenodo.57520
Analysis	
Machine Learning	https://doi.org/10.5281/zenodo.3579166
Artificial Neural Networks	https://doi.org/10.5281/zenodo.3579219
Visualisation	
Theory	https://doi.org/10.5281/zenodo.3579230
Practice	https://doi.org/10.5281/zenodo.3579245
Computational Infrastructures	
Version 2019	https://doi.org/10.5281/zenodo.3579437
Alternative	https://opensciencegrid.org/dosar/
Author Carpentry	
All materials	https://doi.org/10.7907/Z96H4FFZ
Information Security	
All materials	https://doi.org/10.5281/zenodo.3360480

Class	Links
Open and Responsible Research	
Theory	https://drive.google.com/drive/folders/1oH5DAh1QKdCz2fIvu_SZXqHwzK9b2JyP
Ethics feedback and exercises	https://drive.google.com/drive/folders/1gBRub-dnRLljnX3KYeHuxd8BuUNVLiKJ
Research Data Management	
	https://drive.google.com/drive/folders/10moHd2tevP8nKaocAPU7VQwleESSnfr3
Software Carpentry	
Shell	http://swcarpentry.github.io/shell-novice/
git	https://swcarpentry.github.io/git-novice/
R	ttps://github.com/marioa/trieste
Analysis	
Machine Learning	https://drive.google.com/drive/folders/1rBH-PRfHf-TxKZgdwMkV8b1Eq9yMT2x8
Artificial Neural Networks	https://drive.google.com/drive/folders/130LWVhfB3OdqK3EaozuIdbFYvORfLXWq
Visualisation	
Theory	https://docs.google.com/presentation/d/1rtlY9TTIuwhS8zBez9DDtvlmL0GADkvAjm3-rNLJvGk/edit?usp=sharing
Practice	https://drive.google.com/drive/folders/1Lsi9DK4uEB0QaP2GOtJ4GlEPwgL2ssR3
Computational Infrastructures	
Version 2019	https://drive.google.com/drive/u/0/folders/1e_V4gnEzk0zXsr2vrAjN_aw0aPElvubh
Alternative	https://opensciencegrid.org/dosar/
Author Carpentry	
All materials	https://authorcarpentry.github.io/orcid-profile/
Information Security	
All materials	https://zenodo.org/record/3360480#.XUkiIy2Q0UE