Skills Gap Analysis

e-Infrastructures and Data Management in Global Change Research

Vicky Lucas

Human Dimensions Champion

v.lucas@the-iea.org

7th April 2017

1. SUMMARY

The Belmont Forum e-Infrastructures and Data Management group has championed four action themes to deliver sustainable e-infrastructures for global change research. Action Theme four (AT4) is the human dimensions theme and seeks to support training and curricula in data-intensive environmental science for delivery to environmental, social and computer scientists. AT4 is funded in the UK by the Natural Environment Research Council.

This skills gap analysis is the first formal output of AT4. The skills gap analysis is based firstly on a survey completed by current practitioners in the science and data management of global change research and secondly on additional information gathered from complementary global organisations and professionals. The survey includes a catalogue of recommended existing training activities relevant to the Belmont Forum. Analysis of responses to the survey indicated that priority challenges and the skills most in need of improvement are:

- Data complexity
- Data standards
- Data discovery, finding relevant and potential data sources
- Data management
- Overcoming barriers to data sharing, including cultural and interdisciplinary issues
- Improving programming and data analysis workflow
- Improving computational and numerical analysis skills

The survey results and these topics in particular will be the starting point to inform the curricula workshop to be held in late April 2017.

2. INTRODUCTION AND AIMS

The Belmont Forum e-Infrastructure and Data Management (e-IDM) group recognises that, for global change research to flourish, a range of skills are necessary, increasingly so in the broad area of data intensive digital skills. Global change research encompasses the interdisciplinary investigation of the natural world and is increasingly data intensive, requiring significant computing resources and efficient techniques for effective analysis. A short survey was developed, and conducted with participants working in global change research, to establish where skills gaps exist or are emerging. A copy of the survey questionnaire is available in Appendix A.

Highlights of the results of the survey are presented along with other relevant information gathered from discussions with complementary organisations and individuals. This report is an output of the fourh action theme (AT4) of the Belmont Forum e-Infrastructure and Data Management CRA and carried out by funding from the Natural Environment Research Council of the UK.

This report aims to:

- inform discussions on the digital skills necessary for global change research with up to date information.
- collate information on the priority skills, drawn from a range of practitioners.
- list existing relevant recommended training activities.
- identify and learn from existing relevant skills reports and surveys.
- highlight both training relevant conclusions and wider issues.
- recommend next steps on informing curricula to be developed later this year and moving to a 'Belmont Certification' of courses.

The Belmont Forum is a group of the world's major and emerging funders of global environmental change research. The Forum aims to accelerate delivery of the environmental research needed to remove critical barriers to sustainability by aligning and mobilising international resources. The Belmont Forum e-Infrastructures and Data Management Collaborative Research Action is leveraging worldwide conversations on data sharing e-infrastructures to coordinate and promote access to transdisciplinary research data generated by Belmont Forum projects.

Work is being undertaken as four independent but interrelated Action Themes, described in the Phase II Implementation Plan.

- Action Theme 1 (Coordination Office) coordinates activities and liaisons.
- Action Theme 2 (Data Planning) is data policy and planning, to promote active and effective management and stewardship.
- Action Theme 3 (e-Infrastructures) will identify the most critical issues to address, via workshops.
- Action Theme 4 (Human Dimensions) concentrates on human dimensions, supporting training in data-intensive environmental science, including developing curricula.

3. BACKGROUND AND SOURCES OF INFORMATION OUTSIDE THE SURVEY

In 2015 the Belmont Forum e-Infrastructure and Data Management group produced 'A Place to Stand'. The report consists of a set of recommendations on how the Belmont Forum can leverage existing resources to accelerate global change research. A recommendation of the report was to 'support the development of a cross-disciplinary training curriculum to expand human capacity in technology and data-intensive analysis methods for global change research and increase the number of scientists with cross-cutting skills and experience in best practice'. Action Theme 4 developed directly from this recommendation.

'A Place to Stand' identified a number of sources of training and skills needs, one of which was the UK Natural Environment Research Council report on skills needs from 2012. This NERC report prioritised the skills of postgraduates and included as the top three:

- modelling,
- multi-disciplinary skills,
- data management

This Skills Gap Analysis is a progression on 'A Place to Stand' and the NERC work of 2012. This section includes reports recommended via survey respondents, recent work by the US National Science Foundation, other organisations that have come to the fore in compiling this report, the recently formed Copernicus Academy which may offer a format relevant to the curricula and certification, and key thoughts from the Action Theme 3 workshop held in November 2016. The results of the skills survey itself follow in section 4.

3.1. Additional skills surveys identified by respondents

Respondents were asked to list relevant surveys in the hope that these surveys would provide 1) an awareness of the type of data management questions being asked by others and 2) point to reports or findings that would broaden understanding of skill gaps and requirements. Of 15 suggestions, several were actual surveys that were currently closed and several could not be accessed. Others were not subject specific, addressing peripheral topics such as digital competencies in the general population and perceptions of open data. However, one report in particular appears to be a rich resource, and may be considered as further reading to enhance understanding of the current skills and needs climate in data management:

- Ashley, Kevin (2016) *Developing Skills for Managing Research Data and Software in Open Research*. Wellcome Trust. This report focused on four questions: a) determining skills gaps and needs among 1) data specialists, 2) the research community; b) current models for skill development and metrics for effectiveness; 3) the role of data scientists in research institutions; and 4) long term actions to build data management capacity. The report suggests several pertinent conclusions: first, support staff, as opposed to researchers themselves, are often the recipients of data skills-building activities. Second and third, more training material is directed at data skills than software skills, but initiatives that bring together otherwise isolated professionals such as data carpentry initiatives or hackathons prove to be successful, though they often rely too heavily on volunteer support to be sustainable. Fourth, greater coordination among funders is postulated to be an important step in achieving a sustainable, current, and evolving capacity building effort.

3.2. US National Science Foundation

In late 2016 the US National Science Foundation (NSF) released a call for 'Training-based Workforce Development for Advanced Cyberinfrastructure (CyberTraining)'. The overarching goal is to nurture the scientific workforce (Figure 1). The call closed in January

2017. The training activities developed as a result of this call will be relevant to the themes of this report and a source of skills development. Efforts will be made to track the successfully funded activities.

> The overarching goal of this program is to prepare, nurture and grow the national scientific workforce for *creating, utilizing,* and *supporting* advanced cyberinfrastructure (CI) that enables cutting-edge science and engineering and contributes to the Nation's overall economic competiveness and security. For the purpose of this solicitation, advanced CI is broadly defined as the resources, tools, and services for advanced computation, data handling, networking and security.
>
> This solicitation calls for developing innovative, scalable training programs to address the emerging needs and unresolved bottlenecks in scientific and engineering workforce development of targeted, multidisciplinary communities, at the postsecondary level and beyond, leading to transformative changes in the state of workforce preparedness for advanced CI in the short and long terms.

Figure 1: excerpt of the NSF call on training for cyberinfrastructure (Read PDF)

The report most relevant to this work is the Federal Big Data Research and Development Strategic Plan which seeks to expand the community of data-empowered domain experts, critically requiring 'training to build human capacity' and 'as scientific research becomes richer in data, domain scientists need access to opportunities to further their data-science skills, including projects that foster collaborations with data scientists, data science short courses and initiatives to supplement training through seed grants, professional development stipends and fellowships'. These comments indicate the wide opportunities for building capacity via varied training approaches and through incentives with examples of seed grants and project funding with collaboration in mind.

3.3. Organisations

Researching this report and speaking with individuals revealed a number of organisations that have produced reports and analyses relevant to the work of the Belmont Forum e-IDM.

- EDISON is a European project to accelerate data science as a profession. The project has developed a competency framework for data scientists which is a comprehensive guideline for training and assessing professional status and a searchable listing of university programs relevant to data science.
- ELIXIR is the European life science infrastructure for biological information and training is the focus of the UK node. In addition to coordinating training activities provided by academic institutions, ELIXIR has a training materials and events infrastructure project (TeSS), which automates collecting information about training and makes it searchable by the community. This is written as open source code and could be repurposed to digital skills for global change research.
- The Council on Library and Information Resources (CLIR) produced a report on The Problem of Data in 2012 which highlights the importance of data preservation, metadata, curation, ethics and data reuse being integrated into research workflow. The report proposes a network of data specialists who are aligned with disciplines and/or regional or national organizations.
- A curriculum framework for digital curation training has been developed by DigCurV for the long-term management of digital collections.
- Foster is an e-learning platform dedicated to training resources for Open Science.
- Elsevier provide guidance on data management to make data reusable.

3.4. Copernicus Academy

The Copernicus Academy is a network set up by Copernicus, the European initiative on providing open source satellite data and services to stimulate the use of Earth observation. The Academy is an affiliation of like-minded organisations within the community to exchange training materials and promote the use of Earth observation and empower people with skills to use the data and services. Organisations including universities, research institutions, business schools, private and not-for-profit were encouraged to apply for Academy status. The incentive was to be part of a network, with no explicit funding available. It is possible that this model of affiliation under a common cause for the purpose of capacity building may be transferrable to the digital skills of the Belmont Forum e-IDM, for either the provision of training or certification. The Institute for the Environmental Analytics, in association with the UK National Centre for Earth Observation, was awarded Academy status earlier this year and via this connection, the suitability of the Academy as a model will be assessed.

3.5. Action Theme 3 workshop Paris November 2016

Action Theme 3 (AT3) is tasked with identifying critical issues in e-infrastructures and data management. Workshops are used to identify issues, attended by scientific researchers, data managers as well as leaders in related organisations, including the EC, OECD and CODATA. A two-day workshop was held at the end of November 2016 to initiate a funding call on learning from and improving the outcomes of existing Belmont Forum funded projects. The issues aired at the meeting arising within human dimensions were much wider ranging than the more directed survey responses:

- Learning from commercial expertise e.g. data streaming NetFlix, Google search
- Helping companies use data e.g. Landsat, Sentinels, CMIP6 outputs
- Recruiting expertise and increasing the pool of job applicants
- Identifying roles as ‘data organisers’ who facilitate transformations and develop tools and ‘digital scientists’ with domain and computing expertise
- The use of machine or deep learning, it not being clear exactly how they work and that feels ‘disturbing from a scientific point of view’
- Including data management and its principles in honours level degrees
- Moves to automated scientific workflow might be useful but critical to know when human intervention is valuable and this is especially an issue for big data. Automation is a new paradigm from which to do science.
- What does the European Open Science Cloud mean for researchers?

4. RESULTS AND DISCUSSION OF THE SKILLS SURVEY

To gain the opinions from a diversity of current practitioners in global change research, a short survey was designed and distributed to organisations and individuals associated with the Belmont Forum, and also promoted via a blog post, listservs and Twitter.

4.1. Respondents

The skills survey was sent to a range of Belmont Forum contacts and completed in November and December 2016 by 164 people who work in a range of institutions around the world. Of these, 76% were employed by government or universities and 81% came from either North America or Europe, with 6% of respondents from Africa. (See Appendix A for survey text.)

The question of field of expertise (Figure 2) revealed that 'computing, computer science and/or data science' was most common followed by 'atmosphere, climate, cryosphere, hydrology and/or oceans'. Respondents were allowed to select up to two areas of expertise, and 38% chose two fields of which the vast majority elected the computing skills alongside another area of expertise. The high rate of identification with computing skills is reassurance that the survey has reached those who are relevant to e-infrastructure and data management. The field of expertise most commonly indicated in 'other' was information science. The primary role for most respondents (Figure 3) was researcher, followed by data specialists and managers.

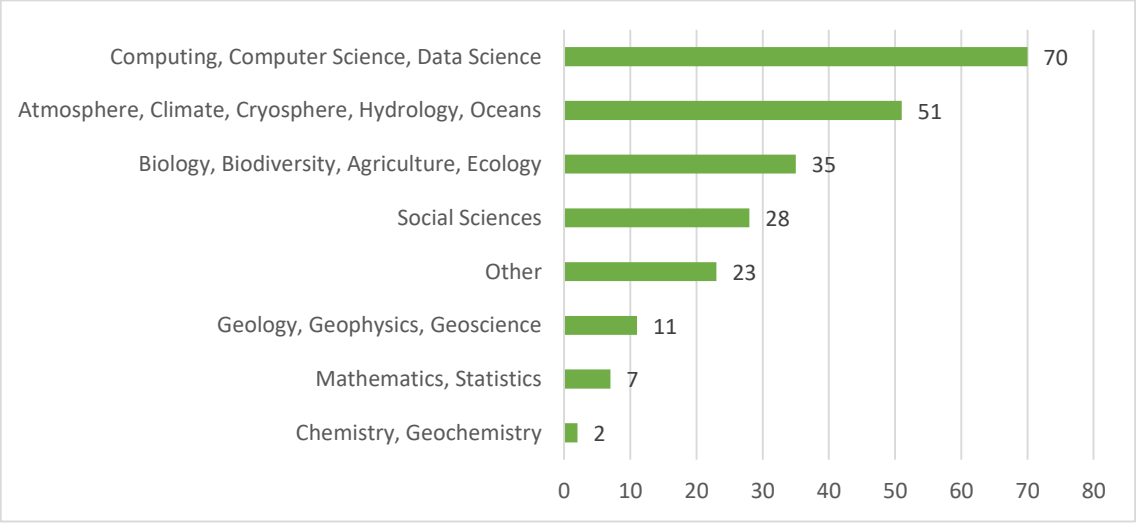

Figure 2: field of expertise of respondents (up to two could be selected)

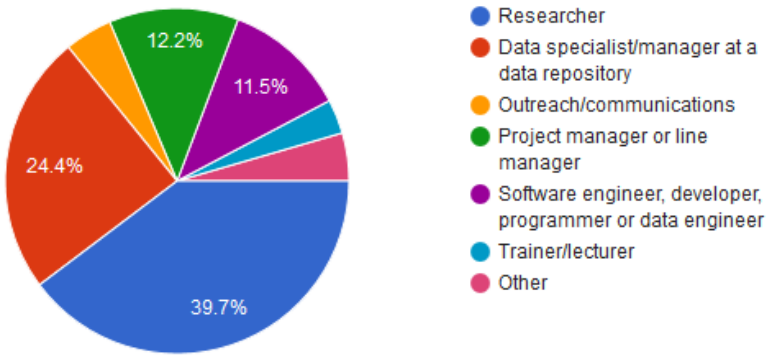

Figure 3: primary role of respondent

Half of all respondents were mid-career, having worked for 6 to 20 years, with 38% more experienced and 11% with up to five years' post-doctoral qualification. The majority, 56%, supervised or managed postgraduates or postdocs and 5% taught undergraduates, the remainder of 39% did not teach, supervise or manage others.

When asked for a rating of 1 (lowest) to 5 (highest), most respondents indicated that they had good technical support and computing infrastructure, 62% rating as 4 or 5 out of 5. There was a minority of 15% who scored this support as 2 out of 5 or lower and from examination of individual responses. On the question of 'how multi- or interdisciplinary are the data you use and the people you regularly work with?' there was a positive response with 63% rating this as 4 or 5 out of 5, and only 11% as a 2 or lower.

The expertise and professional roles of the respondents make the qualitative results of the survey useful to inform discussions with these contemporary views of current practitioners. The responses were dominated by individuals from Europe and North America, so any conclusions should be viewed with this in mind. The survey has not been analysed by continent, expertise, role nor years' experience, but this could be carried out in future and the data will be made available for others to use and analyse.

4.2. Largest challenge in data use

There were a number of questions in the survey investigating firstly the digital skills most necessary for global change research and secondly those skills needing improvement. The opening question in this section asked 'how would you describe the largest challenge you encounter in your data use' and revealed a fairly equal split between a number of issues as shown in Figure 4. The top four issues accounted for 73% of respondents:

- Data complexity
- Lack of data standards and exchange standards
- Finding relevant existing data – knowing what's out there
- Data management and storage

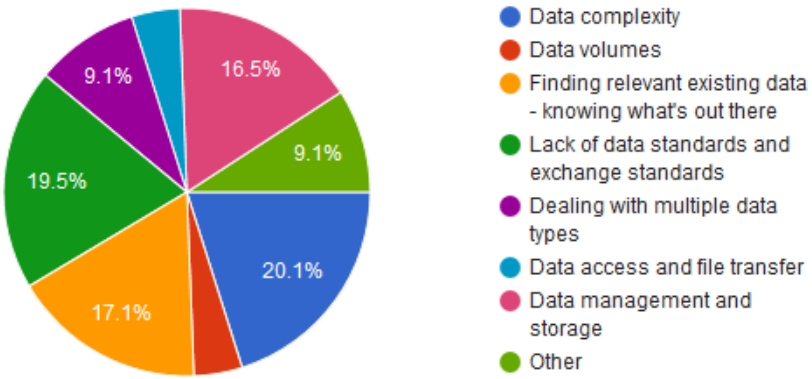

Figure 4: largest data use challenge encountered by respondents

Data complexity points to the varied difficulties including data formats, pre-processing or binning, missing data and noise as well as the challenges of appropriately combining multiple data sources for analysis such as matching spatial or temporal resolution. The lack of data standards indicates the propensity for individual organisations to choose specifications that are incompatible with each other, and this is a barrier to sharing data, discussed later in this section. The issue of finding relevant existing datasets is currently being addressed through persistent identifiers and linked data but is a work in progress.

Data management was a recurring theme through the survey and is discussed later in this section. The issues of data volume and access/transfer were less common problems, consistent with the result that 62% had good technical support and infrastructure, rating it as 4 or 5 out of 5.

The 'other' category of comments provided more specific clarifications on the theme of both data complexity and standards for reuse. Comments included insufficient metadata, absence of documentation and 'poor description of methods'. Taken in combination with respondents reported difficulty finding relevant existing data sets and data management challenges, these comments highlight a need for more training on data management generally, and specifically on the value of capturing detailed metadata from the onset of a research project. Increased inclusion of rich metadata will not only enable effective reuse of data, but also enhance the discoverability of data sets.

An interesting comment was made on the need for 'skills in electronics and programming to design or adapt new tools' as well as knowledge of 'uncertainties in different types of data'. These technical data engineering skills are included as a core competency required to be a data scientist as defined by the EDISON project. The EDISON project is an EU initiative to develop and accelerate data science as a profession. The issues of data complexity can be addressed by familiarity with sensor limits and uncertainties. Another pertinent challenge was expressed as 'a reluctance to publish data'.

4.3. Necessary skills and areas needing most improvement

To explore the types of digital skills that respondents felt most necessary to global change research, up to three could be chosen and the vast majority opted to identify three areas (2.9 per person). The full results are shown in Figure 5 and the three most common responses accounted for 47% of all the selections:

- Data processing and analysis
- Programming (creating clear and robust code)
- Data management

'Data processing and analysis' is the most generic skill that is needed to work with data; the ability to handle and start to make sense of the data. The subsequent two responses aspire to the fundamentals of best practice: that good coding is needed and that data management is a necessary part of the data life-cycle. From examination of individual responses, it was clear that many of those whose own core role was data management also selected data management as one of the necessary skills.

The three least popular responses were perhaps more specialist and were 'use of High Performance Computing', 'using statistical or numerical models' and 'use of big data technologies'. In examining individual responses it was the atmospheric, oceanographic and climate change scientists who were most likely to indicate HPC [as a necessary skill]. On the overall low priority allocated to 'using statistical or numerical models', perhaps this is because such models have been designed to be easy to use or are deceptively easy to use. In the case of the 'use of big data technologies' (e.g. Hadoop, Spark) it may be that this is an emerging issue, only encountered by subset global change researchers to date.

In the 'other' comments on necessary skills, two areas stood out: firstly awareness of data standards with one notable comment that there may be too many standards, and secondly broadly on data quality and 'understanding the reliability of diverse datasets' and 'transforming correctly historic analogue data into electronic formats'.

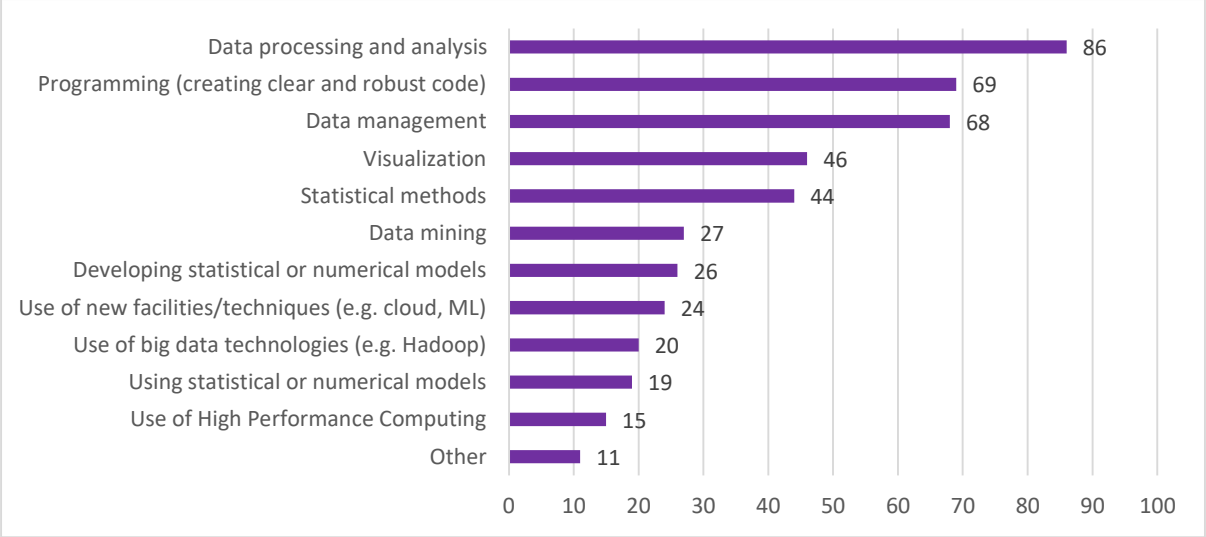

Figure 5: most necessary digital skills for global change research

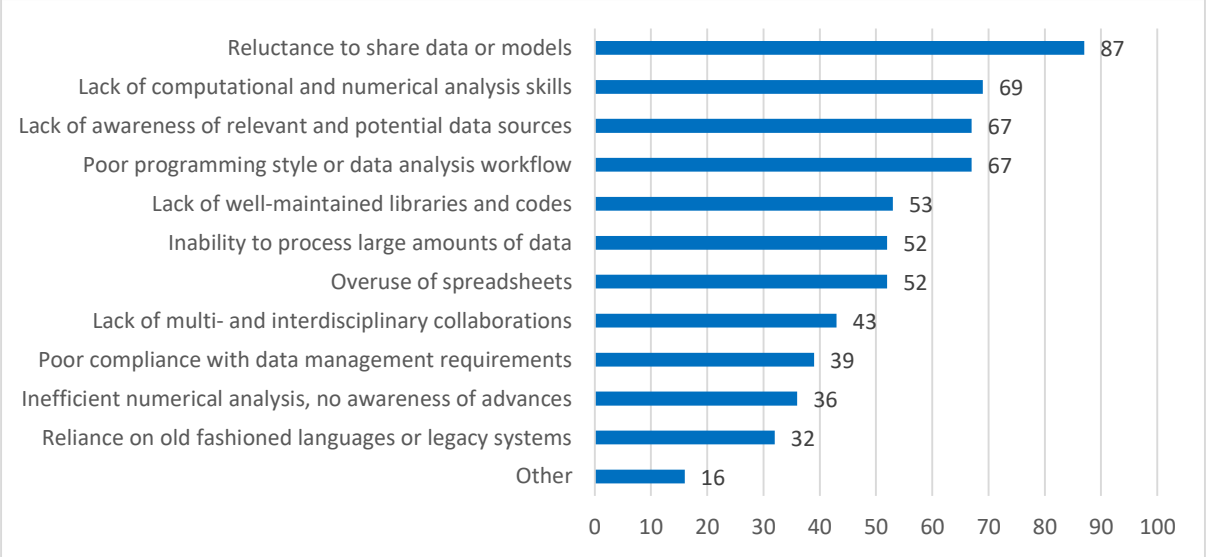

Figure 6: research habits and digital skills that need most improvement

Having established the types of digital skills necessary, the key question was posed on which research habits and digital skills need most improvement (Figure 6). The number of responses was unlimited and the average was 3.7 selections per person. The top four responses (of the twelve available) accounted for 47% of all the selections:

- Reluctance to share data or models
- Lack of computational and numerical analysis skills
- Lack of awareness of relevant and potential data sources
- Poor programming or data analysis workflows

The most popular response, the reluctance to share data or models was identified by about half of all respondents. Reluctance to share is multifaceted, cultural and/or protectionist, as well as potentially rooted in digital skills inadequacy (from fear that errors will be found or general difficulties in sharing relating to data standards). Training could assist with improving sharing and there may need to be a combination of approaches. Training is only one way to influence actions when the simple lack of skills and knowledge is intertwined with

overcoming habits or preconceptions. To improve on the reluctance to share data and models there are several approaches which can be considered. In order of intervention level, they are:

- Messaging – from funders or publishers for top down influence that sharing is beneficial. Contrastingly, bottom up messaging can emerge directly from the scientific community itself via general peer pressure as cultures and technologies change, or via pioneering individuals.
- Communication – the soft end of training, consisting purely of raised awareness: that if there are ways to make sharing happen and people know about them then some sharing will start to occur because of those individuals and groups who automatically see its benefits.
- Training – the next step is to offer training as an additional level of help to assist the already somewhat motivated – those who can see the benefits or who are ready to be convinced of the benefits and just need some guidance which might be because they are pioneering (alone) in their organisation or peers are too busy to help/guide or they simply need the skills and knowledge.
- Incentives – intervention at the macro (institutional or funder) level creates or boosts the motivation to share sufficiently to make it happen.

Three issues followed as commonly identified needs, which were:

- Poor programming or data analysis workflow
- Lack of computational and numerical analysis skills
- Lack of awareness of relevant and potential data sources

This reinforces the finding already expressed that the top necessary skills also feature as skills to be improved. The area least commonly indicated as needing improvement was the reliance on old-fashioned programming languages or legacy systems.

In the ‘other’ comments provided as the skills requiring most improvement, there were several thoughtful additions including the ability to share sensitive data, ‘understanding end to end workflows’ and gaining ‘reproducible results in published papers’, also ‘finding ways to make connections across disciplines, seems to mostly happen by chance’.

4.4. Research group leaders and engaging mid-career researchers

Anecdotal evidence has suggested that some more established and influential researchers who lead groups and peer review, may be lagging behind in the cutting-edge digital skills required for global change research. A question was included to establish a general view on the skills of research group leaders in the areas of data discovery, new analysis methods and new technologies or computing skills. The results area shown in Figure 7. All three response patterns showed an approximately normal distribution and it was clear from examining individual responses that many had chosen the mid-range, 3 out of 5, for all three responses. There was a skew to lower skill in ‘computing and new technologies’. It may be that, given that the survey was completed by those who had generally good technical support and infrastructure (section 4.1), the respondents themselves were part of a generally more highly skilled or resourced environment, including group leaders.

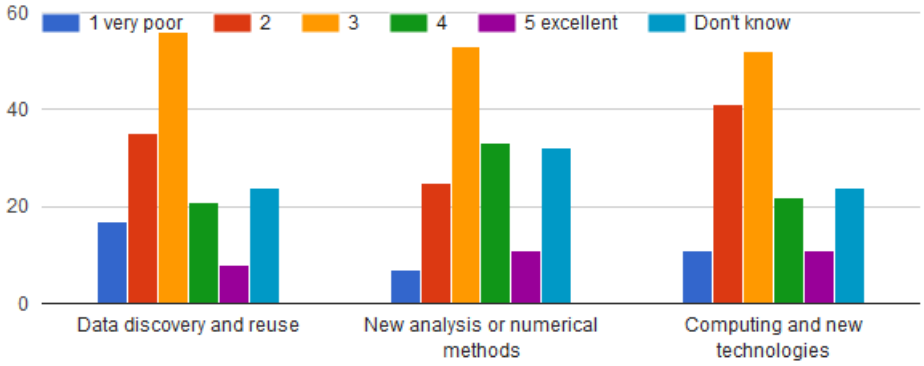

Figure 7: perceived skills of research group leaders

Another question spurred by anecdotal evidence is that mid-career researchers might be harder to attract to training activities than early-career researchers. To address this and to inform future thoughts on how to maintain skills throughout careers, a question was included on how to engage mid-career researchers (Figure 8). Up to three methods could be selected and the average was 2.8 per person. The two most popular responses were:

- Making recognition of digital skills part of career progression
- Providing full financial support for training activities e.g. including travel

The least popular approach was briefings by senior management. The 'other' comments on engaging mid-career researchers was to allow participants to work on their own data or to integrate training with real-world data analysis projects.

Figure 8: best ways to engage mid-career researchers in developing their skills

4.5. Recommended courses

Respondents to the survey were asked to indicate any training in digital skills or global change research that they valued. A summary of responses is listed in section 8 (page 17), and includes recommendations sourced outside the survey. The training activities are divided into online offerings, including Massive Open Online Courses, and face-to-face training, which could be either workshops or summer schools. Online materials range from

documents to webinars to web-based tools to the high quality production of MOOCs with the added element of peer-to-peer learning. All of these materials and courses are continually changing and any course may be a one-off, which makes cataloguing a process that would need to be reviewed regularly or automated, although automation may then result in course listings without the added benefit of endorsement.

Whilst only two respondents recommended certification, shown in section 8, this approach could be a relevant way to address skills development. Professional certification and accreditation was proposed by data managers and information scientists, it could be relevant to scientists, that certain skills could be required to gain professional chartered status or fellowships, although this would require discussions with the relevant institutes and learned societies.

As shown in section 8, the most commonly mentioned type of online offering was MOOCs, although a very small number of people mentioned several different MOOCs. The most common training recommendation was DataONE, the data observation network for Earth. DataONE is an example of a resource that allows data discovery, offers supporting tools and provides training in the form of webinars, screencasts and searchable best practice documentation. Other such links are included at the end of section 8.

The most commonly recommended workshops were those run by the Software Sustainability Institute who will advise organisations on training curricular for computational skills and run the Software Carpentry and Data Carpentry workshops worldwide. The most recommended summer school was that run by CODATA and RDA at the International Centre for Theoretical Physics for the first time in 2016, with the aim of 'train the trainer' and with the materials openly available from the website.

Learning whilst at work from peers or supervisors was recommended by some respondents and general on-the-job training is also an effective way to learn. Another recommendation very relevant to any coder was hackathons and there are many of these available, with generic web lists of events widely available. The benefits of hackathons complements a comment on engaging mid-career researchers, section Figure 8, that working on one's own data would be an advantage and hackathons could be developed to align with this.

In the broad area of online training it is worth noting generic resources as forms of training, as previously discussed with DataONE. The UK Data Service is an example of a resource for social, economic and population data, with guides to datasets, topics, methods and software. Standardisation is also a form of training in best practice and one respondent recommended the CSA air quality project SEFIRA which has produced a common glossary to ease sharing expertise from different fields.

4.6. Discovering courses

The survey asked how respondents find new courses. The most popular method was organisational emails and intranets, followed closely by word of mouth and listservs. The least popular way was via journals. An interesting result was that using trusted or known websites was not a popular way to discover courses, despite many such websites being available as discussed in section 4.5. In the 'other' comments provided on course discovery were Twitter, conferences and newsletters (e.g. OpenAIRE, EUDAT).

A small number of people shared websites where they discover courses which are relevant to digital skills for global change research and most of these were both themselves providers of training and cataloguing related opportunities such as the Digital Curation Centre and the Australian National Data Service, although as mentioned in the previous section 4.5 such listings are often accompanied by training resources and therefore appear under the 'online' heading of section 8 (page 17).

4.7. Additional comments

At the end of the survey, respondents were able to add issues that had not been covered elsewhere. There were thoughtful and thought provoking comments.

- The scope of the survey and the enormity of the task was questioned. At least four of the respondents indicated that there are many challenges for 'global change research' and many issues explored by the survey are critical. One observed 'there are people for whom every issue raised in this survey is the most important so I fear many of the statistics returned here will reflect who responded'. Another added that 'our models are becoming very complex and large: it is becoming less clear which is limiting: conceptual model, data availability or computational power – and few people have the skills to make a robust assessment of all three'.
- Another theme of comments centred on research being reliant on teamwork. There were several elements, firstly that 'the days of individuals having all the relevant skills are well gone' and secondly that 'often the best trainers are colleagues' and researchers do not necessarily need external training. Thirdly, it was noted that it is not necessary for the leader to excel at all skills, rather 'a good research leader recognizes that having a [good] team with such talents advances the science'. And finally, acknowledging that forming and running interdisciplinary teams can be challenging. To address this challenge, a Belmont funded team is using software (www.mentalmodeler.org) which 'helps teams struggle through complex issues'.
- Several comments reinforced that reuse is often difficult but should be considered from the outset. Two roots of this were identified, one being the challenge of integrating natural, social, health and engineering science data thwarted by a 'lack of infrastructure and processes for sharing and managing data to promote reuse across teams, disciplines, borders and time'. The second difficulty is 'there is a tendency for data/tech people to work without direct contact with data users – the more the users influence the e-infrastructure design, the better' which is trans-disciplinary work.
- On training it was noted that 'a new common curricula for data education must be adopted and… should be technology and language agnostic' and that 'scientists need better programming skills and more scrutiny of the quality of [their] code'. A decline in courses on statistics at universities was indicated. A need to make digital skills part of science training workshops was deemed valuable. At least two respondents indicated that access to lists of relevant training would be useful. A final training comment was that skills should not be developed on too low a level as 'it would be very beneficial to automate as much of the research as possible; automatic generation of metadata to support data management, harvesting algorithms for data gathering [and] automatic migration to new formats'.
- Four specific technical issues were raised, worth reflecting on as emerging concerns. Firstly the analysis of text-based data for social sciences, that to develop effective policies it would be preferable to use text mining to make sense of people's motivations and behaviours rather than polls or surveys. Secondly, the issue of using data from several decades ago and how to faithfully digitize analogue sources such as paper or film, that specific and exact procedural standards need to be met to ensure reliable data, and this is not always understood by researchers. Thirdly, the issue of open research data architecture. And lastly, ISO 8000 for data quality.
- The themes of data curation and archiving were identified as missing from the survey. These are important areas, to be addressed in conjunction with AT2.
- The need to liaise with parallel and complementary initiatives to those of AT4 was indicated; synergies should be found with outputs of groups such as EDISON, ELIXIR and, in the case of RDA the interest group on 'education and training on handling of research data'.
- Comments were included on global inequalities, that 'global science is only done by the highly trained and equipped big boys'. A respondent from Africa attested that 'there is little global effort to bring Africa on board to big data'.

5. CURRICULA AND CERTIFICATION

Curricula are planning tools, identifying critical aspects of a subject and setting broad learning goals and level of achievement. Certification is an assurance of quality, be that training or an individual professional. The curricula and certification could designate levels of proficiency, including 'practitioner', 'mentor', 'trainer' and 'specialist'. For example, a 'train the trainer' course may not require a different curriculum but a higher level of achievement.

A fundamental consideration is the need to develop trust in courses, curricula and certification. Continuity is often key for repute and may involve considering a multi-year approach and raises the question of long-term upkeep of websites and other resources.

5.1. Developing curricula

Action Theme 4 of the e-Infrastructure and Data Management group of the Belmont Forum will develop curricula in 2017. Factors to be considered in curricula are:

i. The broad mission and priorities of the Belmont Forum
ii. Acknowledgement of the needs and preferences of the target audience, trainers and delegates, e.g. mid-career researchers
iii. Subject areas
iv. Within each subject area, list of topics with an indication of level of achievement
v. Broad direction on learning philosophies e.g. problem solving, interdisciplinary work, collaborative work, best practice, on-the-job training
vi. Distributing skills in a research team, individuals may not need all skills
vii. A discussion of assessment without being prescriptive e.g. from attendance measures more formal ideas such as certification
viii. Acknowledgement that feedback will evaluate curricula in order to develop

5.2. Curricula workshop Vienna April 2017

The workshop and subsequent supplementary meetings will lead to a defined curricula delivered in September 2017. Priorities for the workshop will be to iterate the skills challenges as identified in this report and distil into the curricula as a solution. The workshop will comprise a range of stakeholders, including research practitioners, data centre managers and informaticists and the Belmont Forum Steering Group. It would be desirable to include those that feature here as good sources of information, including Kevin Ashley (section 3.1), or a developer of the CODATA-RDA summer school or Software Sustainability Institute (section 4.5) or the Federal Big Data Research and Development Strategic Plan which informed the NSF call on CyberTraining (section 3.2).

5.3. Working towards 'Belmont Certification'

Preliminary work is investigating 'Belmont Certification'. It seems expedient to work with an organisation already involved in certification. Communication between AT4 and the RDA working group on 'Certification and Accreditation for Data Science Training and Education' is underway, to find out how certification can be established and whether the data science certification could be extended to Belmont requirements. A discussion has taken with BCS the Chartered Institute for IT who have 80,000 members worldwide. BCS have a skills framework and learning capability model for technical roles, they would review the Belmont curricula and work with the Forum to produce a scheme. The disadvantage of the BCS, in addition to the cost, is that the Belmont Forum may lose some control over the process. It may worthwhile to explore the CMMI Institute's Data Management Maturity (DMM) programme, used by the American Geophysical Union (AGU). The potential model of the Copernicus Academy was discussed in section 0.

6. LESSONS LEARNT, CONCLUSIONS AND RECOMMENDATIONS

6.1. Lessons learnt: reflections on producing this report

Executing the survey

- Issuing a survey seemed a good decision as an efficient way to gain views of many people on specific topics, reaching a wider audience than Belmont Forum associates.
- The skills gap survey questions are a good resource in themselves. The survey had seven excellent reviewers which refined content and produced a good framework.
- The vast majority of respondents completed the entire survey and put thought into replies where free text was available. Optional free text responses generated useful information but time consuming to analyse and was almost too much to consolidate within resources.
- On examining the free text responses, there was a vast array of surveys, training material and courses suggested and with very limited repetitions or overlaps. This diversity in responses mirrors the disparate nature of the skills gaps and remains a significant challenge for establishing coherent curricula.
- The question in the survey on the skills of 'research group leaders' was difficult to word and a number of responses were all middle column (i.e. expressing no preference). If this survey were re-launched, wording of this section should be reviewed and modified.
- The survey was distributed widely, to an appropriate audience reaching many data scientists, but lacked respondents in data curation, librarianship and social sciences.
- For maximum data capture, the survey could have been open for longer and promoted in waves. Numerous tweets and a blog was produced, but direct email and email lists were the main advertisements. Publicity was good but could be enhanced.
- The use of Google Forms for the survey had version control issues, some direct edits were accidentally made to the 'master' form. This confusion was partly a consequence of several reviewers, many of whom had not used Google Forms prior.
- Respondents were asked for contact details and about half supplied email addresses. AT4 does not have the resources to follow up with respondents. Therefore, none of the email addresses were used and, since it creates sensitive identifiable information, perhaps collection of email addresses was not necessary.

Other observations relevant to AT4

- Training is often cited as a need, but exact subject areas and the best method of delivery is always significantly more difficult to define. This report provides evidence for the broad topics. An alternative method to identify and address skills gaps (e.g. the CODATA-RDA summer school) is to use expert judgement to construct the curriculum and course, which is then run by respected organisations and individuals.
- The curricula developed at the workshop will not carry the level of detail of projects such as EDISON data science curriculum. The AT4 work is simply not resourced to that level.
- To train and upskill mid-career scientists is part of AT4 and how to do this effectively needs discussion, especially on the use of e.g. secondments and on-the-job training.
- It might be possible to make courses that are self-sustaining. Data finding, analysis, complexity and management have relevance to data intensive work in industry and thereby provide revenue sources for some courses. Funding may be required to establish courses but not be necessary long term.

6.2. Conclusions

The topics identified by the survey are consistent with existing information on skills gaps.

Training can take many forms, which sometimes overlap, including attending courses, secondments, conferences, discovering web-based tools, professional certification, hackathons and on-the-job. Training is one of a range of ways to influence change, which also include messaging, awareness through communication and incentives. Incentives can include funding projects, internships or secondments that ultimately have cross-pollination and digital skills development as outcomes.

The aims and corresponding conclusions:

- To inform discussions on the digital skills for global change research. Developing, disseminating and presenting the survey provides up to date information to allow further development of plans for the curricula workshop. The raw survey results are a resource that will be made openly available.
- To collate information on priority skills. The survey has indicated that practitioners recognise a number of needs and the topics are listed in the recommendations below. These needs are consistent with existing information on skills gaps.
- To list training activities, which are provided in section 8 (page 17) and discussed in section 4.5.
- To identify and learn from existing reports and surveys. Two documents stand out, firstly from Kevin Ashley on data management (section 3.1) and secondly the Federal Big Data Research and Development Strategic Plan which informed the recent NSF call on CyberTraining (3.2). The two most recommended courses were from the Software Sustainability Institute who advise on training curricular for computational skills and the CODATA-RDA summer school. Individuals from one or more of these projects should be involved in developing the Belmont Forum curricula.
- To highlight training relevant and wider issues. The survey questions asked both the 'largest challenge in data use' as well as the 'research habits and skills that need most improvement'. These questions produced priority training areas which are not purely digital nor domain science nor data science, but are linked with the Action Theme 2 on data management and policy and have more cultural impediments than technological.
- To recommend next steps on informing curricula to be developed later this year and moving to a 'Belmont Certification' of courses

There are many organisations or projects which are community portals offering data sources, documentation and a range of training resources combined in one website. Such portals exist for social sciences, data curation, information scientists and earth observation, but not for such a wide-ranging topic as digital skills for global change research. These portals are great resources, but make it difficult to distinguish between activities and to generate a catalogue of courses. Additionally, courses on offer vary from year to year and an excellent summer school one year many not run in the following year. The training list included in section 8 is a useful start, but would need to be expanded and reviewed periodically, so an automated system of searching the internet may be a more desirable approach, with the 'recommendation' coming from prioritising certain portals for listings such as DataONE or the Digital Curation Centre.

6.3. Recommendations

i. On a training catalogue: work with an organisation who has an existing automated tool for collecting and searching relevant training activities, ideally not one that requires providers or community to upload the events. ELIXIR would be desirable since their code for this is open source.

ii. On the curricula: focus on the priority areas identified as:
   - Data complexity
   - Data standards (relevance to AT2)
   - Data discovery, finding relevant and potential data sources
   - Data management (relevance to AT2)
   - Overcoming barriers to data sharing, including cultural and interdisciplinary issues
   - Improving programming and data analysis workflow
   - Improving computational and numerical analysis skills

   Simultaneously aiming that these topics:
   - resonate with their intended audience
   - remain relevant at least in the short to medium term e.g. up to five years
   - are part of a rationale of open data and enabling reuse of data.

iii. On the curricula: acknowledge that training can take many forms and does not need to be restricted to formal courses (the curricula may be training format agnostic).

iv. On certification: communicate with the RDA working group on 'Certification and Accreditation for Data Science Training and Education' and to keep an open mind on other certification formats and relevant providers.

7. ACKNOWLEDGEMENTS

This work was funded by the UK Natural Environment Research Council

An array of people in the Belmont Forum e-Infrastructure and Data Management group were of great help in iterating the survey structure and detail and their comments produced a clearer and more useful questionnaire. Thanks to Bob Samors, Tina Lee, Mark Thorley and Jean-Pierre Vilotte. Hugh Shanahan of Royal Holloway and Fiona Murphy of Murphy Mitchell Consulting added their valuable perspectives and adjustments to questions.

My thanks also go to the organisations who kindly circulated the survey, including AGU, Future Earth, EarthCube, GEO, DCC, OeRC, IEA, RDA, OKF, University of Reading, CODATA, NERC, CEH and many more, and of course the individuals who spent their time completing the survey.

In the production of this report my thanks go to Rowena Davis of the Belmont Forum coordination group for both assistance on the survey itself and for the analysis on the information provided on other relevant surveys, section 3.1. Thanks for review comments go to Fiona Murphy, Robert Gurney, Bob Samors, Carrie Seltzer and Katie Kinsey, also to NERC comprising Katie Tyrell, Matt Dobson and Mark Thorley.

8. RECOMMENDED TRAINING ACTIVITES

ONLINE	
ANDS	23 Research Data Things, resource for managing data
CMMI Institute	Enterprise data management training and certification
CESSDA	Training on research data management in the social sciences
DataONE	Supporting discovery of Earth and environmental data
DPC	Digital preservation, webinars and resources
ESA	Resources and materials from workshops
ESIP	Data Management Short Course for Scientists
FOSTER	Online courses in Open Science, Open Access, policy & legal
IASSIST	Community repository, sharing resources in social sciences
GEOCAB	Earth Observation, community posted resources
MANTRA	Managing digital data for research projects
MELODIES	Linked and environmental data, visualization
MetEd	Resources for geoscience including meteorology and climate
Research Data Alliance	Data management, repository certification, open science
RRI	Responsible Research & Innovation, online tools, workshops
UK Data Service	Datasets, topics, method and software guides
MOOCs	
Illinois Urbana-Champaign	Data Mining
Illinois Urbana-Champaign	Data Visualization
Johns Hopkins	Data Science
Johns Hopkins	Statistical Inference
North Carolina Chapel Hill	Research Data Management and Sharing
Stanford University	Machine Learning
PRACE	Managing Big Data with R and Hadoop
University of Southampton	Introduction to Linked Data and the Semantic Web
SUMMER SCHOOLS AND WORKSHOPS	
Alpbach (Austria)	Space science and engineering, 41st year
CODATA-RDA	1st run 2016, Unix, R, visualisation, data management
CEDA (UK)	Scientific computing, resources available online, last 2015
CSC (Finland)	High Performance Computing, programming & optimization
ESA - SEOM	e.g. Remote Sensing of the Cryosphere 1st held 2016
Hartree Centre (STFC)	Including summer school on big data, visualization and HPC
ICPSR	Quantitative Methods of Social Research,
NASA	e.g. Utilizing Renewable Energy
NCEAS	Open Science for Synthesis, data science and management
Software Sustainability Institute	Software and Data Carpentry courses
UKDA	Data service, with online resources too
University of Essex	Social Science and Data Analysis
University of Essex	Big Data and Analytics
University of Oxford	Digital Humanities: text processing, linked data, management

PROFESSIONAL CERTIFICATION	
DAMA International	Association for information professionals
COURSE LISTINGS	
DCC (UK)	DCC workshops, reference materials and lists of resources
NCRM (UK)	Methods for social scientists; statistics, Hadoop and more
PHIDOT	Forum. Population modelling, statistical inference

Appendix A:

Skills Gap Survey

Skills for e-infrastructures and data management in global change research

A skills survey which should take 10 minutes to complete. We would appreciate your input by 22 November.

* Required

Background Information

The Belmont Forum e-Infrastructures and Data Management Project (http://www.bfe-inf.org/) recognizes that, for global change research to flourish, a range of skills are necessary, increasingly so in the broad area of data intensive digital skills. Through this short survey, the Belmont Forum would like to establish where skills gaps exist or are emerging. Results will be anonymized.

Global change research encompasses the interdisciplinary investigation of the natural world and how it is changing, from climate change to air quality to ecology and the human interaction with the environment. Global change research is increasingly data intensive, requiring significant computing resources and efficient techniques for effective analysis. This survey is aimed at enabling capacity building in the requisite digital skills for those already working in global change research.

The Belmont Forum (https://belmontforum.org/) is a group of the world's major and emerging funders of global environmental change research. It aims to accelerate delivery of the environmental research needed to remove critical barriers to sustainability by aligning and mobilizing international resources. It pursues the goals set in the Belmont Challenge by adding value to existing national investments and supporting international partnerships in interdisciplinary and trans-disciplinary scientific endeavors. (Read the full Belmont Challenge at https://belmontforum.org/belmont-challenge).

Respondent Information

1. **Employer or affiliation** *

2. **Primary research funder** *
   Mark only one oval.
   - ◯ Government/Civil Service
   - ◯ University
   - ◯ Private/commercial entity
   - ◯ Multilateral organization (e.g. World Bank, UN)
   - ◯ NGO/philanthropic
   - ◯ Other

3. **What is your field of expertise within global change research? (Please select no more than two.)** *
Check all that apply.

- [ ] Atmosphere, Climate, Cryosphere, Hydrology &/or Oceans
- [ ] Biology, Biodiversity, Agriculture &/or Ecology
- [ ] Chemistry &/or Geochemistry
- [ ] Computing, Computer Science &/or Data Science
- [ ] Geology, Geophysics &/or Geoscience
- [ ] Mathematics &/or Statistics
- [ ] Social Sciences
- [ ] Other: _____

4. **What is your primary role within your field?**
Mark only one oval.

- ◯ Outreach/communications
- ◯ Researcher
- ◯ Software engineer, developer, programmer or data engineer
- ◯ Trainer/lecturer
- ◯ Other
- ◯ Project manager or line manager
- ◯ Data specialist/manager at a data repository

5. **With respect to data use and management, what is the level of technical support and computing infrastructure available to you in your work?**
Mark only one oval.

	1	2	3	4	5	
very poor	◯	◯	◯	◯	◯	excellent

6. **In your work, how multi- or interdisciplinary are the data you use and the people you regularly work with?**
Mark only one oval.

	1	2	3	4	5	
not at all	◯	◯	◯	◯	◯	very diverse

7. **How many years have you been working professionally (including graduate work)?** *
Mark only one oval.

- ◯ Early Career (1-5 years)
- ◯ Mid Career (6-20 years)
- ◯ Been Around (20+ years)

8. **Where do you live and work? (Please use "otherif your funder and/or field work is elsewhere.)**
Mark only one oval.

- ◯ Africa
- ◯ Asia
- ◯ Australasia
- ◯ Europe
- ◯ North America
- ◯ South America
- ◯ Other: _____

Training Opportunities

Data intensive digital skills which are relevant to global change research apply to the full data life-cycle, from discovery to analysis to insight to management to publication.

9. **If you teach, supervise or manage researchers, please indicate their education level:** *
Mark only one oval.

- ◯ Post-PhD
- ◯ Masters/Doctoral candidates
- ◯ Undergraduates
- ◯ I do not teach, supervise or manage researchers

10. **In your field, please list any training in data intensive digital skills or global change research that you value e.g. Masters', summer schools, online. Include links if possible.**

_____

_____

_____

_____

_____

11. **How do you discover new and upcoming courses? (Check all that apply.)** *
Check all that apply.

- ☐ Search engine (e.g. Google/Bing)
- ☐ Organizational intranet or internal emails
- ☐ Word of mouth
- ☐ Journals or scientific publications
- ☐ Listservs or social media from learned societies
- ☐ Trusted/known websites (please check "other" field and insert URLs)
- ☐ Other: _____

Identifying Relevant Skills

12. **How would you describe the largest challenge you encounter in your data use?** *
Mark only one oval.

- ◯ Lack of data standards and exchange standards
- ◯ Finding relevant existing data - knowing what's out there
- ◯ Data access and file transfer
- ◯ Data volumes
- ◯ Data management and storage
- ◯ Data complexity
- ◯ Dealing with multiple data types
- ◯ Other: _____

13. **What do you consider to be the three digital skills most necessary for global change research? (Select up to three).** *
Check all that apply.

- ☐ Programming (creating clear and robust code)
- ☐ Data processing and analysis
- ☐ Statistical methods (e.g. regression, Bayesian, data assimilation)
- ☐ Using statistical or numerical models
- ☐ Developing statistical or numerical models
- ☐ Data mining
- ☐ Visualization
- ☐ Use of new facilities and techniques (e.g. cloud computing, machine learning)
- ☐ Use of big data technologies (e.g. Hadoop, Spark, Storm)
- ☐ Use of High Performance Computing
- ☐ Use of data bases and data base technologies
- ☐ Data management
- ☐ Other: _____

14. **In your experience, what are the research habits and digital skills that need the most improvement? (Select all that apply.)** *

Check all that apply.

- [ ] Overuse of spreadsheets
- [ ] Poor programming style or data analysis workflows
- [ ] Lack of computational and numerical analysis skills
- [ ] Inefficient numerical analysis or lack of awareness of advances in these technologies
- [ ] Inability to process large amounts of data
- [ ] Lack of well-maintained libraries and codes
- [ ] Lack of awareness of relevant and potential data sources
- [ ] Reliance on old fashioned programming languages or legacy systems
- [ ] Reluctance to share data or models
- [ ] Lack of multi- and interdisciplinary collaborations
- [ ] Poor understanding of or compliance with funders' research data management requirements
- [ ] Other: _____

15. **In your opinion, do research group leaders have appropriate digital skills in the following areas?**

Mark only one oval per row.

	1 very poor	2	3	4	5 excellent	Don't know
Data discovery and reuse	◯	◯	◯	◯	◯	◯
New analysis or numerical methods	◯	◯	◯	◯	◯	◯
Computing and new technologies	◯	◯	◯	◯	◯	◯

16. **In your opinion, what are the best ways to engage mid-career researchers in developing their skills? (Select up to three.)** *

Check all that apply.

- [ ] Messaging from funders
- [ ] Briefings by senior research managment
- [ ] Make the training activity a required part of career development and progression
- [ ] Making recoginition of digital skills part of career progression
- [ ] Providing full financial support for training activities e.g. including travel
- [ ] Target the training activity as an opportunity to meet co-funded project colleagues face-to-face
- [ ] Provide time-flexible or on-demand training activities such as webinars or online courses
- [ ] Provide ways to try new tools and technologies via incremental engagement e.g. demonstration software
- [ ] Other: _____

Final Thoughts

17. **Please list any relevant surveys or analyses on digital skills carried out in the last three years of which you are aware. Include links if possible.**

18. **Please share any other ideas or experiences on digital skills and global change research skills that were not covered in the questions above.**

19. **May we contact you regarding your survey responses? Your details will not be passed to third parties and any results presented from this survey will be anonymized.** *
Mark only one oval.

- Yes *Skip to question 20.*
- No *Stop filling out this form.*

Contact Information

20. **Please enter your email address.**