October 2016

Developing Skills for Managing Research Data and Software in Open Research

Kevin Ashley

wellcome

Cite this as: 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

Developing Skills For Managing Research Data And Software In Open Research

A DCC Report for the Wellcome Trust

Executive Summary

Despite a widespread acknowledgement that we face a skills crisis in relation to data, both in the field of research and more widely in society, there is less detailed evidence available than we might hope for on the details of the problem. Only one quantitative study dating from 2011 has identified both the number of skilled people required and the current ability to meet the demand. Parallels from other domains in the past show that the growth in graduates required is a challenging but not unachievable target, and that the gap could therefore be closed within five years but there is a great deal of uncertainty around all of the numbers involved.

There are a number of analytical frameworks which have been developed or are under development which categorise the skills required by specialist data scientists, other researchers and those who support them in library and IT roles. These frameworks help to classify existing training and educational resources, and hence to identify gaps that require to be filled. It seems likely that, when ready, these data science skills and capabilities frameworks under development will gain widespread acceptance, but their current absence has not prevented universities around the world from creating data science courses at many levels.

There are many existing initiatives around the world producing and delivering training for online and classroom-based consumption. The most successful have demonstrated their quality through growth, reuse and positive feedback from users. Far more material is at present directed at data skills than at software skills, and there is also arguably a preponderance of material directed at research support staff as opposed to researchers themselves. Sustainability questions exist regarding the very successful software and data carpentry initiatives, which rely on volunteer labour for delivery. Community events that bring together otherwise-isolated professionals are important in skills development, and funders should seek to support the relatively modest costs involved.

Career structures vary greatly. There are long-standing problems for the researcher with particular strengths in data analysis, management or software development which new initiatives such as CRediT are only beginning to address and which will only partly solve the problem. There are fewer concerns for those who see themselves primarily as data scientists or who work as data managers within university services or data centres. Indeed, the problem for the research community might be that, at present, the data scientists see their ideal career paths as being outside the research domain.

Coordination amongst funders, and clarity about that coordination, is a clear recommendation which has delivered results already. Universities are asked to do more to educate their researchers and provide access to appropriate services. Researchers themselves also have a responsibility – and a growing obligation – to utilise the training and infrastructure appropriately.

A review of the landscape - introduction

In looking at the literature and talking to those close to the issues, we concentrated on the 4 questions suggested by the Trust in its request for proposals:

> **[A] What evidence exists of potential gaps and needs for skills and capacity in data management (both in terms of current and anticipated future demands) - both among**
> **(i) specialist data scientists/managers; and**
> **(ii) the broader research community?**
>
> **[B] What initiatives and models exist to develop skills in data and software management in both of these groups both at institutional, UK and international level? And what evidence exists on their effectiveness?**
>
> **[C] What career structures exist for data scientists in research institutions, and what is the broader employment market context for these skills?**
>
> **[D] What key longer-term actions have been proposed for funders and institutions to build the skills and capacity in data management to meet future needs?**

There is little quantitative evidence for [A], and the little that exists is not specific to the research domain. It is, however, possible to draw conclusions for research from this evidence. Although there is a lack of quantitative evidence on the size of skills gaps, there are a good number of analytical frameworks that help us understand the types of skills required for different roles and the ways in which those skills might be acquired.

There are a number of sources of good evidence to address the first part of question [B]. There are many initiatives and models for training researchers and specialists who work with them in data and software skills, and a number of groups have sought to itemise these. There is very good coverage of the availability of open training resources and of international initiatives to provide researchers and others with relevant skills. There is patchier evidence on the takeup of these within individual institutions. The second part of question [B] is poorly addressed by the current state of knowledge. We have little evidence on the long-term effectiveness of these initiatives; at least in part, because it is too early to tell. What we do have is evidence of the popularity of many of them, which is high, as is the level of satisfaction of those who attend training events and webinars.

The evidence available for question [C] is, at present, more anecdotal than quantitative. It tells us that there are two divergent stories in terms of career paths. One offers greater potential rewards but is likely to take the practitioner out of the research domain; the other allows the practitioner to stay within the research domain, but in a career path that is perceived as being undervalued and under-rewarded in comparison to peers who took the other path. The first route is available to those characterised as data scientists – whose data and software skills are at the fore, with domain knowledge taking a secondary role. The second is the route more familiar to domain specialists whose primary role concerns software or data. There are, however, promising signs that institutional promotion criteria are starting to take open science skills into account and the emergence of mechanisms such as the CRediT taxonomy may be of relevance to those whose major contributions are in data and software management for research.

In addressing question [D] the recommended actions for funders and institutions are distinct in quality and quantity. There are more concrete recommendations directed at institutions than at funders. We have also chosen to summarise what some funders are doing already. In some cases it is sensible for more funders to replicate these activities, ideally in a coordinated fashion, so as to create a consistent picture between countries and between related research domains. In other cases we note that an action, carried out once, does not need to be replicated by other funders. Some studies, such as that produced by the EDISON project [8], fall into this category.

On roles

Throughout this report it is necessary to distinguish between at least three broad groups of people. The first are **data scientists**. These will have a range of specialist skills in statistics, data analysis, machine learning, database systems and software development. They are likely to self-identify as data scientists first, and are less likely to be closely attached to a specific domain of application.

The second group is those whose career has been within a specific research domain but who, over time, have become specialists in data management or software. They may not self-identify with these roles on a permanent basis, even when the teams that they are part of see their role as essential to the research. They are far less likely than the data scientists to recognise that their skills may be generic, with an application beyond their own research field. We will use the term **data managers** to refer to this group in the report, although their actual role may be much broader, and may have more to do with software management than data management.

The final group is a heterogeneous group of **research support staff** who work in groups such as library or IT services providing services and support to researchers in their institution. This group has recognised for some time that dealing with research data (and, much more recently, research software) is an increasingly important part of their function, and that they need to acquire relevant skills.

We identify these three broad characterisations because many of the questions have different answers for each group, but the distinctions between the groups remain relatively constant for each question. It is, therefore, a useful if somewhat blunt and imprecise classification. In places, we will refer to other groups, particularly to **researchers** as a whole.

Gaps and requirements

One of the most widely quoted numbers with regard to the skills gap comes from a 2011 study by the McKinsey Global Institute [3]. It predicted two large gaps between demand and supply in the USA alone by 2018. The smaller gap, of 140,000 to 190,000 people, is of those with ‘deep analytical skills’ that we might otherwise refer to as ‘data scientists.’ The larger gap, of 1.5 million people (in the USA by 2018) is of managers and analysts ‘with the know-how to use the analysis of big data to make effective decisions.’ In a research context, we might think of these as researchers, but they will also include some of the data managers and also those outside research domains where our research is intended to have application and impact. These are large numbers and they apply only to a single country in what is now, five years on, the relatively near future. The figures have been quoted in a number of contexts as evidence for the size of the skills gap facing the research community. Barend Mons, chair of the High Level Expert Group on the European Open Science Cloud, has done so on a number of occasions (see [4] for an example) and they are part of the justification for the European Data Science Academy [5]. However, the numbers need to be treated with caution if we are to use them as a basis for understanding the research gap in a UK context.

The most obvious problem is that the McKinsey report is not specific to the research sector, but addresses demand across society. Similar reports, such as that from the Centre of Economic and

Business Research (CEBR) for the UK [7] focus exclusively on the commercial sector. They are still relevant to research, for two reasons. One is that we can make a starting assumption that the change in demand for these skills in the sectors examined by McKinsey and others is similar to the change in demand that will occur in the research sector. The other is that even if the research sector's requirements are static, increasing demand from elsewhere will reduce the availability of data scientists to the research sector.

The second concern with the McKinsey report is that the underlying data and the models used were primarily for the USA. Some sections of the report do contain data on other nations, however, and this provides some useful information which may help us extrapolate for the UK. The gap of 140,000 to 190,000 data scientists predicted by McKinsey comes from two other numbers – a raw increase in the number of jobs for data scientists of approximately 160% over 6 years, and an increase in the supply of those with relevant skills of about 100% over the same period. That increase comes from two sources – a larger number of graduates being produced in the USA, and a greater level of immigration of skilled people. Both of these would present greater challenges in the UK. At the time of the baseline figures used in the report (Eurostat 2008), the UK's per-capita rate of production of such graduates was already one of the highest in the world, with only Poland and Romania achieving higher rates. The capacity for increase in new graduates is thus arguably lower than it is in the USA. In the current political climate, it is also unlikely that the UK will be able to use immigration to the same extent to plug any skills gap in this area. These factors, combined with the draw of high salaries in industry and in overseas markets such as the USA, could mean that the UK's skills gap for these highly-skilled individuals could be even larger. (The CEBR report, which looked exclusively at demand in the commercial sector, predicted 58,000 new jobs in this area between 2012 and 2017, but did not examine increase in supply of graduates.) Nonetheless, initiatives have been undertaken by governments, funders and universities to address increased demand, and these are identified in the next section.

It may be thought that achieving still higher rates of growth in the production of data science graduates is an unattainable target. An example from another field may illustrate what is possible when student demand exists, even without any significant funder support. In a phenomenon often dubbed 'the CSI effect', the numbers of students on forensic science courses showed huge rates of increase between the late 1990s and the decade which followed. [1] UK student numbers rose by 158% over a five-year period from 2002. Student numbers on a course at West Virginia University rose from 4 in 1997 to over 500 in 2006. Those represent annual rates of increase of 21% (UK) and 71% (West Virginia.) This has been achieved despite the fact that the job market for these graduates is at best static, and in the UK has been in decline. Universities were able to respond to student-driven (market) demand, and were criticised in some quarters for doing so. By contrast, the annual rate of increase needed to deliver the results anticipated by the McKinsey report is 12.2%. There is clearly room for more rapid growth if universities are confident that their courses will be filled.

One further estimate is worthy of note. The High Level Expert Group report on the European Open Science Cloud [2] estimates that 500,000 people with this level of expertise will be required 'within a decade' – i.e. by 2026[1]. This is to support a research community in Europe that is currently 1.7 million strong, and a wider innovation community of 70 million people. The report makes clear that

[1] The draft report was widely circulated in early 2016 and made available for public comment in April 2016 at the address noted in the references section. However, at the time of writing (August 2016) the draft had been withdrawn from that page and had not been replaced with a final version. These observations are based on the public draft.

most of those consulted in its preparation felt that, without action, this number would not be achieved. What the report does not do is identify the size of the gap.

The second gap, of ‘data-savvy’ individuals, is of greater interest. In research, this will often mean those researchers with a sufficient understanding of these techniques to be able to know when they can be utilised to further their research goals. In some cases, the researcher’s own knowledge is enough to apply the techniques, especially when they are encapsulated in tools and services. In others, the researcher is well-placed to have collaborate productively with an expert data scientist who is not a domain expert in the researcher’s field. This gap can be addressed in multiple ways – during education but also later in the career as part of continuous professional development activity. In the following section we identify a number of initiatives which are addressing this gap.

More work has been done to examine skills requirements in more detail analytically, without quantifying the size of the requirement. Some statements assert the need for such skills without giving much further detail. These include a recent report from a meeting of researchers at the University of Cambridge [10] which notes that “…it requires effort and skills to make research open, re-usable and discoverable by others” and calls for discipline-specific training in data management and open research to address this problem. Researchers noted that the current skills gap has two effects; the data they find is often not as reusable as it might be, and they themselves lack the skills to make their research data and software reusable in turn. Similar sentiments are expressed by the UK Research Concordat on Open Research Data [9] whose 9th principle is that “Support for the development of appropriate data skills is recognised as a responsibility for all stakeholders.” The Concordat is more specific on the responsibilities of institutions to ensure that researchers can acquire these skills, and this is addressed in a later section.

A 2015 paper [11] by Liz Lyon and Aaron Brenner of the University of Pittsburgh goes further in categorising the different roles and associated skills required for open research. It lists data analysts (a group corresponding to ‘data scientists’ in this report), data archivists, data engineers, data journalists, data librarians and data stewards/curators, and identifies activities associated with each role. It also contains a comprehensive list of earlier work identifying skills requirements as well as associated capabilities of institutions and research communities. This concept of community capabilities is an important one, and is overlooked by some other analyses. It is not enough for researchers to acquire skills. They can only use those skills effectively in an environment which supports the exercise of those skills and which provides the services and other complementary skills necessary for their effective use. For example, it is not enough for a researcher to know how to prepare data for effective reuse from a domain data repository if no such infrastructure is available to them.

More detailed work has been produced on the skills requirements for data scientists. This includes work by the EDISON project[2] which has produced drafts of a data science competence framework[8] as well as a model curriculum for data science and a body of knowledge document, both also currently available in draft for comment. EDISON’s work is particularly relevant for this report because it looks at data science with research applications as the primary use case, with business applications treated as a special subset or alternate domain. The competence framework is the key document, with the model curriculum, body of knowledge and eventual professional certification process all building upon it. It builds extensively on previous competence frameworks that are used in the computer science and IT professional fields. For example, the EDISON framework utilises three high-level groupings of competencies from the NIST framework, and adds two more: “Data

[2] www.edison-project.eu

Management" and "Scientific Methods." It is also aligned with the European e-competence framework and standards from other bodies such as ACM. The report's authors express an intent that this will increase acceptance of the framework by educators and students alike.

It is too early to say whether frameworks such as that from EDISON will help us better understand skills gaps in future. They have been designed primarily to define curricula for educators. To use them to understand skills gaps would require those attempting to recruit people with these skills to assess job descriptions and roles against the pre-defined frameworks. This practice is common in some areas of the IT industry (as well as other areas of the job market) but is not widespread in research.

Finally, some understanding of the skills thought to be required can be gained by looking at schemes used to classify existing training courses and materials. These are particularly strong in relation to research support staff, partly because of the relative wealth of material produced that has been aimed at that group over the past decade. Again, these mechanisms do not quantify the skills gaps, but they provide an indirect qualitative indication of where some people perceived the skills gaps to be. Amongst these, the classification produced by the FOSTER[3] project [12] is the most complete, and itself builds on an earlier scheme used to classify research data management training materials produced by the JISC-funded *Data Management Skills Support Initiative – Assessment, Benchmarking, Classification* (DAMSII-ABC) project [13]. The FOSTER analysis shows that whilst most of the training resources identified are multi-disciplinary, of those which focus on specific disciplines humanities has the greatest number (almost 50% of the discipline-specific material), followed by social sciences (30%), natural sciences (17%) and engineering and technology (1%). One other classification of note carried out by FOSTER relates to the level of knowledge imparted by the training. On a three-point scale of:

1. Introductory – aware of;
2. Intermediate – able to;
3. Advanced – applies;

52% of material fell into the introductory category, and only 12% into the advanced category. The opinion of the FOSTER project, and of the authors of this report, is that this distribution does not accurately reflect the nature of the skills gap, and that there is a need for more advanced materials aimed at all of the target groups and for more material aimed at discipline groups such as engineering which are presently under-served.

Initiatives and models

A wealth of initiatives, past and present, address different aspects of the skills gaps. As the previous section on the size of the gap indicates, few have been informed by detailed information on the size of these gaps. Nonetheless, a number of the initiatives described here have produced outputs, in the form of training courses and training materials, which receive consistently high ratings by those making use of them. We can conclude that they may not be filling all of the gaps, but those that they are filling are real and are being addressed well.

We begin with what might be described as meta-initiatives: activities which themselves aim to collate, identify, classify and/or promulgate the work of others. The FOSTER project, referenced in the preceding section on skills gaps, is of greatest relevance for this study. A two-year project funded by the European Commission under the FP7 programme between 2014 and 2016, FOSTER's focus was

[3] www.fosteropenscience.eu

the development of Open Science skills in the research community, characterised by the project's funders, the European Commission, as a 'downstream activity'. It was complemented by an 'upstream activity' project, PASTEUR[4], which addressed Open Science from the perspective of funders and research organisations. Both projects were led by the same team at the University of Minho and shared a number of partners, ensuring that their work was closely linked. FOSTER identified existing training resources for researchers in open science and developed and delivered new training to address gaps or deficiencies in existing materials. A portal, accessible via the project website, allows these resources to be navigated via the classification scheme [12] mentioned in the previous section. Since the project's remit was Open Science in general, not all of the resources are directly relevant to this report. Some, for instance, are about making a general case for Open Science, and others address understanding of relevant funder policies. The bulk of the resources described deal with data management and related skills; FOSTER is relatively poor in material which addresses software skills for Open Science, and its classification scheme does not explicitly address them.

In addition to collating and creating training materials, FOSTER had an active dissemination and uptake function. Institutions throughout Europe could bid for funds to run a local FOSTER workshop or training event; this would then be delivered by a combination of staff from the project and colleagues local to the institution. The intent was to encourage reuse and adaptation of the materials in local contexts, and there is some evidence that this was a successful strategy. Material has been translated into other European languages and further training days run locally without European funding after a visit by FOSTER staff. Variations on this model for increasing training uptake have been used successfully by other initiatives described in this report.

The second phase of JISC's Managing Research Data programme, which was undertaken in the UK between 2012 and 2013, included a strand dedicated to producing training materials in research data management. The materials were produced and trialled by individual projects, with oversight and coordination being provided by the DAMSII team at the Digital Curation Centre [13]. Most of the packages produced were intended to be discipline-specific; many were intended to be incorporated into the training offered to early career researchers, perhaps by being embedded into existing research methods courses. Disciplines covered included social sciences, health sciences, social anthropology and performance arts. The coordination project ensured that the material was classified against a standard framework and collected in a single repository for later reuse. This was initially JORUM, but the imminent demise of that UK service prompted a relocation of the collection to Zenodo. That collection is no longer specific to the Jisc-funded programme, but is promoted by the RDA Interest Group on Education and Training in the Handling of Research Data (RDA IG-ETHRD) [16] amongst others.

The classification of skills that was utilised by DAMSII-ABC and FOSTER informed – and was informed by – a more general effort by Vitae to produce an information literacy 'lens' on its influential researcher development framework. [14] Although relatively high-level, this work identifies a number of skills relevant to open research, open data and open software, and relates them to a generic set of skills and competencies applicable to research in all domains. The information literacy lens expects, for instance that researchers should understand "the risks to research information/data over time and of operating in virtual environments" and should possess the ability to "assess and mitigate these risks." It includes specific requirements relating to open research processes: researchers should understand "the importance of information/data sharing and accessibility to maximise opportunity for collaborative research, further subject and enhance own profile" and also

[4] www.pasteur4oa.eu

“how sharing and making data accessible aids synthesis and facilitates new research.” It expects that researchers will understand “the need to manage, share and curate information/data ethically”. It does not explicitly mention skills related to software, but does include requirements that imply at minimum an understanding of how to use software effectively, and, in some domains, how to create it. These include an understanding of the techniques of data analysis and a number of references to the use of tools for analysis, sharing and interdisciplinary research which all recognise that the skills required need constant updating throughout a research career, as the tools and technologies themselves evolve. Explicit recognition of this need for ongoing professional development is absent in a number of the other models of skills described in this report. It is important, therefore, to view the Vitae framework not just as something which has informed later developments but one which has direct relevance to any future plans to develop research skills in the areas of data and software in open research.

The Research Data Alliance[5] (RDA) has an interest group dedicated to skills development, the Interest Group on Education and Training in Data Handling (RDA IG-ETHRD) [16]. This group at present functions primarily as a meta-initiative, to exchange and collate information on initiatives in training and education for skills needed throughout the data lifecycle. It has three aims which are of direct relevance to the issues being examined in this report, including credit and career paths. These aims are to:

- enable the setting of quality standards for appropriate education and training programmes aimed at researchers and the professionals that support them, at all career stages;
- encourage the recognition of data skills amongst employees, employers, and professional bodies.
- prepare the ground for practical applications applying these standards in educational environments

The interest group is still a work-in-progress, but has already enabled a more global view of existing initiatives than the other meta-projects described so far. Amongst its intended outputs are a categorised list of training materials with information about audience, type, duration and availability. This list is currently available as a Google spreadsheet[6]. There is also an ongoing RDA task force producing a list of skills sets for different actors in the research field: research librarians; research administrators; research infrastructure managers / operators; researchers. As with many RDA interest groups, it suffers not from a lack of expertise but from a lack of funded time from many of those involved to move the group’s work forward more quickly.

We now move to considering specific initiatives with directly useable or applicable outputs and results. Some are aimed primarily at existing professionals (researchers and research support staff) who did not have the opportunity to acquire the relevant skills in their pre-career education, or whose skills require updating. These take the forms most common to continuous professional development – online courses or short (one day to one week) on-site courses. Other initiatives are aimed at those still acquiring qualifications, at levels from undergraduate to doctoral.

One of the outputs of the JISC-funded RDM training programme is worthy of particular attention due to its continued impact. This is Research Data MANTRA [17], an online self-paced course on creating, managing and using research data, which was produced by the University of Edinburgh. MANTRA is aimed at both early-career researchers and information professionals, with different routes through the material for each group. It is openly licensed and the content has been developed

[5] Rd-alliance.org
[6] https://docs.google.com/spreadsheets/d/10RTW-nZk0x_mpQw2VAlttcc656MV9EeCaDe2lM4umb4

using open standards, allowing it to be embedded elsewhere in standards-compliant virtual learning environments such as Moodle. The initial development involved consultation with researchers in three disciplines – geosciences, social and political sciences, and psychology. The content has continued to be developed by the university since external project funding ceased five years ago, and it has been reused and adapted by a number of other institutions worldwide. One of these examples is as part of a data management MOOC [18], developed in association with the University of North Carolina at Chapel Hill, which has already been completed by hundreds of students since its launch earlier this year.

The Digital Curation Centre is also reusing material from the RDM training programme as part of its data management training programme [24]. This includes content aimed at research support staff, researchers at all career stages, and at data managers, with the intent that matching expectations will be created in each group as a result of the training. Once grant-supported, the DCC's training is now financially independent and self-sustaining and is provided on-site around the world as well as through open courses which can be attended by anyone. On-site courses are designed so that they can be used to build up local capacity to deliver such courses in the future.

Another training initiative produced in the Netherlands has also demonstrated its value through longevity, influence on others and adaptability. "Essentials 4 Data Support" [19] is aimed at those who must support researchers in managing research data and is now a product of the national initiative Research Data Netherlands. Its origins lie in an earlier course, Data Intelligence 4 Librarians, developed by a consortium of Dutch universities with a shared research data service (3.TU, now 4.TU[7]) That course influenced training developed in the UK and elsewhere.

One of the most successful initiatives of recent years has been the set of programmes which began with Software Carpentry[8] and has since spawned other initiatives including data carpentry and library carpentry. Software Carpentry is distinctive for a number of reasons. It was the first initiative to focus on the software skills needed in research, and is still one of few to do so. It relies a great deal on volunteer effort, and it has a growth and dissemination model comparable to that employed by FOSTER, but is more formalised and arguably more effective. In the four years between late 2011 and late 2015 (the latest data available), it ran 500 workshops, trained 16000 researchers and had built up a cohort of 450 qualified instructors [21]. It is now run by an independent not-for-profit foundation in the USA with financial backing from grant-giving organisations such as the Sloan Foundation, a number of universities in the USA, Europe and Australasia, large research organisations such as the Jackson Laboratory, and research infrastructures such as ELIXIR. Its financial model assumes that instructors give their time for free, although expect their other costs (travel, subsistence) to be covered by those wanting the training. There are questions over the long-term viability of such a structure, but there are no signs of it failing yet.

Data Carpentry[9] is a much younger offshoot of Software Carpentry, and shares the same model, many of the same financial backers, and many staff and instructors. Both organisations offer two-day workshops aimed at practising and early-career researchers to give them skills in data or software which they can then develop further themselves. Each is building up an increasing library of 'lessons' from which the two-day courses are constructed.

[7] www.4tu.nl
[8] Software-carpentry.org
[9] www.datacarpentry.org

The Software Sustainability Institute, a partner in Software Carpentry, has also undertaken a number of other actions with the aim of increasing skills and knowledge amongst software developers in research. These include a fellows programme [25], which funds individuals to attend events and then to spread knowledge of them amongst peers, and a number of events designed to build communities, including the Collaborations Workshop [26] and the Research Software Engineers conference. The community events address a problem noted by other organisations (such as the DCC) who work with data managers and research support staff with data responsibilities, namely that such individuals are often isolated in their institution without any local peers. Only in the largest laboratories or groups will there be a local community of research software developers or data managers. Skills development and acquisition depends not just on the availability of education and courses but on the ability to hone skills and acquire new ones through interaction with peers. Other organisations recognise the value of these peer networks in providing examples of good practice as well as advice on the value of future initiatives. One such example is the nascent group of Open Data Champions[10] announced by SPARC Europe in autumn 2016, which builds on its successful Open Access Champions model.

A joint RDA-CODATA group has been working to develop a summer school on data science and cloud computing with a particular focus on the developing world. The intended audience alone makes this initiative distinctive, and it has received backing from a number of other organisations including The World Academy of Sciences (TWAS.) As a summer school, the course is two weeks long, significantly longer than those offered by Software Carpentry and similar initiatives. The curriculum includes elements common to software and data carpentry, but also includes Open Science, research data management and curation, the use of cloud platforms and the use of packaged AI techniques. Only one pilot course has taken place so far, in August 2016, and the curriculum is likely to undergo further development. The pilot was extremely popular and anecdotal feedback from both attendees and instructors was extremely positive.

Inspired by the carpentry model, the Mozilla Foundation is creating a set of lessons aimed at the broader issue of Open Science, a term to which they give a broad interpretation which includes the development of open educational resources (OERs) [22]. The material is still under development and there is as yet no experience with delivery, so it is too early to be confident of success.

Universities around the world are responding to reports of a current and future skills gap in data science with the creation of masters and doctoral programmes in data science. Although initiatives like the EDISON project are still working to produce a concrete, widely-agreed definition of what a data science curriculum should contain, many practitioners feel that there is no time to lose, and that they have a sufficient understanding of the requirements to produce courses which will enable their graduates to succeed in a field where demand is likely to exceed supply for some time. Both observations (on the need for speed and the level of understanding) are broadly correct. There is already a good deal of commonality between curricula for data science courses in various parts of the world, even if there is a difference in emphasis and detail. Some courses focus more on computer science skills related to large databases, parallelisation or data transformation, whilst others place greater emphasis on the statistical skills required to draw reliable information from large volumes of often imperfect data. A smaller number may also offer the opportunity to learn data science within a particular disciplinary context, perhaps because of financial support from a specific funder, but most prefer to stress the general applicability of the skills learned. Some universities – Edinburgh is one example – go further and aim to embed data skills throughout undergraduate training as well, and a 2015 report from Universities UK [20] contains a series of specific recommendations on how this

[10] http://sparceurope.org/news_opendatachampions/

should be extended. These recommendations include a requirement that students be taught not only data analysis but also data management, based on an observation that this was generally lacking in existing provision.

The lack of education about data management applies more generally to the majority of existing curricula at masters and doctorate level, although it is a recognised requirement in the draft frameworks developed by EDISON and comparable initiatives. The existing courses focus primarily on only the second of the following scenarios:

- If I wish to acquire certain items of knowledge, what data should I aim to collect and what techniques should I use to analyse it?
- Given a particular data collection, what knowledge can I extract from it and how?
- Having extracted some knowledge from a data collection, how can I ensure that my data and methods are available to others (including my future self) either to understand my own work or to undertake new, unrelated work?

One gap in the programmes can be observed anecdotally by the surprise frequently expressed by those undertaking data-intensive research in industry, finance, health and many other sectors on the time consumed by data curation – making data fit for analysis – rather than on the analysis of data itself. This is a long-recognised truth which at present it seems is being rediscovered by each new cohort of data scientists rather than being inculcated during their education.

It is too early to be certain about the effectiveness of the majority of degree programmes in data science. They are addressing the skills gap, but we do not yet know to what extent they have closed it. Those that have been running long enough to produce graduates are seeing very high levels of employment amongst those graduates, but this is to be expected at a time when demand and supply of skills are so ill-matched. We can be reasonably confident that they are doing good work, although there is undoubtedly room for improvement in the curriculum which will require continual attention in the coming years. Some defects (such as those relating to data management) are already recognised and should be dealt with; others will require firmer agreement on what the common constituents of the curriculum should be.

Career structures

The lack of career development for those working on software development or data management within research groups is a problem that has been recognised for over 30 years, but which until recently was not being addressed in a consistent way. One of the authors of this report was a software developer in medical research over 35 years ago, and the problem was recognised by the Medical Research Council (MRC) at that time. Their response was to organise and fund periodic workshops for developers, bringing together often isolated individuals from laboratories around the UK. This helped with skills development through community building, but did not address the underlying career structure issue. At that time the most effective way of overcoming that was for an individual to cease being classified as research or technical staff, and instead become a member of administrative staff. Their specialist skills now became valid criteria for rapid promotion. This required creative management, as well as an individual willing to relinquish any aspiration to be a researcher who happened to write software, and instead accept a supporting role in research.

There is evidence that the issue goes back even further for data managers. CODATA[11] was established over 50 years ago to recognise the particular skills and talents required of those dealing with data in the sciences. Although not explicit, there is an implication in the statements around its founding that these individuals were not sufficiently recognised within their domains, and once again required a community of similar people to interact with in order to develop those skills.

These statements apply primarily to those who desire a career in research, working within research groups. There is evidence of some change at the highest levels; it is now not unknown to see Open Science skills cited in recruitment for university chairs, for instance [28]. The general concerns, however, are not unique to data managers and software developers; they apply to many other techniques and skills essential to support research such as electron microscopy [27]. The problem is at least three-fold:

- career advancement can be hindered because promotion criteria such as publication counts are geared towards those who focus on discipline-specific skills;
- peer esteem and self-esteem can be affected because those with data and software skills are perceived as playing supporting roles to those who actually conduct research, as opposed to being seen as collaborators and hence co-authors;
- funding structures may end up creating support roles which are more likely to be supported through (inherently unstable) project funds than through ‘core’ (and implicitly stable) funding; the opposite, however, can also be true.

Some advances are now being made to address the first two criteria by giving adequate recognition to the scientific contributions made by those with data and software skills. Considering data and software itself to be a first-class research output is one step. DataCite is one of a number of schemes that has helped boost acceptance of data in this way by providing a means to assign identifiers to datasets that parallels that used for research papers. Despite the fact that it has many drawbacks when applied to dynamic data collections, it has helped changed the thinking of many in the research domain about data as a scholarly statement. Others are now applying the same techniques to software and other scholarly outputs such as workflows and methods. In addition to citable source code in software repositories, citable virtual machines can allow a number of these scholarly outputs to be combined within a single entity.

[11] www.codata.org

The EOSC report [2] recognises the value of these new forms of scholarly output but also recognises that the transformation will not be complete until those who assess research and researchers give full credit for these other forms of scholarly statement. Familiar identifiers are a necessary but not sufficient step to realise that aim. The expert group's consultation which preceded the report noted that data generators had historically always received greater credit than data analysts and that the divide was now greater than for the wet-lab analysts of the past.

There is some evidence that change is taking place, at least amongst those who assess researchers for promotion. One of this report's authors has periodically been an external referee for academic promotions where the candidate's primary scholarly contributions have been in creating well-curated data and enabling others to do so. These are the exception rather than the rule at present, but nonetheless do demonstrate that change is possible. A more worrying picture emerges from the perspective of research assessment. The periodic exercise in the UK to assess research quality in universities, most recently referred to as the Research Excellence Framework (REF), depends to a great part on the assessment of selected scholarly outputs for each department put forward by the universities being assessed. Those who administer the REF have made clear for some years that curated data is an acceptable output for assessment. However, there are many within the UK scholarly community who believe that this is not the case. Those who know that it is theoretically acceptable believe that in practice the academic review committees will not know how to assess it and that therefore it is unacceptably risky to put data forward. A great deal of thought and effort goes into the selection of outputs for assessment by universities, since their quality has a significant effect on research funding for some years to come. The result is that most universities behave conservatively, putting forward papers in single disciplines of the type with which they are most familiar. Similar concerns apply to software. As a consequence, only a handful of datasets have been submitted to the REF exercises over the past six years. The REF is particular to the UK context, but the lessons for behaviour change have worldwide relevance. It can be seen that it is not enough for assessors to change their criteria; there needs to be a significant awareness-raising initiative for those being assessed and potentially incentives to change behaviour in what is submitted for assessment; cultural change does not immediately follow legislative or policy change.

Some recent work should help with the recognition problem, but it will take some years to have a demonstrable effect. The CRediT contributor roles taxonomy [29] is one such effort. Developed through CASRAI as a result of a workshop jointly hosted by Harvard University and the Wellcome Trust, CRediT defines a number of roles that can contribute to scholarly outputs including data curation and software development. Recent work by Davidson [30] has extended this to recognise the contributions made by those such as library staff who support researchers in the practise of open science. As of August 2016, some examples have already emerged of the CRediT taxonomy being used on scholarly publications.[12]

It appears that the problems of career development are not so acute for the other professional groups we have considered in this report. Those who work in the various research support roles, whether as library, IT or research admin staff, have no special barriers towards career advancement in their areas. In fact, at present, their skills will set them apart from their peers and make potential career advancement more likely as more and more institutions set up research data services which depend on staff with existing skills. They face the same ultimate barriers to advancement as anyone else in these fields – that at some point one has to cease being a data curator or software developer and accept a

[12] See, for instance, http://journals.plos.org/plosbiology/article/authors?id=10.1371%2Fjournal.pbio.2000074

role primarily defined by management and strategic development if one wishes to advance beyond a certain level.

Those working within large research data centres also have relatively attractive career paths open to them since these centres are staffed primarily by people with similar skillsets, and where researchers with more traditional career paths are external to the institution. This model is not universal but is common enough to provide a choice of career paths for most data and software professionals.

On the evidence available at present, those defined as data scientists have the fewest problems with career progression. They may not be able to advance to the most senior posts in some research institutions (where those posts are reserved for people whose strength lies in a particular discipline), but they do not lack opportunities elsewhere. As the McKinsey report shows, we still face a large skills gap in this area in society as a whole, and demand for data science skills is extremely strong in the worlds of finance and industry. The concern of many in the world of research is that we will lose many of the data scientists needed to domains – or indeed to research organisations in other countries – which can afford to pay them a great deal more.

Longer term actions – funders and institutions

Some of the actions identified here are drawn from external sources which we identity within the text. However, the bulk are the conclusions of the authors, drawn from the evidence gathered in the production of this report and from our own previous experience. For ease of reference, all recommendations are numbered. The accompanying text makes clear if the recommendation is drawn directly from an external source (as is the case with R1) or produced by the authors as a consequence of external evidence (as with R1.)

> **R1: Promotion boards and related activities in research institutions should do more to recognise data and software skills as criteria for advancement and show through example that this change has been effected.**

Most of the recommendations we identify are directed at institutions, not at those who fund research. Some are implicit; for instance, the EOSC HLEG report [2] observes that lack of recognition for data skills by promotion boards in institution is a barrier, but it does not include a specific recommendation to institutions to address this. This is partly because the report's audience is the European Commission and hence the recommendations are solely those which apply to the Commission. We have constructed **R1** as a consequence. The report does include two recommendations for the Commission which, indirectly, would have an effect on institutional behaviour:

> **R2: (EOSC HLEG Recommendation I3): Fund a concerted effort to develop core data expertise in Europe**

> **R3: (EOSC HLEG Recommendation I6): Make adequate data stewardship mandatory for all research proposals**

The Commission has already gone some way to acting on I6, having extended what was a pilot on open data in Horizon 2020 to apply to all new proposals from 2017. However, this does not fully implement the recommendation as worded. At present, projects can argue that one or other exemptions to open data apply to their work and, if successful, it is not clear what requirements on data stewardship will be placed on them.

> **R4: (UUK recommendations 1-4) Support schools to implement the recommendations for education in data skills for 11-18 year olds contained in the UUK report.**

> **R5: (UUK recommendation 2) Embed data analysis across more disciplines through extending models such as Q-step**

> **R6: (UUK recommendation 3) UK Stakeholders to actively promote the data analyst degree apprenticeship option, encouraging employers to offer it and young people to apply for it.**

> **R7: (UUK recommendation 5) Funders to use top-slicing to establish interdisciplinary research and skills development programmes**

The UUK report [20] makes a number of specific recommendations directed at schools and colleges, universities and industry. Those directed at schools include some which will be difficult to realise without dedicated funding, including the embedding of data analysis in other subjects and the establishment of extra-curricular activities, such as summer schools, in which data plays a central role. Funder action is likely to be necessary to help schools realise some of these aims. We have summarised these requirements as **R4** above. Those recommendations directed at universities by contrast are accompanied by recommendations to specific funders to support them. For instance, RCUK is asked to identify specific funding to build on the success of the Q-step initiative to support the embedding of quantitative analysis across all disciplines. Other recommendations are particular to this report, such as the one that suggests that data science graduates require more work to give them business and soft skills. Whilst valid, it is unclear that such requirements are specific to data science and related disciplines. We have extracted recommendations **R5** to **R7** as being of particular relevance to this review.

> **R8 Support and recognise the development of individual skills in data and software across a range of educational and professional interactions instead of purely within the context of a specific educational setting**

There are also recommendations from within the profession on how to use existing frameworks to develop teaching and training in open science skills. These include proposals such as [31] which shows how to extend skills frameworks to consider the development of individual capabilities through a range of educational interactions. This has parallels with the existing Vitae framework for data literacy [14]. [31] focuses in detail on education within the university for research support staff; the UUK report, by contrast, considers the pre- and post-university context, but only for the data scientist and related roles. We have synthesised these views as recommendation **R8**.

> **R9 (Concordat principle #1) Funders to support the costs associated with open research practice**

> **R10 (Concordat principle #1) Funders to ensure that institutions foster research environments which recognise the value of open data and provide staff with skills and infrastructure necessary to practice open research**

The UK concordat on open research data [9] contains a number of requirements buried within its principles which can be read as being even stronger than recommendations, with language such as 'will' being used instead of the more common 'should.' For instance, principle #1 ends with requirements on researchers, their employers and funders:

> **Employers of Researchers** will foster a research environment which recognises the value of open data and will seek to provide appropriate access to infrastructure systems and services to enable their researchers to make research data open and usable, having due regard to value for money. They will also recognise good data management as an important aspect of researchers' duties.

> **Funders of Research** will support open research data by appropriately acknowledging and supporting its costs, and by supporting the wider agenda with appropriate policy and investment activities.

Principle 6 notes that:

> The importance of training in research data management cannot be overstated as an enabler of open research data, and all researchers should receive such training at an early stage in their careers, along with subsequent updating as appropriate.

but it is not explicit about where responsibility for the training provision lies, although the language in principle 9 lays this responsibility on the employer. However, it also says that

> Individual researchers must also ensure their own data skills are at a level sufficient to meet their own obligations whilst understanding the benefits to themselves of a higher level of understanding.

Finally, principle 9 ends with a more gently-worded joint action for funders and research institutions regarding career paths for data scientists:

> The specialised skills of data scientists are crucial in supporting the data management needs of researchers and institutions. Research institutions and funders should work together to help build under-pinning capacity and capability in this area, and to attract and retain such specialists by developing well-designed and sustainable career paths for them.

We have argued that it is not the data scientists who have a problem with career paths, but there are others involved in data and software management to whom this recommendation could usefully be applied within universities.

> **R11 Funders to work with the European Commission and other bodies to enable the creation of data compentencies and skills exchange hubs**

Other recommendations for European funders in particular are not specific to the Open Science domain, but consider digital skills within the wider context of the Digital Single Market. However, these recommendations are at least partially applicable to the research domain and, if broadly implemented, could benefit it. The next few examples are not strictly recommendations, but statements from the Commission itself about its intended future actions and the rationale for them. In doing so, however, The Commission is also recommending coordinated action by others including the governments of member states and other public and private actors. They come from "Digitising European Industry - Reaping the full benefits of a Digital Single Market" [32]. After recognising the nature of the skills gap, one recommended action is the creation of 'data competencies and skills exchange hubs' in which practical work with real datasets will allow skills transfer from the more-experienced to the less-experienced. This is captured in **R11** above, which is a specific instance of **R14 and R15** below. The European Commission also sees a role for itself in bridging techniques emerging from data-intensive research into market applications, and also for hastening the creation of a Europe-wide accepted set of competencies in data, which it is accomplishing through projects such as EDISON.

The remainder of the recommendations here are drawn from our own observations. The first, and most important, is for funders to work harder to coordinate actions, funding and policy across national and disciplinary boundaries. Such coordination can take a number of forms.

> **R12 Funders to make greater efforts to coordinate policy relating to data and software and to communicate the results of that coordination and the commonalities in implementation and compliance**

On policy, funders need to work harder to coordinate policy relating to data and software and to make clear the results of that coordination. Lack of coordinated policy leads to greater waste by researchers and research organisations in complying with policies which vary without good reason (the complexity in journal Open Access policies is one imperfect analogue.) It is also important to make clear that some policies which appear to be different are fundamentally the same. For instance, RCUK has a single set of principles regarding data which are then translated into more specific statements by each of its constituent councils. NERC and ESRC operate their own data centres, and hence their policy statements explicitly require the use of the infrastructure that they fund; other councils do not operate similar infrastructure and therefore use less prescriptive language. Yet all of these policies are addressing the same fundamental requirement about ensuring that data of lasting value is made available for the long term in an environment which enables reuse, and where a permanent identifier is attached. Universities understandably seem to treat the funder policies as independent statements requiring separate compliance exercises instead of focussing on the commonalities which underlie those policies. This is not just a UK issue; similar observations can be made about US funder policies which ultimately derive from a single directive from the White House Office of Science and Technology Policy. Funders such as NSF already had funder-level policies which were compliant but are different in form and each of NSF's funding boards then translate the statements into a more discipline-specific form; universities and researchers again focus on compliance at the lower level instead of the higher one.

> **R13 Funders to coordinate methods for assessing research quality in ways which make clear the value assigned to data and software outputs and impacts**

Funders can also work to coordinate the way they assess research quality, demonstrating that they are all assigning value to data and software outputs and their wider impacts, as well as to more traditional outputs such as publications, and doing so in comparable ways.

> **R14 Funders to coordinate more trans-national and interdisciplinary funding actions which develop and recognise data and software skills in research**

> **R15 Funders to find better mechanisms to allow cross-border funding of open data and software infrastructure for research**

Funders can also assist by coordinating funding actions across national and disciplinary boundaries. The Digging Into Data[13] programme and the Belmont Forum e-infrastructure data management programme[14] are two good examples of ways that funders can act in concert. The first is an interdisciplinary and international research programme which brings data science techniques to the social sciences and the humanities. It has enabled cultural institutions to recognise digital collections as data when they might not previously have done so and has achieved wider impact because of its cross-border nature. The Belmont Forum programme, meanwhile, coordinates work on skills development and infrastructure provision to ensure more seamless worldwide access to data resources necessary to understand and make headway on issues around climate change. There are many similar examples where we need to perceive data infrastructure as a global resource but where the funding comes from one or two national bodies, often on a project basis. At present, it can be difficult for a funder in one country to support infrastructure in another, even if that is clearly the most efficient solution for all involved. Large infrastructures such as CERN use treaty-based mechanisms to circumvent these issues, but these are not applicable to any but the very largest of cooperative undertakings.

[13] www.diggingintodata.org
[14] www.bfe-inf.org

R16 Funders and institutions to support the development of current and existing peer networks which enable skills development and skills transfer

R17 Funders to work with initiatives such as Software Carpentry which rely on volunteer labour to ensure the development of a long-term financially and organisationally sustainable operational model

There are specific activities that funders could do more to support. They can allocate funds to support the community building typified by the SSI’s Collaboration Network events, the DCC’s Research Data Management Forum series and SPARC Europe’s Open Data Champions. Support can include funding the time needed to create and maintain networks, subsidy of events, travel support for researchers and promoting the networks to researchers that they fund. Funders should also work with initiatives such as Software Carpentry which rely on volunteer labour to identify whether this is sustainable as the activity grows and, if not, help them transition to a model which is financially more secure.

References and further reading

1. Where do all these forensic science students go? (John Cassella, March 2014) <http://blogs.staffs.ac.uk/johncassella/2014/03/13/where-do-all-these-forensic-science-graduates-go/>
2. Report of the High Level Expert Group on the European Open Science Cloud < http://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud>
3. Big data: The next frontier for innovation, competition, and productivity (James Manyika et al at McKinsey Global Institute, May 2011) http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation
4. Science as a Social Machine (Barend Mons, February 2016 keynote at IDCC) < http://www.dcc.ac.uk/events/idcc16/programme-presentations>
5. New Academy to help Europe meet Big Data Skills Gap (Open University, January 2016) <https://www3.open.ac.uk/media/fullstory.aspx?id=28449>
6. BIS – seizing the data opportunity https://www.gov.uk/government/publications/uk-data-capability-strategy
7. CEBR – data equity < http://www.sas.com/offices/europe/uk/downloads/data-equity-cebr.pdf>
8. Data Science Competence Framework (EDISON project, 2016) http://edison-project.eu/data-science-competence-framework-cf-ds
9. UK Concordat on Open Research Data (2016) < http://www.rcuk.ac.uk/documents/documents/concordatonopenresearchdata-pdf/>
10. Could Open Research benefit Cambridge Researchers? (Lauren Cadwallader, Joanna Jasiewicz and Marta Teperek, August 2016) https://unlockingresearch.blog.lib.cam.ac.uk/?p=817
11. Liz Lyon & Aaron Brenner (Feb 2015): Bridging the Data Talent Gap: Positioning the iSchool as an Agent For Change ; International Journal of Digital Curation 10(1) http://dx.doi.org/10.2218/ijdc.v10i1.349
12. D2.1 Technical protocol for rich metadata categorisation and content (Joy Davidson et al, FOSTER project, June 2014) https://www.fosteropenscience.eu/images/documents/D2.1_Technical_protocol_for_rich_metadata_categorization_and_content_classification.pdf
13. DaMSII-ABC outputs (DCC, 2013) < http://www.dcc.ac.uk/training/damssi-abc>
14. Information literacy lens on the researcher development framework (Vitae, 2010) < https://www.vitae.ac.uk/vitae-publications/rdf-related/information-literacy-lens-on-the-vitae-researcher-development-framework-rdf-apr-2012.pdf/view>
15. Nelson, M. (2016). Scaffolding for data management skills: From undergraduate education through postgraduate training and beyond. Purdue University Research Repository. doi:10.4231/R7QJ7F9R
16. RDA IG-ETHRD (Interest Group on Education and Training in the Handling of Research Data) < https://rd-alliance.org/groups/education-and-training-handling-research-data.html>
17. RDM MANTRA (EDINA and Data Library, University of Edinburgh, 2011-) <http://datalib.edina.ac.uk/mantra/>
18. RDM MOOC https://www.coursera.org/learn/data-management
19. Essentials 4 Data Support (Research Data Netherlands) http://datasupport.researchdata.nl/en/
20. Making the most of data: data skills training in English universities (UUK, 2015) http://www.universitiesuk.ac.uk/policy-and-analysis/reports/Documents/2015/making-the-most-of-data-training-skills-english-universities.pdf

21. Software Carpentry – About <http://software-carpentry.org/about/>
22. Mozilla Open Science 101 https://github.com/OKScienceDE/Open_Science_101
23. RDA-CODATA group on summer schools in data science and cloud computing in the developing world https://rd-alliance.org/groups/rdacodata-summer-schools-data-science-and-cloud-computing-developing-world.html
24. DCC training http://www.dcc.ac.uk/training/
25. <https://www.software.ac.uk/fellowship-programme>
26. Collaborations Workshop 17 https://www.software.ac.uk/cw17
27. EPSRC electron microscopy roadmap https://www.epsrc.ac.uk/files/research/electronmicroscopyroadmap/
28. "Changing hiring practices towards research transparency: The first open science statement in a professorship advertisement" (Felix Schönbrodt, January 2016) http://www.nicebread.de/open-science-hiring-practices/
29. CRediT http://casrai.org/credit
30. Davidson, J. (2016) Fostering open science practice through recognising and rewarding research data management and curation skills. In: Bisto, C. and Raju, R. (eds.) LIS Education and Research in a Dynamic Information Landscape: Proceedings of the Library and Information Studies Centre 75 Years Commemorative Conference. University of Cape Town Libraries: Cape Town, pp. 63-75. ISBN 9780799225266 (doi:10.15641/0-7992-2526-6)
31. A proposed scaffolding for data management skills (Megan Sapp Nelson, June 2016) https://www.openconf.org/IASSIST16/modules/request.php?module=oc_program&action=summary.php&id=33
32. Digitising European Industry - Reaping the full benefits of a Digital Single Market - COM(2016) 180 final (EC, April 2016)

October 2016

Version 1

Wellcome exists to improve health for everyone by helping great ideas to thrive. We're a global charitable foundation, both politically and financially independent. We support scientists and researchers, take on big problems, fuel imaginations and spark debate.

**Wellcome Trust, 215 Euston Road,
London NW1 2BE, UK
T +44 (0)20 7611 8888, F +44 (0)20 7611 8545,
E contact@wellcome.ac.uk, wellcome.ac.uk**