Institute for Environmental Analytics

Looking ahead: extended recommendations for digital skills in data-intensive global environmental change research

1. Summary

The Belmont Forum e-Infrastructures & Data Management community noted in 2015 that a new data literacy was needed for the 21st Century and a "curriculum was required to expand human capacity in technology and data intensive analysis." The e-I&DM Project, through its Human Dimensions and Capacity Building work stream (Action Theme 4), has created an extensive body of work to establish the curriculum.

The recommendations compiled below are the distillation of work carried out by the AT4 team in 2016 and 2017, beginning with a skills gap survey and report, followed by a workshop on relevant curricula and culminating in a recommended curricula framework and priority actions. This document builds on that body of work and presents more information and a wider rationale for improving digital skills for data-intensive global change research, including implementation ideas and a preliminary estimate of the financial costs of specific actions. The report is written for the Belmont Forum partnership of funding agencies, and is presented for information and use by individual agencies, rather than for formal adoption or action by the Belmont Forum members.

The recommendations are grouped under the following headings:

1. Definitions and assumptions
2. Capacity building
3. Format of delivery
4. Attracting and retaining learners
5. The curricula in different countries: global applicability
6. Costs of training activities
7. The role of the Belmont Forum

The top three recommendations for effectively upskilling researchers are:

i. **Endorsement of the curricula**. Clear and active communication by funding agencies that the curricula are vital skills and a valuable reference for providing training activities.

ii. **Explore sharing existing short courses between partner agencies**, especially that align with the curricula.

iii. Consider **funding priority training activities identified by the curricula**. Priority core topics are 'Environmental data: expectations and limitations' and 'Interdisciplinary data exchange'. The Principal Investigator briefings are also a priority.

2. Recommended curricula: digital skills in data-intensive global change research

The recommended core modules (Appendix, page 10) are designed to enhance skills of domain scientists specifically to make data handling more efficient, research more reproducible and data more shareable – including visualisations for end-users. The five core skills emphasised comprise programming, particulars of environmental data, visualisation, data management and interdisciplinary data exchange. A number of optional modules are suggested for more established researchers which would be useful introductions to widen their data skills and provide new options for examining and processing data, such as machine learning and object-orientated programming.

Two additional modules would be to brief Principal Investigators, providing an overview of data management and skills.

Of the core curricula, the two skill areas addressed least by existing training are 'Environmental data: expectations and limitations' and 'Interdisciplinary data exchange'. Therefore, moving forward in these areas would have the most beneficial impact among scientists and researchers. Materials on the former are likely to exist within university Masters' curricula, so, of the two, 'Interdisciplinary data exchange' deserves greater attention, especially if taught as suggested in a mixed class of environmental scientists, social scientists, and engineers.

3. Capacity building

Capacity building at an individual level promotes the development of conditions that allow individuals to enhance knowledge and skills and to be able to engage in learning. At an institutional level, capacity building supports the promotion of sound management and policies. For digital skills in data-intensive global change research, the curricula and recommendations enable capacity building by fulfilling the recommendations:

i. Individuals, funders and training providers can identify priorities for skills and knowledge development.
ii. Individuals can access courses from partner agencies, promoting relevant learning practically and quickly by using the strength of the partnership.
iii. Funders and training providers can develop training targeted on specific skills gaps which have been shown to be hard to address within the environmental science domain.

4. Extended recommendations and additional information

In addition to the top three recommendations of the curricula report, a range of information and recommendations have been gathered through this work that might be a useful pool of ideas and references for future work. The sources for these ideas include the free-text responses to the skills survey, the curricula workshop, first-hand experience running IEA courses relevant to academics, industry, and graduate students, as well as keeping abreast of contemporary training information (e.g., Kevin Ashley's report).[1]

1. Multi-, inter- and trans-disciplinary

Throughout the curriculum development activity, the issue of clarity of definitions was often raised, so definitions as assumed in the AT4 documents are:

- In the wider world, multi-, inter- and trans-disciplinary are sometimes used interchangeably. In AT4, multi-disciplinary means different disciplines working together, mostly domain scientists who carry out their own work and may liaise occasionally with other domains, often in natural science (e.g., weather expertise liaising with ecology). Inter-disciplinary means domain physical and/or natural scientists working closely with a range of mathematicians, engineers, computer scientists, and social scientists, mostly with the focus on publishing papers and pursuing research. Trans-disciplinary means physical and natural scientists working closely with

[1] Ashley, Kevin (2016) Developing Skills for Managing Research Data and Software in Open Research. Wellcome Trust. https://dx.doi.org/10.6084/m9.figshare.4133916

all stakeholders including any other scientists or researchers, policymakers, the not-for-profit sector, the fine arts and the public, and doing so from the start of a project to produce a range of outputs (data, papers, policy analysis and recommendations, etc.).

- For e-I&DM, inter-disciplinary work is of primary importance when focussing on digital skills for data-intensive research. Interdisciplinary work is key for sharing datasets, methods, and research outcomes across disciplines – from digitising historic records, to citizen science initiatives to using climate model output as input to economic models. Whilst trans-disciplinary work is of great value and a priority for the Belmont Forum, it may produce some digital issues, but is unlikely to have data-intensive needs other than communicating outputs and concepts in an accessible way.

2. Capacity building

Capacity building refers to both individual and institutional changes to develop skills, with varying levels of compulsion.

- Training and capacity building spans a range of activities and levels of direct encouragement, from on-the-job learning to hackathons to open online videos to conferences to summer schools to compulsory training for PhD students to continual professional development requirements for members of learned societies. There is no single training panacea that will work for everyone at all levels. The recommended curricula delineate between skills needed by research scientists and those needed by Principal Investigator. Also, core and optional skills reflect the different roles a researcher/scientist may play in a project. Additionally, from the skills survey, the two most popular ways to encourage mid-career researchers to upskill were '*making recognition of digital skills part of career progression*' and '*providing full financial support for training, e.g., including travel*'.
- The core curricula provide a thorough basis for contemporary digital skills for data-intensive global change research. It is quite possible that competent and successful researchers may not be extremely effective at all of the core skills, so there is a capacity building opportunity to raise awareness of this fact. Having worked closely with a software engineer on the content of the programming, software development and object orientated modules, including running courses on these topics, domain scientists who code and even teach programming may not demonstrate rigor, and whilst the iterative nature of research programming may not require software engineering approaches, it is always worth knowing best-practice. Python is now used extensively in research and it inherently offers multiple ways to do the same thing, so awareness of best-practice is particularly pertinent.

3. Format of delivery

The format of training delivery was explored during the curricula workshop, presented in the curricula report, and are directly reproduced here. The format should be considered a key element of the curricula. In delivery of the curricula, emphasis should be on *practical sessions with real-life examples, reproducibility of research, and interdisciplinary work highlighting data sharing*. Further recommendations include:

- Use real-life examples and, if possible, set aside sessions to deal with an individual's real research data. Design training acknowledging what a researcher needs.

- Encourage attendees to bring their own devices, if possible, with the intention that software and examples installed will work after the training. Dissemination is possible in other ways and the fewer barriers to continued experimentation the better.
- Training needs to be used soon after delivery or it will be forgotten. Good online resources for post-course reference, using a follow-up webinar presentation, and/or making a trainer available on a specified day for email questions, calls, or other troubleshooting methods are ways to make the training survive into the workplace.
- Emphasise the tangible benefits of the efficiencies (e.g., in programming, data management), collaboration between countries and interdisciplinary data sharing.
- Online delivery is a good way to complement face-to-face training, for example, bringing all attendees to a known common level before starting a course. Online has the asynchronous advantage, addressing time-zone issues and availability, with discussion boards then useful to promote peer-to-peer learning.
- A helpful format for snippets of online training is to host alongside data repository access. For example, UK Data Service provides short videos on the same web page as the data access.
- Teaching of a curricula need not be on consecutive days and breaks between learning is an advantageous way to include further learning and project work.
- Assessment is a positive way to ensure that a module has been understood (outcomes-based) and certificates and certification should only be provided if there is, at the very least, an assurance of complete attendance.

4. Attracting and retaining learners

The mechanisms for attracting and retaining learners was a significant feature of the curricula workshop and was presented in the curricula report. The first six items below are directly reproduced here from the curricula report and the remainder are additional recommendations and observations.

- Exploiting conference training slots is a good way to attract a diverse group of domain scientists for short taster training sessions (one day or less).
- The use of ingress, kick-off meetings, and mid-term meetings of funded projects is a way to gain a wider audience than by self-selection. Separate training activities could be aimed at potential future applicants for upskilling and connecting with the widest community.
- The team approach: training should not be compulsory for everyone on a funded project, but selection of individuals for particular upskilling could be encouraged, for example, via nomination by the Principal Investigators.
- Provision of training aligned with the curriculum serves as messaging from the funding agencies that these skills are necessary and valued.
- Training could be linked around particular research challenges, including using data from other domains, such as hackathons.
- Some funded projects have held summer schools (e.g., GOTHAM) and it would be beneficial to include elements of the curricula as part of the school.

Additional to curricula report:

- It is useful to acknowledge that providing targeted training at the point of need is a key way to capitalise on learner motivation and an example of performance support.

YouTube videos and the Kahn Academy use this approach and it reinforces the efficacy of the UK Data Service model of short videos next to the data access. The 'immediate need' advantage would also be present when providing, training related to a funder requirement included in a call for proposals (e.g., data management planning).

- Online training is attractive for the wide accessibility and longevity of courses and modules after the initial outlay. However, online training, especially open online courses, are notorious for high drop-out rates, and learner sign-up can be one or two orders of magnitude higher than those who complete a course.
- As noted in the curricula report, there is a wealth of openly accessible online training on data management available, it is motivating researchers to engage with the training and put the ideas into practice that is the greater challenge. To tailor general data management training to specific domains would be one way to help engagement.
- But advantages and emerging trends for keeping learners engaged are mobile-compatible, high use of video/audio and gamification. Gamification can be thought of in three ways - role-playing (learning whilst 'playing'), games as a way of learning (e.g., including puzzles, hidden rewards for completion), and using one's competitive nature such as leader boards and badges for completion.
- There are lots of courses which do not charge for attendance. Many European projects, for example, contain summer schools and one day workshops as part of the project and many online courses, of course including MOOCs, do not charge attendance fees. Therefore, a new course which requires fees to be paid needs to have a truly desirable unique selling point such as being on an emerging topic, or a niche subject or a skill which is in high demand. This is discussed further in the costings section.
- When there is no charge to attend a course, the expectation of quality and usefulness is still as high as if it were a paid course, since the learner has given up their time to be there. Furthermore, for unpaid classroom courses, it is likely that 10% to 40% of the attendees will either cancel at short notice, not turn up, or not attend the entirety.
- Marketing is important to ensure good attendance, even if there is no charge for the course. Effective marketing will require effort, especially to extend to an audience beyond one's usual links, such as a different scientific domain. Seeking out venues and meetings where these new audiences abide is a useful first step.

5. The curricula in different countries: global applicability

Having attended the CODATA-RDA summer school for research data science, aimed at developing skills for individuals in low and middle income countries, it is apparent that the skills required for data-intensive research are very similar worldwide. The research data science topics taught at the summer school would be equally applicable to anywhere in Europe or North America. Whilst the requirements for researching data-intensive global environmental change are different from research data science, the underlying skills that need to be taught are very similar and it is the emphasis which is different. There is no need for separate curricula in different countries since the fundamental skills needed are sufficiently similar. It is likely that research infrastructure and institutional support may vary, these variations can be between a range of countries, not necessarily corresponding to GDP. Furthermore, the ever-increasing availability of open source code, open data, and improving access to remotely hosted services, all assist with limiting inter-country differences.

6. Costs of training activities

When comparing different training methods it is useful to consider costs. Endorsement and messaging is always cheapest and this simple information dissemination can be enough to start changes in some people.

More formal training requires a high initial outlay. When considering face-to-face classroom training, the two major costs are teaching staff preparation time and accommodation for the learners. Developing training materials is time consuming and, for new material by someone with existing expertise, can take up to a ratio of 10:1. A brand new one day course, complete with lesson plans, presentations, interactive sessions and structured exercises with model answers, can take up to ten days to prepare, especially for technical topics such as programming where learners require a high level of exact materials. If existing materials are reused then only delivery time need be considered. The teaching room, accommodation and food for residential training is a significant cost and, per attendee, can in the UK range from £100 minimum which might necessitate shared bathrooms, to £200 per person per 24 hours for a three star venue. Many conference facilities charge per person in this way and the teaching room is included in the delegate rates. Venues will provide free Wi-Fi, which may not be fit for thirty students to use simultaneously, and only specialist venues will provide computer labs in the package.

The costing shown in Table 1 assumes development time of 5:1 and two trainers required for delivery, giving an overall ratio of 35:5 or 7:1. For costing, a trainer day rate minimum of £400 (~€450) and maximum £800 (~€900) has been assumed. Administration is required for arranging bookings, sending out information and liaising with the venue. The payment of travel can vary depending on distance and whether this is part of the package. Based on a class size of 30, the estimate in Table 1 shows that a new five day summer school is likely to cost at least £1000 (~€1130) per attendee.

Table 1: estimated simplified costs for a five night residential course, 30 attendees, 2 trainers

ACTIVITY	DAYS	MIN	MAX
Training preparation	25	10,000	20,000
Training delivery (2 trainers)	10	4,000	8,000
Administration	3	£600	1,200
Residential venue and food	5	16,000	32,000
Trainer travel	-	0	1,500
Attendee travel	-	0	9,000
TOTAL (GBP)		£30,600	71,700

The costs of developing online courses can vary markedly. The major costs, once the learning platform is in-place and mastered, are preparing materials, transferring materials to the platform and developing necessary interactive sessions along the way, and the cost of professional video content. One day of filming at one location is typically about £4000, including post production, which might produce a maximum of 60 minutes' usable video and for very high quality video the cost may be as much as £1000 per minute of finished video.

Table 2: estimated simplified costs for one day (eight hours) of online training

ACTIVITY	DAYS	MIN	MAX
Training preparation	20	8,000	16,000
Transfer to online system	10	5,000	9,000
Administration	3	600	1,200
Video and audio with editing	-	3,000	15,000
Web hosting per year	-	600	1,200
Graphics	3	900	2,400
TOTAL (GBP)		£18,100	£44,800

Having recently developed online courses of several hours of learning time, the costs were high, even without paying for some contributors' time to appear on camera. The ratio of approximately 20:1 for development time compared with the finished materials was remarkable but necessary. The perfection expected of online material involves iterations on content and look, along with very carefully prepared materials to ensure they are effortlessly engaging and varied for the learners. Graphs and diagrams might need to be redrawn, along with bespoke infographics or animations, requiring the services of design professionals. The estimate in Table 2 does not include the cost of the learning management system itself, assuming an open source solution or the system provided as part of the transfer to online costs. Nevertheless, with less stringent production standards and the inclusion, for example, of video of classroom sessions with little or no editing, costs could be lower than the minimum indicated in Table 2. Once in place, an online course can teach hundreds.

Neither the classroom nor online costings shown here include marketing expenses. Promotion requires significant time, effort, imagination and, ideally, an existing well-developed network of contacts. A newly developed course is likely to have to run at least three times before it develops its own momentum to attract more attendees by reputation and awareness of the relevant community.

To encourage face-to-face courses to be sustainable it may be useful to factor in transition from funded to fee-paying. It is likely that this model is feasible for skills of the AT4 e-I&DM curricula since many would be applicable to research staff in a variety of industries. A course could be capped at 90% funded for the first run, with the remainder paid by attracting one or two fee-paying attendees from industry, tapering to 30% for the third run with a greater number paying fees. Training providers may be unsure whether courses funded by public money should or could have non-academic attendees, so this would have to be made clear from the outset. The compulsion to include fee-paying attendees in the very first course would assist with keeping the content relevant to industry and academia.

Secondments and on-the-job training do not require the high initial outlay of classroom or online training, but obviously require the 'learner's' time to be paid, albeit on joint projects useful work, such as published papers, could still be an output and the costs are therefore difficult to assess.

7. The Role of the Belmont Forum in Digital Skills Training

The Belmont Forum is well positioned to influence research communities. The recommendations below arose from the three existing AT4 outputs and are collated here for convenience and full context.

Recommendations from the skills gap report:

- Partner agencies may prefer to recommend the curriculum framework rather than a list of courses, since the curriculum is likely to change more slowly than individual courses.
- If maintaining a list of courses is beneficial, automated systems could be set up to create a list which is always up to date, finding courses via an internet keyword search. An example of this is TESS by ELIXIR which has developed open source software, in Python, to create a training portal.

Recommendations from the curricula workshop:

- Require Belmont Forum-funded project scientists to attend certain training activities, for example, on data management.
- The Belmont Forum is in a powerful position to provide thought leadership in the form of position papers. Relevant topics could include:
  - The importance of open data and examples of how research has been improved with open data along with open enterprise examples.
  - Encouragement of basic knowledge of the research data lifecycle (including the role of trustworthy data repositories) and best practice in data management.
  - Recommending standard vocabularies and encoding as appropriate; including persistent identifiers.
  - Showcasing and endorsing researchers who promote and carry out data sharing, especially across scientific domains.
  - The need for academic institutions and publishers to recognize and value open research and collaboration for hiring, promotion, and publication.
- The Belmont Forum can target a funding call on the re-use of data, especially involving discovery of data outside one's domain. There could be conditions in the call on collaboration and exchange of staff as secondments.

Recommendations from the curricula report:

i. Endorsement of the curricula. Clear and active communication by funding agencies that the curricula address vital skills and are a valuable reference for shaping training activities.
ii. Explore sharing existing short courses, especially that align with the curricula, among agencies.
iii. Consider funding priority training activities identified by the curricula. Priority core topics are 'Environmental data: expectations and limitations' and 'Interdisciplinary data exchange'. The Principal Investigator briefings are also a priority.
iv. Provide opportunities or direct funded groups to access training, ensuring these resources are widely and openly available.
v. Decide whether the advantages of 'Belmont Certification' are sufficient to resource a scheme and provide funds for organising, administering and quality assuring a certification process for the medium to long term.
vi. Consider funding training activities based around curricula as part of the process of implementing Collaborative Research Actions.
vii. Consider targeting a funding call on the re-use of data, especially incorporating data outside of one's domain.

The recent website refresh and upcoming funding activities will also give the Belmont Forum a high profile in the next year and this exposure could be used to direct learners to information on the curricula and be aligned within the existing communication strategy.

5. Conclusions

The recommendations to the Belmont Forum for the November 2017 plenary are simple and highly actionable by funding agencies together or individually. The recommendations fulfil the capacity building remit of Action Theme 4. The extended list of recommendations and additional information presented in this report are useful for greater detail on implementation and informing the discussion on developing digital skills in data-intensive global change research.

6. Acknowledgements

The skills gap analysis, curricula workshop, recommended curricula topics and this recommendations report were all funded by the Natural Environment Research Council of the UK and prepared for the e-Infrastructures and Data Management Project of the Belmont Forum. I would like to thank all in the e-I&DM for their support and input to Action Theme 4, it has been enjoyable working with such a delightfully dedicated group.

Vicky Lucas 12th November 2017
Human Dimensions Champion, e-I&DM and
Training Manager, Institute for Environmental Analytics, v.lucas@the-iea.org

APPENDIX – Details of curricula

A = first year PhD; B = final year PhD; C = postdoc, D = mid-career, Principal Investigators.

CORE

i. **Programming for data intensive research (for those who already code)** (A) – 4 DAYS
- Unix/Linux, shells, syntax and using the command line
- Coding standards and 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** (A) – 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** (A and B) – 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** (A and B) – 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

v. **Interdisciplinary data exchange** (B, C and D, 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** (C and D) 3 DAYS

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

vii. **Object orientated programming** (C and D) 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** (C and D) 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** (A to D) 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