Outline: Data Skills Curricula Framework

(endorsed November 2017)

CORE

i.*Programming for data intensive research* (for those who already code) – 4 DAYS

- Unix/Linux, shells, syntax and using the command line
- Coding standards & style guides e.g. PEP 8
- Version control using GIT, individual and team
- Commenting and documenting code
- Modularising code
- Domain specific data formats and libraries
- Debugging and basic error handling
- Introduction to testing and design
- Open source licensing and code
- Importance of metadata
- Optional mention of parallelising code
- Programming language agnostic, but most likely taught in Python
- Introduction to Notebooks, e.g. Jupyter, for sharing code and documents

ii.*Environmental data: expectations and limitations* – 3 DAYS

- Uncertainties in environmental measurements
- Instrumentation examples, calibration, precision
- Spatial data - issues for gridded data and projections
- Cleaning data and dealing effectively with missing or corrupt data
- Combining datasets from multiple sources e.g. points in time and space vs averages in time and space
- Using representative data – assessing suitability of other sources of data e.g. from a different location or using modelled instead of measured data
- Using numerical model outputs e.g. climate simulations
- Getting data into a usable format; documenting data workflows
- Using historical data e.g. transcribed hand written records and scanned documents
- ISO 8000 on data quality and mention of other relevant standards
- Metadata, provenance and documentation

iii.*Introduction to visualising environmental data* – 2 DAYS

- Data presentation and labelling - tables and plots
- Plotting data for analysis
- Data visualisation for papers and presentations
- Scripts for reproducible figures

iv.*Data management* – 2 DAYS

- Research data, collected by you and collected by others, formats
- Data management plans
- Journal and funder data policies
- Reproducibility and organising data: file names, README, structures and versions
- Metadata for describing, finding and making data reusable
- Citing and publishing data
- Data security including documentation, backing up, checksum and wider issues
- Best practice for data including ethics, transparency and data protection
- Persistent identifiers and introduction to research objects
- Data sharing, preservation, licensing and trusted repositories
- Data licensing using machine-readable standards (e.g. Creative Commons licenses)
- The European INSPIRE directive and similar initiatives

Full Data Skills Curricula Framework report is available at www.bfe-inf.org.

v. *Interdisciplinary data exchange* (mixed classes engineers, social and environmental scientists) – 2 DAYS

- Sharing and open data – data standards and metadata, interoperability standards
- Introduction to semantic vocabularies across domains and ontologies e.g. Envo
- Syntactic ways to encode data
- Discussion of uncertainties e.g. surveys, model output
- The use of expert judgment in environmental science e.g. emission pathways
- Practical collaborative project work e.g. hackathon or bring your own data
- Thinking of end-users in project design and throughout the project lifecycle
- Data and software citation and publication
- Reusability and reproducibility across domains

OPTIONAL

vi.*Software development ideas for scientific coding* - 3 DAYS

- Design methodologies, diagramming and data structures
- Unit and integration testing
- Requirements capture
- Error handling and debugging

vii.*Object orientated programming* - 3 DAYS

- Object orientated programming: analysis design and implementation
- Design patterns and advanced design methodologies
- Exception handling and exception classes
- Testing strategies, testing classes
- Debugging with classes

viii.*Introductory data science topics* - 1 DAY each

- Relational and non-relational databases
- Advanced visualisation
- Machine learning
- Data mining, including text mining
- Research computational infrastructure - cloud, HPC etc. (some country dependence)

ix.*Data organisation* - 1 DAY

- Best practice on research science workflow, intro to Taverna, Kepler or Vendor
- Intermediate to advanced use of Jupyter or other notebooks for sharing

PRINCIPAL INVESTIGATOR

x. *Data management plans and data repositories* (<0.5 DAYS)

- Expectations of funders, publishers, repositories and other stakeholders
- Data life cycle and the requirements of data management plans
- Legal considerations
- Data security, privacy and sensitive data
- Team organisation for effective data management – designating duties

xi.*Overview of skills for data intensive research* (1 hour briefing)

- Software and computing skills as a team resource
- The Hybrid or Digital Scientist - cases e.g. from Lesley Wyborn or Bryan Lawrence
- Giving credit to supporting roles e.g. CASRAI CRediT

---

*The **e-I&DM Project** is funded by Belmont Forum member countries to advise and provide recommendations to the Belmont Forum agencies regarding policies, programs and procedures to accelerate open data sharing, data reproducibility, data curation, and other aspects of long-term, full-path data management and access. This effort integrates four interrelated Action Themes and collaborations with related international initiatives being undertaken by CODATA, Future Earth, Group on Earth Observations, ICSU-World Data Systems, Research Data Alliance (RDA) and others.*

For more information, please contact Rowena Davis at the e-I&DM Coordination Office: *rowenaidavis@email.arizona.edu*