Key Perspectives

Consultants in scholarly information

THE SKILLS, ROLE AND CAREER STRUCTURE OF DATA SCIENTISTS AND CURATORS: AN ASSESSMENT OF CURRENT PRACTICE AND FUTURE NEEDS

REPORT TO THE JISC

July 2008

Prepared by:
Alma Swan and Sheridan Brown
Key Perspectives Ltd
48 Old Coach Road
Playing Place
Truro
TR3 6ET
UK
+44 1392 879702
aswan@keyperspectives.co.uk
www.keyperspectives.co.uk

CONTENTS

1. EXECUTIVE SUMMARY

This study was commissioned by the JISC to specifically address two recommendations from the report by Liz Lyon on data management in the UK (Lyon, 2007). The main aim of the project was to examine and make recommendations on the role and career development of data scientists and the associated supply of specialist data curation skills to the research community.

The nomenclature that currently prevails is inexact and can lead to misunderstanding about the different data-related roles that exist. We have attempted to reconcile in section 3.1 the definitions offered by authoritative organisations and the practical experience of people working in the field. We distinguish four roles: data creator, data scientist, data manager and data librarian. We define them in brief as follows:

- Data creator: researchers with domain expertise who produce data. These people may have a high level of expertise in handling, manipulating and using data
- Data scientist: people who work where the research is carried out – or, in the case of data centre personnel, in close collaboration with the creators of the data – and may be involved in creative enquiry and analysis, enabling others to work with digital data, and developments in data base technology
- Data manager: computer scientists, information technologists or information scientists and who take responsibility for computing facilities, storage, continuing access and preservation of data
- Data librarian: people originating from the library community, trained and specialising in the curation, preservation and archiving of data

In practice, there is not yet an exact use of such terms in the data community, and the demarcation between roles may be blurred. It will take time for a clear terminology to become general currency.

Data science is now a topic of attention internationally. In the USA, Canada, Australia, the UK and Europe, developments are occurring. It is notable that the vision in all these places is that data science should be organised and developed on a national patern rather than relying on piecemeal approaches to the issues.

Researchers in general are becoming much more aware of the issues that data-based research raise. Some already possess considerable skills in handling and managing data (so-called 'native data scientists'), but even those less experienced in this regard show an interest in learning more. They turn, in the absence of a data scientist in their circle, to the institutional IT services or library for assistance and advice. Some UK universities are now beginning to offer taught master's courses in data management which may help to raise the general data skill level. Just as data centres have been training data scientists for some time now and accepting that they will eventually leave for other jobs, thus helping to diffuse data skills into the research community, so increasing numbers of researchers with postgraduate training specifically in data-related matters will do the same.

Data scientists have usually ended up in their role by accident rather than by design, though this is changing as more data science posts are created. They may be qualified for their role by either being a domain expert who has acquired specialist data skills in the course of their career, or by originating as a computer scientist who has acquired domain knowledge over time. Most data scientists currently in post say they have learned their skills on the job because of the lack of proper training opportunities and the cost (in time and money) of attending suitable events. Although until recently there has been no tight specification for qualifications the trend now is increasingly for postgraduate training in informatics to be required. In practice, data scientists need a wide range of skills: domain expertise and computing skills are prerequisites but ‘people skills’ are also valued since a major part of the role is in translating the needs and practices of the researchers for the computing experts (people we have defined as data managers) and to some extent vice versa.

There is no defined career structure for data scientists and this is a major problem that must be resolved if the UK research community is to be properly supplied with data skills. Data scientists may be in tenured jobs in universities and data centres, or they may be employed on short-term research contracts. Those in tenured roles in universities may be on a variety of career grades, from technical through service or academic-related grades to full academic grades. There is no consistency across the system at present. The lack of job security is an issue in encouraging and retaining data scientists and demand currently far outstrips the supply of skilled people. Another issue causing some degree of disaffection amongst data scientists (or would-be ones) is that they can feel undervalued, a result of the lack of professionalisation of their role and of a formal, organised career structure.

People in data science roles face a big, continuing challenge in remaining properly skilled up. Data matters are moving very quickly and they need to stay abreast of general developments and developments specific to their field. In some disciplines there are international workshops that serve to assist in this, but even here these are not always enough. Data scientists favour the idea of continuing professional development in the form of regular short courses on specific topics that are ‘of the moment’ and hope that such a system will become an accepted part of their role.

As regards the question of whether there is value in extending data skills within the undergraduate curriculum, there is a dichotomy of views. Whilst many people consider this advantageous – data scientists themselves think that the earlier basic data skills are instilled in future researchers the better – many people teaching undergraduate programmes say that they are full enough as they are without adding specific data skills modules. They also point out that in disciplines where data handling skills are very pressing, the undergraduate curriculum already has elements (such as teaching how to construct and use simple relational databases) included within it. It looks likely that further data skills training will naturally become part of undergraduate training as things evolve over time, in ways appropriate to each discipline.

The role of the library in data-intensive research is important and a strategic repositioning of the library with respect to research support is now appropriate. We see three main potential roles for the library: increasing data-awareness amongst researchers; providing archiving and preservatin services for data within the institution through institutional repositories; and developing a new professional strand of practice in the form of data librarianship. In the US, advances are already being seen in this respect as the library community aligns with the demands of the data deluge and organises to provide data archiving and preservation skills formally via library school education. There is a fledgling advance in this area in the UK, too. There are, however, not enough specialised data librarians yet. In the UK there are thought to be just five at the moment, something that will need to be changed quickly. One reason why there are so few so far is a parallel with the situation for data scientists – there is no recognised career path. Attracting well-qualified – that is, pre-qualified in specific domains so that an understanding of the data structures and uses in a domain comes as a given – is also difficult at present. And, in the US, which is further along the path than the UK, a lack of suitable internships for data librarian trainees has also been identified as a factor hampering training in the profession: this may yet also prove to be an issue here.

The main recommendations from the study are as follows:

1. Recommendations regarding data skills development in research domains (RD):

Recommendation RD1**:** Major research funders in the UK should work with universities and research institutes to define properly and to formalise the role of data scientists, and to develop the means by which the work of data scientists can be recognised and remunerated.

Recommendation RD2**:** These same bodies should work together to create the conditions that support data science, foster its study and encourage professionalisation of the role.

Recommendation RD3**:** The JISC and other organisations that commission original research should take forward a study (or studies) that cover the following issues:

- A description of the role played by data scientists and the value of the contribution they make to research
- Examples of data science careers
- The development of a set of practices that represent good practice in data science

Recommendation RD4**:** The relevant bodies (HEFCE and the research councils) should consider the establishment and funding of a network of trainers with the skills to deliver short postgraduate training courses to researchers covering the fundamentals of data management, thus building basic data science skills into the research process. Some of the research councils have laid the foundations for this with their requirements for a data plan in grant applications.

Recommendation RD5**:** The research councils and other research funders should consider whether, as part of the grant application and award process, they should require at least one member of the project team to be nominated as the project's data scientist. This person should be required to attend a short course covering the fundamentals of data science and management. Research councils should consider the extent to which accrediting valid courses and proof of attendance is necessary.

2. Recommendations regarding data skills development in research libraries (RL):

Recommendation RL1**:** The research library community in the UK should work with universities and research institutes to define properly and to formalise the role of data librarians, and to develop a curriculum that ensures a suitable supply of librarians skilled in data handling.

Recommendation RL2**:** The JISC should consider supporting the development of the International Data curation Education Action (IDEA) working group. This group is well-placed to play an important advisory role in the development of appropriate curricula for future data librarians, particularly those coming through the library and information science route.

3. Recommendations regarding data skills development in general (RG):

***Recommendation RG1*:** Because there are already a number of players active in the data area there is potential for exploiting synergies in respect of data skills training. It is recommended that a study scopes this potential, looking in particular at the activities of the UK Data Archive, universities or research groups where data science is advanced, library schools, the Digital Curation Centre and IDEA (the International Data curation Education Alliance). The study might also look internationally at initiatives in the US, Canada and Australia.

2. INTRODUCTION AND METHODOLOGY

"Opening a 5th dimension through cyberinfrastructure is the revolutionary force of the digital age … individuals, groups, organizations and nations that don't embrace the 5th dimension will fall behind in the digital age" (Christopher Greer[1], 2007).

This study was commissioned by the JISC to put into effect two of the recommendations in the report *Dealing With Data: Roles, Rights, Responsibilities and Relationships* (Lyon, 2007) which gave an overview of the UK scene with respect to digital research data. Amongst a host of recommendations drawn up in this report were the following two:

> *REC 34. A study is needed to examine the role and career development of data scientists, and the associated supply of specialist data curation skills to the research community.*
>
> *REC 35. JISC should fund a study to assess the value and potential of extending data handling, curation and preservation skills within the undergraduate and postgraduate curriculum*

The recommendations above were made in the context of the 'data deluge', the growing amount of digital data pouring out of research efforts across the disciplinary spectrum. So-called 'big science' or 'e-science' is always put forward as the reason why data issues should be paid more attention, and it is true that big science generates a lot of data: around 80% of all research data are produced by three areas – high energy physics, meteorology and astronomy. Nonetheless, 'small science' also plays its part in the data deluge and its contribution is also increasing. All these data need to be managed and looked after. The skills to enable re-use – often in ways the original creators never imagined – are critically important for the progress of research. Data scientists bring these data handling, manipulation and curation skills to the research community and data librarians provide archiving and preservation skills to ensure safe custody for data outputs.

The importance of data science and data care is brought into clearer focus when research data are considered from a long tail perspective. Data produced in so-called big science projects tend to be *relatively* more homogeneous and straightforward to curate from a technical perspective. The data outputs from small science, on the other hand, tend to be extremely heterogeneous and requiring of unique procedures to create or process the data. In short, data that reside in the long tail are difficult to curate, re-use and preserve and yet have much potential value. And small science data are certainly expensive to produce: a recent analysis of National Science Foundation (USA) grants for biological research awarded in 2007 showed that 44% of the total funds awarded went to small science projects (those with a value up to $350,000)[2]. Providing the means to unlock the potential re-use value of the data produced by small science projects is an important challenge faced by data scientists and data librarians.

Methodology

A multi-faceted approach was adopted for this project with a bias towards qualitative techniques in order to ensure we were able to explore people's perspectives in sufficient detail. The primary research incorporated a series of fifty-seven semi-structured in-depth

[1] Christopher Greer, Office of Cyberinfrastructure, National Science Foundation

[2] BP Heidorn, Graduate School of Library and Information Science University of Illinois at Urbana-Champaign & the NSF, June 2008 (personal communication)

personal interviews together with four focus groups to adduce the views of data scientists (including those embedded in research groups and those working for data centres and research councils), librarians, library technologists and educators. Focus groups and interviews were carried out in England, Northern Ireland and Scotland. We sought the views of researchers from a wide range of different subject areas including systems biology, astronomy, chemistry, archaeology, geology, ecology, rural economy and land use, and a number of other fields in the social sciences. This process was underpinned by an online survey of data scientists and by thorough desk research. We also participated in a two-day workshop in Washington DC to discuss the development of digital curation curricula, which was attended by experts from the USA and the UK, as well as meetings organised by the JISC to promote open communication between JISC-funded projects working on data-related topics.

Part of the challenge of this project has been the requirement to distil the findings into an incisive document or around thirty pages and to restrict the number of recommendations to a maximum of ten.

We thank those who generously gave us their time and views in this exercise. They are all busy people but participated willingly in the interests of scholarship.

Alma Swan and Sheridan Brown
Key Perspectives Ltd
Truro, UK
1 September 2008

3. OVERVIEW OF DATA SCIENCE ISSUES

3.1 Some definitions

As soon as we began this project it became apparent that the issues of what to call whom and who does what have not yet shaken down into common terms of usage. The project sponsor used the term 'data scientist' for the role under study, and attributed to that role the tasks of *data handling*, *curation* and *preservation*. In our study, however, we found that whilst those people who considered themselves data scientists may do all these three things, they place the greatest emphasis on the first of the three – data handling – but did not *necessarily* consider themselves data curators or data preservers. In many cases these are discrete roles carried out by persons with a high degree of specialism.

There are, then, two spectra: one spectrum is the range of tasks carried out with or on digital data and the other is the range of role titles people may acquire. The former is dealt with in Section 4.1: here we try to come to some workable conclusion as to how to label roles for this study.

The US National Science Foundation's *National Science Board* came up with one definition in a report published in 2005 (NSF, 2005), calling for the creation of data scientists who are:

> *"the information and computer scientists, database and software engineers and programmers, disciplinary experts, curators and expert annotators, librarians, archivists, and others, who are crucial to the successful management of a digital data collection"*.

This is very inclusive, embracing people with pure research roles through computer scientists to members of the library community. The report goes on to say that these people:

> *"conduct creative enquiry and analysis; enhance through consultation, collaboration and coordination the ability of others to conduct research and education using digital data collections; be at the forefront in developing innovative concepts in database technology and information sciences, including methods for data visualization and information discovery, and applying these in the fields of science and education relevant to the collection; implement best practices and technology; serve as a mentor to beginning or transitioning investigators, students and others interested in pursuing data science; and design and implement education and outreach programs that make the benefits of data collections and digital information science available to the broadest possible range of researchers, educators, students and the general public"*.

It further distinguishes the role of data scientist from three other roles:

- *Data authors: the scientists, educators, students, and others involved in research that produces digital data. They include domain scientists, educators and students who have a vested interest in the research generated from the data*
- *Data managers: the organizations and* **data scientists** *[our emphasis, to denote the potential for confusion] responsible for database operation and maintenance and a reliable and competent partner in data archiving and preservation*
- *Data users: the larger scientific and education communities, including their representative professional and scientific communities*

Having studied data management in practice in the research community in the UK, especially with respect to practice in HEIs as required by the project sponsor, we would not make these distinctions in this way. In our view the roles most clearly distinguish themselves as:

- Data creators or data authors: researchers with domain expertise who produce data. These people may have a high level of expertise in handling, manipulating and using data, gained through experience and as a result of need or personal interest.
- Data scientists: people who work where the research is carried out – or, in the case of data centre personnel, in close collaboration with the creators of the data – and conduct all or a number of the functions described in the NSF's definition above including, in many cases, being data creators themselves. In origin and training they may be domain experts, computer scientists or information technologists and their career development may have required them to assimilate skills from a discipline from which they did not originate. So, a data scientist in systems biology may be a biologist by origin who has acquired very considerable computing skills and a data scientist whose background is in software engineering may have acquired a considerable degree of biological knowledge. Some data scientists told us that an important part of their role is to be a 'translator', communicating the needs of the data creators to data managers (see below) and working with the data managers to ensure that data are stored and accessible in a usable way
- Data managers: people who are computer scientists, information technologists or information scientists and who take responsibility for computing facilities, storage, continuing access and preservation of data. They liaise extremely closely with data scientists, ensuring that the right technological facilities are available for the research group to be able to carry out its work effectively. Some data managers described their role as data 'plumber', piping data from one place to another, ensuring data flows operate properly and that valuable data are not lost
- Data librarians: originating from the library community, trained and specialising in the curation, preservation and archiving of data. Originally, the term data librarian seemed to be confined to librarians dealing with social science data, but the title now encompasses people with data skills in all disciplines. It is a particularly important area as institutions begin to develop digital repositories for the collection and curation of their research outputs. Datasets are part of those outputs, an institutional repository is a natural home for them and the repository is usually in the care of the library. Whilst 'big science' has its (international) data centres and some research councils in the UK provide national data storage facilities, 'small science' will need to be provided for by institutions. Even if a third player in the form of a national data service does materialise, there will still be the need for local facilities for data that do not qualify, for one reason or another, for inclusion in that data centre, and data librarians will be the custodians of that body of data

We recognise that in current practice these terms are not used precisely as we have defined them here. People we view as data scientists are in jobs called data manager or data specialist. It will take time for the terminology to standardise but for the purpose of this report we will adhere to the definitions given above. It is also important to note that the boundaries between these roles are currently very fuzzy. Pragmatism, need and personal preferences have been the main drivers so far in the area of role development. We would expect some change as the role of data scientist becomes more common and as institutions recognise new responsibilities for data outputs, thus driving the need for the research community and the library to take a shared approach to managing data.

3.2 National approaches

Whatever tags the people who manage data are called by, they are being given increasing attention by governments and the wider research community. We report here a brief overview from selected countries.

In Canada, lacking any system of national data archives on the UK model (such as UKDA) and with a reluctance on the part of the Canadian Government to establish any new institutions, strategies for data have become an issue of finding existing institutions to take responsibility. The research councils are carrying out consultations and Library and Archives Canada (the merged National Library and National Archives) has conducted an 18-month study and published a report (Canadian Digital Information Strategy, 2007). CARL (Canadian Association of Research Libraries) is also looking at the issue, along with a task force called Research Data Canada coordinated by CISTI (the Canadian Institute for Scientific and Technical Information), from the perspective of trying to tie in data strategies with those for institutional repositories.

The Australian Government is establishing the Australian National Data Service (ANDS - http://ands.org.au/) which should be fully operational during the last quarter of 2008. This has four main programmes of activity: Developing Frameworks which is concerned with national and institutional policies, Providing Utilities dealing with discovery and machine-to-machine services, Seeding the Commons which is building a data commons for Australian research and Developing Capabilities responsible for improving data management human capability across the country. ANDS is a consortium of Monash University, the Australian National University and the Commonwealth Scientific and Industrial Research Organisation (CSIRO).

In the USA the Interagency Working Group on Digital Data (IWGDD), which represents 22 Federal agencies including the National Science Foundation, has been working to develop and promote a strategic plan for the preservation of and access to digital data. The final report is due later this year, but the draft strategy calls for the creation of a national coordinating body for digital scientific data preservation and access as well as efforts to maximise the accessibility and utility of digital data. With respect to education and training, the IWGDD recommends three key policies: first, that education and training activities be integral to all Federal science data investments; second, that the national coordinating body promotes the coordination of education and training across Federal agencies and education, research and technology sectors; and third, that the national coordinating body promote data science and management as career paths with appropriate recognition and reward structures.

Here in the UK, in spring 2007 HEFCE issued a shared services call, as a result of which a project was started to look at the feasibility of a UK Research Data Service (UKRDS). The project is expected to report in December 2008. In addition, the Digital Curation Centre, funded by JISC and established in 2004, has a range of activities under its umbrella including the delivery of a data skills summer school later this year (2008). The JISC has also commissioned a number of studies on data, some of which touch upon the roles of data scientists, and the Research Information Network has recently published a study on how researchers create, publish and share data (Brown and Swan, 2008).

On a broader European front, data management and sharing are part of an infrastructural approach across Europe. ESFRI (the European Strategy Forum on Research Infrastructures), established in 2002 at the behest of the European Council, published a

Road map for pan-European research infrastructures in 2006 and will convene a conference in December 2008.

4. THE ROLES AND CAREERS OF DATA SCIENTISTS IN THE UNITED KINGDOM

4.1 Introduction

There is an assumption that data scientists have a major role in bringing computing skills to the experimental team[3]. This is true, but before we discuss these skills and how they are acquired, it should be noted that they are not *necessarily* needed by more computing-oriented researchers. Some researchers have themselves considerable database and computing skills gained from studying relevant modules during their undergraduate training, people who fit the description 'native data scientists' coined by Liz Lyon in her 2007 report (Lyon, 2007)[4]. Some have also done Linux certification courses and others have pursued Java certification. It is as well to keep in mind that the computing skills of the current researcher community spans the whole gamut from none at all in some individuals to very accomplished in individuals at the other end of the curve.

Indeed, given the scarcity of people in data scientist positions, researchers commonly have to find their own way towards data skills and commonly do this by approaching their institutional or departmental IT services for advice and help. This may be necessary even in well-funded areas of science. For example, we studied four systems biology groups during this project: three have a data scientist in post but one of them does not. In this group, researchers liaised with computer scientists (variously called computing officers, IT support or similar), leaning on them heavily for advice and practical solutions to data handling problems. In science departments such support may be quite readily to hand but in other disciplines or in less well-endowed institutions this is not so likely. It seems certain that data handling and management skills are going to become highly sought-after.

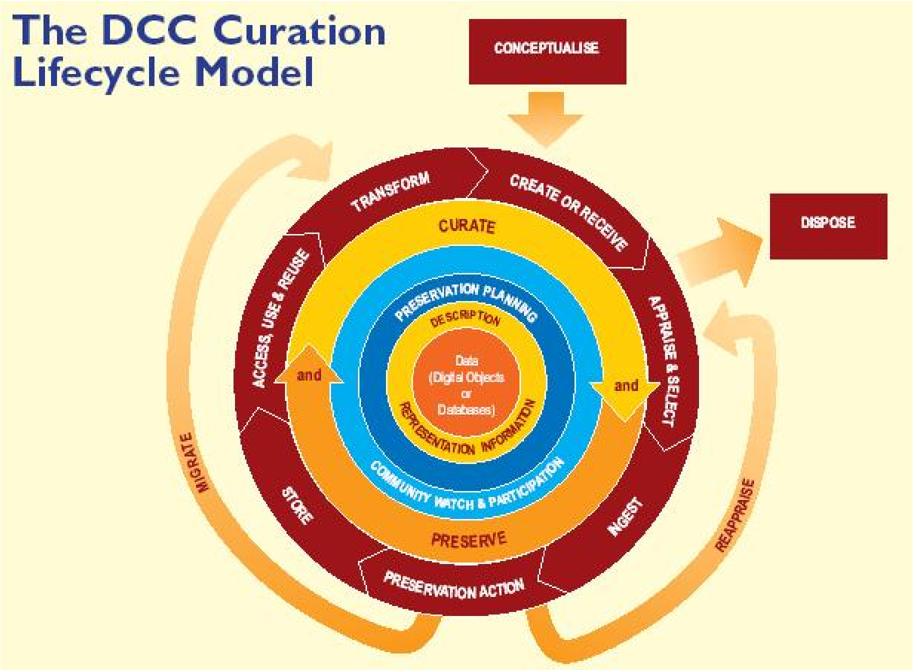
4.2 What data scientists do

Notwithstanding the current paucity of data scientists and the 'fuzziness' we ascribed to the present system of roles and responsibilities, a fairly clear picture of data science as a professional role is emerging.

With respect to what 'official' data scientists do, it is helpful to take a data life cycle approach and in doing that refer back to the definition of a data scientist produced by the NSF and referred to in Section 3.1. A visualisation of the data curation life cycle has developed by the Digital Curation Centre and is reproduced below. Naturally, given its provenance, it has a strong focus on curation and preservation activities – those that sit in the cycle from '5 o'clock' onwards.

[3] We use the terms 'experimentalist' or 'experimental team' to indicate those involved in generating or creating research data. In disciplines outside experimental science, data are created in other ways and are of different form, but for simplicity we are using the term to indicate the researchers involved in data creation across the board

[4] "… the 'net generation' is more open to data sharing. Indeed, it is likely that over time, "native data scientists" … will emerge through the acquisition of relevant skills learnt through the standard educational curriculum" (Lyon, 2007, p55)



Using our broad definitions of roles from Section 3.1, we identified the following with respect to where the various roles sit in that cycle:

4.2.1 Data scientists

Data scientists in research groups broadly focus on the tasks from 12 o'clock to 4 o'clock – *conceptualisation*, *creation*, *access and use*, *appraisal* and *selection*. They may also go further, depending on the discipline and the type of research community in which they operate and the norms and requirements of that community. For example, in one of the areas we studied in depth, systems biology, the need for preservation and storage activities is limited because there are large public access databases to which researchers submit their data and which guarantee professional archiving and preservation processes for the datasets thereafter. In other disciplines this is not the case and data scientists may be responsible for considerable preservation efforts locally. In these cases there is an institutional perspective though not necessarily yet an institutional view on the issues involved.

One thing to bear in mind is that data scientists tend to use terminology slightly differently to the Lifecycle Model: specifically, they refer to the modification, annotation, reduction, derivation and other manipulations of data as they pass from the rawest stage through to other stages, at any of which they may be accessed and used by experimentalists, as curation. In common parlance, curation is a term normally associated with the term preservation (as in the Data Lifecycle Model above), implying activities that are undertaken with a view to looking after data in the longer term. In an attempt to reflect accurately the activities undertaken at all stages of the data life cycle, therefore, we avoid the use of the term curation as far as possible in this report and where we do use it we aim to reflect those activities that data scientists and experimentalists are engaged in as they process and use the data.

Data scientist skills are also extremely important before the start of the data creation process – during what the DCC's Lifecycle Model terms *conceptualisation*. In experimental disciplines the data scientist aids in experiment planning and design, advises on how to collect data in optimal ways, the types of data that are advisable to collect, may train data creators in the use of laboratory machines and how to work with the proprietary software that these employ, and assists in planning the project protocols to derive the greatest data-related benefits from the intended investigation. At this early stage, too, the data scientist works with the researchers to write a data plan in project proposals, advises on the requirements regarding data from funders, and helps to ensure that the work goes ahead in a way that conforms to good data management practice and fulfils any obligations placed on the research group by external or institutional players.

During experimental processes, the data scientist may also determine how to compare or cross-register two types of data outputs (say, confocal microscope images with X-ray images), and thus contribute insights to the research process that complement those of the experimentalists. The final thing to point up here is that embedded data scientists are frequently called upon to assist in data access and re-use – a task that appears further round the DCC's life cycle. Their skills are needed here because access and re-use of datasets not produced by their own research group may require specialist script-writing and/or data transformation skills that the researchers do not necessarily have. The data scientist provides not only the instrumental skills at this point but will also be able to develop such skills didactically within the research group as these occasions arise.

Data scientists in data centres serving specific parts of the research community may do all these things – or just some of them – and they also pay much attention to the tasks itemised further along in the life cycle, from '5 o'clock' onwards. The scale of data centres' operations often necessitates a higher degree of individual specialisation, where a data scientist may be responsible for just one part of the data curation process. Although data scientists in these circumstances develop highly focused skills, the skills are often closely associated with a particular process or dataset and may sometimes have limited value in terms of transferability. That said, data centres are normally able to offer their data scientists a variety of different work opportunities and even the possibility of doing original research alongside their data science roles. This variety is important for maintaining people's interest and enthusiasm over the medium term. Data centre data scientists also have access to a much higher level of professional backup than is available to those working in research groups. Data centres normally have a wealth of collective experience, a great deal of expertise distributed among staff members, and access to data managers and other technology-based resources. Finally, although most data centres are required to bid for operational funding periodically, established data centres are well-positioned to continue to be funded, and they are also well-placed to leverage opportunities for additional funding for particular projects. This means that data scientists working in data centres typically enjoy greater job security than may pertain to other working environments.

4.2.2 Data managers

Data managers, by our definition, are individuals with specialist skills in computational science, experts in database technologies, and are responsible for ensuring that data produced and needed by the research team are properly stored, curated and preserved. They manage the systems for backing-up and refreshing data, for format migration where appropriate and liaise with the data scientists (and sometimes the experimentalists directly) over what systems need to be in place and the kind of data that will be looked after. Data scientists often assume a 'middle man' role between the experimentalists and the data managers, and translate the data needs and problems of the experimentalists into issues that the data managers can resolve.

4.3 Data scientists' qualifications and career path

It is apparent that for many people who currently have data science responsibilities as a core part of their role, this is not the result of careful career planning and selective training. Typically their career route is characterised by accident rather than design. The following scenarios were regularly cited by participants in our study:

- Research groups. Within research groups, an individual is nominated by the Principal Investigator to be responsible for the data science function. Typically these people will have demonstrated some aptitude for information management and they may opt to go on and make it their main role if the appropriate opportunities arise. Sometimes the appointment will not be welcomed wholeheartedly: researchers near the beginning of their careers tend to view such appointments as a distraction.

- Data centres. When interviewed, many data scientists who work within data centres said they did not consciously choose their current line of work. Many started out with the intention of working in a data centre for a year or two before moving to a more permanent position in the public or private research arena, yet they are still with the same organisation several years later. Why do they stay? For those with a research background, people cite the attraction of being involved with the research community (without the requirement to teach) but they also like the stability of working for a single employer where they are not required to apply for grants, to work on short-term contracts or to move around the country to keep their career moving.

- Subject-specific institutions. Among the managers of institutions we interviewed, there have been two sets of problems. The first is the difficulty of recruiting data scientists with the right combination of domain, technical and interpersonal skills. Invariably there is a requirement for additional training once somebody has been appointed. The second – and most important – problem is that the researchers at all levels in the institutions neither fully understand the need for data scientists nor do they value data scientists. This, coupled with the lack of any obvious career path leads to dissatisfaction and data scientists leaving. It appears that, in some institutions at least, data scientists are spending more time trying to explain and justify their role rather than getting on with the job. This is an issue for senior managers at institutions and their funders to address.

Whilst there is some variation in the skills that are possessed by data scientists it is true to say that all these individuals now possess at least a substantial competency in the domain in which they operate. For example, in systems biology and associated fields like genomics most data scientists are 'from the domain', having migrated into a data science role from a place at the bench. Many embedded data scientists have doctorates in their domain. Indeed, many of them still consider themselves to be active members of the experimental team whilst also carrying the responsibility for data-related matters for the group and thus specialised in this sense. A few have arrived in their posts from a non-domain background, most usually in computer science or information science. This state of affairs appears to hold true across most disciplines.

During the course of our research we encountered few people who had been specifically trained in data science skills. Most of them have acquired their specialist data science skills on-the-job in an *ad hoc* manner. That said, it seems that requirements of candidates for *new* data scientist posts, or replacements for data scientists leaving their existing posts, are more than likely to include some informatics training, thus beginning the formalisation and professionalisation of the role. Here is an extract from a recent advertisement for a postdoctoral scientist for a data management [their terminology] role in a systems biology research group:

> *The ideal candidate will have a background in the physical or computational sciences, a Ph.D. or equivalent in bioinformatics or related discipline and should have extensive and relevant experience of database technologies, software development, biomolecular simulation or microscopy related imaging data. The successful applicant will have a strong publication record, and a proven track record in the use of database management with respect to biomolecular data. Experience of Molecular Dynamics data would be an advantage.*

In certain disciplines informatics is a specialised and well-advanced part of the postgraduate curriculum. Two examples of this specialisation are bioinformatics and chemoinformatics, both of which are offered at masters level in many UK universities. But just as informatics training is not a necessity (yet) for a data scientist, it is also not the panacea. Data scientists we spoke to were of the opinion that informatics training, at least to master's level, does not provide the full depth of domain research knowledge needed for a data scientist role. And one director of a research institute who teaches a postgraduate informatics course said that the course does not fully equip graduates for a data science role. Further personal development and learning whilst in the job are needed.

The majority of data scientists who participated in the online survey for this study have a master's degree and around one third have another relevant post-graduate qualification. As to the mix of skills required, our study revealed that there is no overall agreement among data scientists on whether it is essential for them to have a degree or higher qualification in the subject areas they serve.

Incumbents in the role of data scientist may be domain experts who have picked up advanced computing skills, or they may be computer or information science experts who have assimilated domain knowledge on the job. The former – at least in the experimental sciences – are the most common but not overwhelmingly so and computer scientists can and do assume data scientist roles by assimilating the required subject knowledge over time. We came across one data scientist in a research group, a computer scientist by training, who had spent two weeks embedded in a subject-relevant research group (before he sought his present job) to familiarise himself with contemporary domain-specific concepts and methodologies.

Do data scientists need to have some research experience in a relevant field in order to be effective? The answer from data scientists who have a research background is a resounding "yes". They argue that it would be very difficult to do their current jobs without a detailed understanding of the subject area they are working with. This view is based on the requirement to understand the particular characteristics of information associated with particular subject areas, but also on the basis that having an in-depth knowledge of a subject area facilitates communication between the data scientist and the other members of the research group. It has also been reported that researchers (with a track record and several publications to their credit) who become data scientists tend to be respected by their peers, and that this parity enables the data science process to be conducted more smoothly and efficiently than would otherwise be possible.

In many ways this is the ideal scenario, one where researchers take a career detour to focus on data science. One problem appears to be that researchers often do not want to take such a detour and, of those that do, some may wish at some point to return to full-time research. This is not because they don't recognise the value of data science, but because they want to get back to focusing on the challenges presented by research. Lucky ones get the opportunity to carry out both data science and research work in their job.

On the other side of the coin, there are data scientists who argue that it is not necessary to be a subject expert in order to do the job effectively. There are some fundamental data science skills that are generic in nature, such as dealing with confidential research, data description and metadata, software, copyright and intellectual property rights, and data storage. Although this is may be so, the core issue is that of effective communication between a data scientist and their research colleagues. There is a general acceptance that, for instance, people trained in an arts-based subject might find it challenging to deal with astronomical datasets and that the period of time required to acquire the necessary insights and skills would be detrimental to the a research group's schedule and budget. A compromise position would be that data scientists should have degree-level education in a relevant subject area. Indeed, this parallels the position increasingly being adopted by library schools that offer digital curation as part of their master's level library science courses.

From a practical perspective, as demand for competent data scientists grows, so it will become necessary to cast the net as wide as possible. Subject knowledge is important, but so too are technical skills and people skills. Subject knowledge deficiencies can be offset by people with the facility for effective communication. Data scientists who have the personal attributes necessary to win people over and facilitate effective interaction within the research group can be very productive.

We must consider also the question of technical and computing aptitude. Data science does require a good degree of understanding of issues like data formats, digitisation, data storage, data access and security. The role normally requires practical computing skills as well as a sufficient knowledge of technical terminology to be able to communicate effectively with, for example, an institution's computing centre staff. Our study indicated that nearly half of data scientists believe it to be essential for people in their role to have a technical or computing background.

Our online survey of current data scientists also showed that the data science community is evenly split on whether people skills are more important than technical skills for success as a data scientist – but then people's opinions are often predicated on their own experiences and whether their own strengths lied toward the technical or people skills end of the spectrum. It is uncommon to find people who are excellent at both. We came across several examples of instances where people whose background was primarily computing and information technology became sufficiently familiar with the subject area of their specialist institutions that they were deemed to be effective data scientists. There are some in senior positions who hold the view that it is perhaps better to employ someone with a technical background and teach them the fundamentals of the subject area rather than vice versa. The downside, of course, is that this process of familiarisation can take considerable time – possibly years. In some fields of research there are greater numbers of highly educated people than jobs available on the market. In these circumstances employers are often able to find candidates who offer degree-level content knowledge plus master's degrees in computing or information technology skills.

4.4 The positions data scientists hold

Data scientists could conceivable work in any situation where research data is produced. A list of the main types of work situation is presented below. As part of the study current data scientists were asked what sorts of roles were appealing to them from a professional fulfilment perspective. These results are, of course, born of personal preferences and backgrounds, but give a useful overview of the current perceptions of data scientists with respect to the type of organisation that appeals to them. The three most popular choices were:

- Data scientist in a subject-specific research institution
- Data scientist with a research council
- Data scientist embedded in a research project team

The three least popular choices were:

- Data scientist affiliated with a computer centre
- Data scientist with a Data Support Service
- Data scientist in a large data centre

Occupying the middle ground are data science jobs affiliated with a library or with a small data centre.

4.5 Job security

The data scientists we studied fall into two main categories regarding their security of tenure. First, there are those who are employed in tenured posts in universities, research centres or data centres. Second, there are those employed on short-term research grants.

4.5.1 Tenured data science posts in universities and research centres

Some data scientists hold fully tenured posts in UK HEIs. Their positions vary. Some universities have specific job grades that they assign to such posts and they may be aligned with either the academic grades or the technical grades. One example is the *professional service* grade at Imperial College London. Those on quasi-academic grades (we use this term because they may not always be the same as 'academic-related' roles that are common in UK universities) may find it difficult to move into academic or academic-related positions because they are not producing publications, the usual currency for measuring 'worth' in academic research. In other universities, tenured data scientists may be on academic grades. The evidence suggests that these are few in number at the moment. As data-based projects become more common and data science grows in importance in universities we would expect some changes to universities' thinking on this issue: the need to recruit and retain highly-skilled professional data scientists will inevitably become more pressing.

4.5.2 Data scientists on short-term contracts

Many data scientists are employed on short-term contracts. These are not always single-term: they may be renewed or arrangements may be made in the institution hosting the project for the post to be included in further rounds of grant applications, as has been the case in some of the university-based astronomy data centres, for example.

Short-term positions may be perfectly appropriate: a discrete piece of research is to be done, it is carried out, the data are properly managed and curated during the course of the project and stored somewhere safe for future re-use. In many areas of academic research, and in many other areas of academic research, such an arrangement, with the data being deposited in an appropriate public databank such as GenBank, may be a satisfactory model. In other cases, exemplified well by the current situation in astronomy, very senior scientists remain on rolling contracts indefinitely. These are highly-skilled individuals, experts in data science, with skills that have taken years to acquire. They are crucial to the operation of the research group yet while they are very highly regarded within that group, they have little 'official' academic status and no defined career track.

4.6 The supply of data science skills to the research community

The role of data scientists is set to increase in importance as the value of research data skills becomes more apparent to the research community and as funders seek greater returns on their investment in the research process. At present, it is doubtful that there will be sufficient numbers of appropriately skilled and experienced data scientists to meet the growing need in the short to medium term. Currently, demand outstrips supply. We were told anecdotally by two groups that they had advertised posts but been unsuccessful in making a suitable appointment. This is not because the job is intrinsically unappealing: at least half the data scientists who participated in our survey believe their career to be professionally rewarding, and around one quarter enjoy considerable autonomy in their jobs. There appear to be a number of fundamental problems with the career structure within which data scientists in the UK currently work. The key problems identified by participants in our online survey are listed below:

- Two-thirds of data scientists disagree with the notion that it is obvious what formal training is required to pursue a career in data science.

- Many data scientists do not think that formal training is available for those wishing to pursue a career in data science.

- More than half of data scientists feel that the role of data science is neither understood nor respected in the research community.

On the whole, research teams tend to do three core things: write grant applications; conduct research; and write papers for publication. The issue of data science comes much further down the list of priorities and in some instances is ignored. The natural corollary is that the role played by data scientists is not necessarily highly valued and in many of the instances we investigated for this study the role of data scientist is reported to be viewed as a cost burden rather than a benefit, despite the fact that data scientists have an increasingly important role in all three core activities listed above. The lack of understanding of the importance of data science by many researchers and the concomitant shortage of professional respect for the role of data scientists working in research teams can be reflected in their terms and conditions.

Other disincentives to choosing and pursuing a career in data science given by our online survey respondents include:

- Few current data scientists agree with the idea that there are many permanent data science jobs available in the UK.

- Over one quarter of data scientists think that data science jobs tend to be short-term (and half don't know).

- Data science jobs tend to be on technical rather than academic grades, and only 10% of data scientists believe that, on the whole, data science jobs are remunerated fairly.

And on the issue of career progression:

- Two-thirds of data scientists believe that there is no well-defined pathway for people who wish to pursue a career in data science.

There is no formal career structure at present, and certainly not compared with a career as an academic researcher or a technician. Instead, people working in data science in universities do a specific job with limited opportunities for career progression in the traditional sense. People can, of course, enrich their roles through the acquisition of new skills but when they decide to leave their current employer the move tends to be sideways – to a similar position in a different organisation – because the scope for career advancement in terms of hierarchical grade and salary level is very limited. In fact data scientists report that the development of very specialist skills tends to tie them closer to their current employer because they perceive those skills to be non-transferable. This may change over time as those skills and experience are more widely valued and needed.

The situation is somewhat different in dedicated data centres where staff may be on Public Service Grades and where the possibility of career advancement exists – though in reality the pyramid is relatively flat and it can take many years to progress up the ranks. The process can be accelerated for a talented few, but often younger recruits tend not to stay long given the dissonance between their career aspirations and the reality of organisational constraints.

There is some evidence that researchers with data science experience may have a "value added" competitive edge in the jobs market. In recent times most research councils in the UK have become more insistent that the research teams they fund pay closer attention to data planning and to data science in general. As the need for effective data science practice gradually permeates the research community, so data science skills will become more valuable – not least since they are in such short supply at the moment. As we have already noted, however, some researchers who have responsibility for data science issues at present would prefer to return to their original career – doing research – on a full-time basis. This tendency will naturally serve to limit the pool of data scientists with research backgrounds. There are others, however, for whom the attraction of research has lost its allure and for whom the relative stability of a data science job is appealing. In this context it was reported, for example, that researchers wanting to return to work – possibly part-time – after a spell of childcare are attracted by data science positions.

5. TRAINING PROVISION

5.1 Introduction

Data scientists are very pragmatic when it comes to the mode of training they perceive to be important in motivating them to attend training courses. In particular, according to our online survey, they value the following:

- Hearing about actual experiences from practitioners
- Acquiring practical skills immediately relevant to their work
- The provision of hands-on, practical exercises

Data scientists also acknowledge that the skills they have now will not be adequate in a few years' time, if that. Things are moving very quickly in this area and a programme of continued upskilling will be necessary if data scientists are to stay abreast of developments in their field.

Data scientists have a limited appetite for formal assessment of their participation in training courses, and only about one quarter of current data scientists think that receiving formal certification or accreditation for completing a training course is important. This bias towards practical training which people can apply to their work situations reflects data scientists' tradition of learning “on the job”. We discuss in more detail below their preferences.

We also lay out an overview of other types of training provision, not only for data scientists but also for the associated supply of specialist data curation skills to the research community. Our own study for the RIN, JISC and NERC (Brown and Swan, 2007), and the recent study on data practices in Australian universities (Henty et al, 2008), both provide evidence that researchers are in overall favour of sharing data, recognise the importance of looking after their data better and in general wish to learn more about developing good data skills but need help to do so. These are not practices that researchers can easily pick up in passing and they understand this. Moreover, there are data-related skills that the data scientists themselves do not necessarily themselves need to practice. The archiving and preservation of data for the longer term can be seen as more properly part of the library's role, and where the institutional repository (or, in some cases, a separate institutional data repository), part of the library's domain, comes into its own. This, of course, means that relevant data skills must also reside within the library community. We discuss both these issues in this chapter.

5.2 On-the-job skills development for data scientists

Data scientists do not arrive in a job fully equipped with all the skills to take them forward *ad infinitum*. They need continually to update themselves on new developments and techniques. The majority of data scientists who participated in the survey for this study say that they learnt their trade “on the job”; indeed our survey showed that two-thirds of respondents were self-taught. They can acquire skills on the job by a variety of means. These may be very formalised, such as specific training courses run by the manufacturers of laboratory machines, through occasional short courses, to day-to-day pragmatic rising to new challenges brought before them by the experimentalists. The notion of specialised short courses is received with considerable enthusiasm in the data science community and we discuss this further in Section 5.4. In some organisations, notably data centres, data scientists can call upon the expertise of more experienced colleagues when they need advice.

Given that data science requires a mix of technical skills, people skills and subject knowledge – an uncommon and valuable combination of attributes – it is interesting that such a large proportion of the existing community of data managers have taken this rather informal on-the-job route to skills acquisition. Their answers in the onlline survey gave some indications as to why, identifying the following two reasons as being particularly important.

- Lack of suitable training opportunities (and lack of knowledge about them)

It is the nature of data science that each organisation – whether it is on an institutional or small research group scale – has different needs and priorities. The people appointed to be data scientists often perceive that their needs are so specific or unique that the training courses that are currently available do not adequately address these needs. Indeed, nearly half of the people we contacted said there was a lack of training opportunities at the level they require. This is in spite of the fact that most data scientists say that they are prepared to invest time attending training and skills course for data science or digital curation. Just over one third of participants in our poll of data scientists said they would be prepared to attend courses of more than five days duration and some of the people we spoke to during the study were registered as participants at various research data skills events that are running over a number of days in Europe during this summer. So some people welcome the idea of longer courses, but overall the preference would seem to be for intensive short courses, a notion that seems well-received by most data scientists.

- Cost and location of training events

Two of the most common factors militating against training are lack of time and lack of funding. Even if suitable training courses are available, the costs of travelling in terms of time and money are significant obstacles for around one third of data managers. For a smaller proportion, the lack of organisational support is perceived to be an obstacle to attending external training courses.

5.3 Formal postgraduate training

In addition to specialised informatics training which is increasingly offered by UK universities, other domain-specific master's degrees do include some elements of data science. Most taught master's degrees in disciplines with a high level of digital data generation now include aspects of data handling, data care and data curation as part of research practice teaching. In addition, PhD programmes will always include the required level of training in the specific practices of the domain relating to information collection, recording, manipulation, storage and so forth. All these things can be expected to persist and increase as data skills become increasingly important.

We distinguish between two constituencies that could undertake formal postgraduate training in data science: first, there are those who wish to become data scientists themselves; second there is the main body of researchers who, whilst not aspiring to leave research proper and enter data science as a career, can benefit from upskilling in data-related matters.

5.3.1 Training for data scientists

As stated earlier, most of the data scientists who responded to our survey have a postgraduate degree in a discipline relevant to data science such as computing or informatics (though we encountered many others who had domain-training only and had learned to be data specialists on the job). The most relevant postgraduate qualification is a master's degree in informatics. More and more universities are offering these. In some cases a broad-brush informatics degree is delivered by the computing or information science department[5], but in a number of instances these days the course is domain-focused, the most common examples being bioinformatics and chemoinformatics. As more disciplines migrate into data-intensive research areas other domain-related informatics courses will increase in number.

The important thing about such discipline-oriented approaches to training in informatics is that data manipulation and use does have a strong disciplinary nature. Datasets generated in molecular biology research, for example, are dissimilar enough amongst themselves for their integration (mashups) and cross-interrogation to pose severe challenges to even skilled data scientists in that field. They bear no resemblance at all to datasets in artificial intelligence, astronomy, anthropology, archaeology or area studies and the techniques used to manipulate datasets in a particular discipline may only be similar in the most basic of ways to those used in another. Discipline-specific courses are therefore a needed approach as research becomes increasingly data-intensive and data-handling skills become part of a researcher's toolbox.

5.3.2 Training for researchers

There is a clear need to provide training in data skills for all researchers at postgraduate level. This type of training has three potential benefits: first, it helps researchers understand the importance of the data lifecycle and the role they need to play if research data are to be generated, handled, curated, archived and preserved successfully; second, it may provide the impetus for some researchers to go on to become data scientists themselves; third, this type of training will equip new data scientists with the basic knowledge they need to get started.

Although courses specifically about data science and management are not yet commonly found, they are starting to appear. Two examples are the new master's course offered by the Centre for Computing in the Humanities at King's College London[6] and the MSc in Information Management and Preservation taught by the Humanities Advanced Technology and Information Institute based at the University of Glasgow[7].

Non-degree-level courses are also becoming available. The UK Data Archive has already developed a set of training and support materials that cover the key areas of data science and management. Eight modules cover subjects such as: developing consent agreements; anonymisation techniques; data description and metadata; data formats and software; copyright and IPR; data storage, backups and security; digitisation and providing access to data. The materials were designed for researchers who produce and work with social and economic science and humanities research data, but the concepts are generally applicable to all research data. The UKDA has already run a number of training events and more are planned, but ultimately the materials could be adapted by trainers in different subject areas and delivered to researchers across the UK.

The Digital Curation Centre also has an important role to play in the delivery of postgraduate training. The DCC is currently finalising a set of modules that focus on the data life cycle model. Although Digital Curation 101, as the course will be known, is

[5] For an example, see the University of Strathclyde: http://www.gsi.strath.ac.uk/
[6] http://kcl.ac.uk/iss/cerch/teaching/
[7] http://www.hatii.arts.gla.ac.uk/imp/index.htm

primarily designed for bench scientists and information specialists, the modules could be adapted to be relevant to researchers across the disciplinary spectrum.

Finally, a working group was recently convened to discuss developments in the provision of data curation training and education, drawing on current best practice from the UK and the USA. The group – known as the International Data curation Education Action (IDEA) working group – is exploring how best to move forward with developing and delivering appropriate training and education, and how to promote digital curation as a profession in its own right. The goal is for IDEA to widen participation and continue its work through future meetings and ongoing cooperation between members.

5.4 Continuing Professional Development (CPD)

People already in data science posts require specialist training in particular topics. More generic approaches will necessarily be of limited value, but helping data scientists overcome some common challenges in each domain would be very helpful. Facilitating their coming together would be a major part of the task, for there is much they can learn from each other. For example, data integration is both a huge challenge and of huge importance. At present, learning to integrate datasets of different origin is something done on the job in an *ad hoc* manner or by swapping experiences and notes on practice with other data scientists. Formalising such interactions and exchanges by the provision of workshop-type events would be a positive contribution to spreading advanced skills through the community.

There is an argument, well made, in the data science community that data issues are moving so fast that periodic updating is much more effective than an early, intensive training with no follow-up. To this end, various organisations have already begun to offer training courses and workshops on data-related topics. Some varied examples of events are:

- The Digital Curation Centre has recently convened the first meeting of the Research Data Management Forum[8], a group that comprises data scientists and data managers from across the UK in any discipline. The Forum provides the opportunity for data professionals to exchange experiences and best practice
- The Pan-American Advanced Studies Institute in Cyberinfrastructure for International Collaborative Biodiversity and Ecological Informatics[9]
- The Evry workshop on Data Integration in the Life Sciences[10]
- The ERASysBio Initiative's Data Management Summer School[11] (part-funded by BBSRC)

In a UK context, the data science community looks favourably upon the notion of a CPD programme that addresses their professional development needs. Data science certainly is a fast-moving field and the chance of some formal instruction in topics of interest, plus the opportunity to get together and discuss their experiences and practices, is a popular notion.

A number of players are appropriate for delivering this sort of specialised CPD training. The Digital Curation Centre is one and in areas where the honing of computing skills is needed The British Computer Society is another potential deliverer of CPD training, as is the National e-Science Centre. The type of CPD training that data scientists favour is focused,

[8] http://www.dcc.ac.uk/events/data-forum-2008/
[9] http://ciara.fiu.edu/eco/
[10] http://dils2008.lri.fr/
[11] http://www.erasysbio.net/

intensive, short courses on contemporary topics. It was suggested by people we spoke to that the suppliers of such training might consider bringing in experts in specific disciplinary data issues – such as integrating datasets or on the manipulation of specific software – to deliver such courses. Half-week to full-week courses are favoured most.

The provision of CPD to people specialised in clinical data management is already well organised and may provide a model for professionalizing data science in other fields. The Association for Clinical Data Management[12] provides an evolved programme of education and training (including NVQ certification) in addition to networking opportunities such as conferences, special interest groups and technical meetings.

Alongside the provision of such training, there needs to be recognition at institutional or funder level that a long-term commitment to it will be required. Data science moves fast and those riding the bow-wave will need periodic but persistent upskilling.

5.5 The undergraduate curriculum

It is true that as time passes, data handling skills will become more and more a part of the researcher's basic skillset. 'Native data scientists' will become the norm in all disciplines. Already, many undergraduate courses include lectures or modules in data-related topics. These may be fairly basic, such as teaching the use of Microsoft's Excel programme or basic statistics, but everyone in data science agrees that it is never too early to start people thinking about the big picture. Indeed the data scientists we spoke to were of the opinion that if every researcher had even basic data management skills then most of the current day-to-day problems for data scientists would evaporate, leaving them to deal with the more challenging and specialised tasks.

There are two views on the issue. First, some people think that the undergraduate curriculum is crowded enough now without adding more compulsory topics to it. These people were in a minority of those to whom we spoke, but their point is valid. In some disciplines, at least, there is the feeling that there is barely time to cover sufficient disciplinary information to merit awarding a degree in the discipline and that research-related skills, important though they are, should appear in postgraduate taught courses or as part of PhD programmes. The other view is that data skills should be viewed as a fundamental part of the education of undergraduates in the same way as basic statistics, laboratory practices and methods of recording findings are. It is likely that this latter view will prevail eventually. There are basic data skills and principles that can be taught generically – relational databases, XML, the principles of curation, documenting work, task-tracking and so forth. These may find a place in the undergraduate curriculum, possibly at the time of the research project in those disciplines that include such a thing.

5.6 The role of the library

The library and information science community should have an important role to play in the data science arena, particularly in delivering awareness and understanding of data issues and the importance of good data science and data curation. There are generic data handling and management skills that are native to librarians and can be taught as part of the basic research skills training in an institution. After all, the fundamentals of data science can be taught and subject expertise can be acquired over time. There are also

[12] http://www.acdm.org.uk/

other roles that libraries can play here. We suggest that three of the most relevant ways in which the library community might influence developments are:

- Training researchers to be more data-aware
- Adopting a data archiving and preservation role
- The training and supply of data librarians

5.6.1 Training researchers to be more data-aware

This will require a shift of emphasis because, although for example library schools in the UK already collaborate with academic departments in teaching general research-related skills (such as information literacy) and sometimes in very specialised programmes (such as chemoinformatics), they have generally seen themselves as more concerned with information than data. In addition, involvement in the research programme is admittedly more difficult to achieve than involvement in undergraduate education. Libraries usually offer information literacy programmes to undergraduates, for example, but it is uncommon to see these penetrating the research programme and, if they do, they are very rarely compulsory. Nonetheless, it is likely that the data deluge will change things. The new Australian study (Henty et al, 2009) clearly indicated the research community's interest in, and desire to know more about, good data practice and our own work in this area corroborates this (Brown and Swan, 2007). We have also ehard from librarians we spoke to for this study that they are increasingly being approached by researchers for advice practical help with data management. Libraries should be gearing up to raise data-awareness because the demand for this is already growing.

5.6.2 Adopting a data care role

The growing need for data archiving and preservation capacity offers libraries a strategic opportunity to reposition themselves with respect to research. There is much discussion and planning on this topic in the library community (see, for example, Steinhart et al, 2008) and how libraries can best serve the e-science agenda (Martinez, 2007; Carlson, 2006). Many have accepted the challenge of developing institutional repositories and the natural extension of these is into the realm of data. Many librarians are repositioning their libraries to take on the role of caring for data on behalf of the institution and data scientists we interviewed believe that libraries should indeed be responsible for data archiving and preservation. They believe this would free their own time to focus on working with researchers on different (domain-specific) data challenges. There is a difference between archiving and preservation activities and data science in all its manifestations, though, and senior librarians caution that subject or liaison librarians will have a limited amount to offer in terms of data science itelf. Nonetheless, there is much scope for interactional learning between library staff and data scientists and it is hoped that this will become a norm as the academic community as a whole shifts into new ways of working, exploiting potential synergies. An assessment of the potential role of university libraries in harvesting and curating datasets is beyond the scope of this study though it is looked at by the DISC-UK DataShare project[13].

5.6.3 The training and supply of data librarians

Library educators have an important role to play in planning for and delivering appropriately skilled people to meet the latent demand for data librarians to manage the libraries' potential data curation role). Yet very few library and information science schools currently teach the skills that future data librarians will need.

[13] http://www.disc-uk.org/datashare.html

Of the 55 institutions that teach Information and Library Science (ILS) in the United States, for example, only a handful include any digital curation content in their courses. A significant amount of effort has been invested in the development of a digital curation curriculum[14] at the University of North Carolina, Chapel Hill. The same team are organising a conference to move the debate along.[15] The team at the Center for Informatics Research in Science and Scholarship[16], University of Illinois at Urbana Champaign, has also been working to see how best to prepare librarians for a data curation role. In the UK, the library school in Sheffield participates in teaching a short course in chemoinformatics but in general most library schools include a digital data management as an element in the MLS course rather than offering a special option in this area. Although there are some individuals in the UK who are called data librarians, it is thought these currently number around five.

There appear to be a number of reasons why library and information science schools are not yet producing librarians with good digital curation skills in anything approaching significant numbers:

- There is currently no clear career trajectory for data librarians and for graduate students who have to pay for their tuition and living expenses it would require a leap of faith to deviate far from regular library and information science school curricula. Many potential students know very little about the subject area, what the courses will be like, the likelihood of finding relevant work when they graduate, and this uncertainty can make such courses too risky for some. Although a lot of valuable work has been done to develop the foundation of a digital curation curriculum, there is still a way to go before this is distilled down to a canon of topics. This explains why even leading library schools have to “soft pedal” the data curation aspect of their general Master of Library Science (MLS) courses. By just including elements of digital curation in an MLS course, students are able to hedge their bets. Library schools are having to actively promote the potential benefits of studying a course with digital aspects to prospective students, including the prospect of a career in “library information technology” – this being a known quantity.

- Library and information science schools in the US that have attempted to teach graduate courses with a concentration in digital curation have experienced problems trying to find suitable internships for their students. Work placements with a genuine digital curation element can be very difficult to find within an institution – if they exist at all. Until more libraries and institutions are engaged with digital curation, the lack of internships will continue to be a bottleneck in the teaching and learning process.

- The library and information science schools with experience of teaching courses with digital curation content have determined that, in order to do well, students need at least a degree-level grounding in the area of scholarship they plan to work in. The library school graduate employed to undertake a digital curation role must have an understanding not only of the data but also the work practices of the researchers with whom they are likely to be working. Clearly this makes the process of attracting and recruiting students more difficult. Some people argue that since librarians have, in the

[14] http://ils.unc.edu/digccurr/aboutI.html
[15] 15 http://www.ils.unc.edu/digccurr2009/
[16] http://cirss.lis.uiuc.edu/index.html

past, developed into the role of subject experts, that they should be able to do the same in the digital data arena. Others, however, say that over recent years the trend has been towards seeing the role of subject librarian being replaced by that of liaison librarian with a focus on working with departments to deliver what they need – which may include collection issues and skills training – but which does not imply or require a particularly developed level of subject knowledge.

6. DISCUSSION

These are early days in data science. There is a data science protocommunity in the UK that is beginning to gel into something cohesive, aided by initiatives such as the DCC's Data Managers Forum. Much lies ahead in terms of community-building and professionalisation.

Data scientists, working in close collaboration with data creators, will have an increasingly important role as funders and the research community come to appreciate the value of curating and re-using research data. Ideally, data scientists will have a research background together with a technical aptitude and finely-tuned advocacy and interpersonal skills. People with this combination of attributes together with data science experience are in short supply and employers often have difficulty filling vacant data science positions.

The supply of appropriately qualified data scientists must be addressed. In the first place funders and leaders in the research community need to promote the value of the role played by data scientists to the research community at large and thought must be given to recognising data science as an academic career in its own right. At present the potential for career progression and the scale of incentives and rewards is limited.

There will also be an increasingly important role for data librarians who will curate, preserve and archive digital data. For many, this is a logical extension of the core mission of libraries which serve scholarly communities and extends naturally from the research-supporting role of institutional repositories. Librarians working in this field will liaise closely with domain-specific data scientists. Work by library educators to produce a digital curation curriculum is well under way although at present digital curation skills are not widely taught by schools of library and information science. The senior management of libraries will need to consider how best to foster the nascent field of data librarianship by devising appropriate career structures.

The discussion around the need for libraries to take a proactive stance highlights one of the main problems at the moment. There are numerous stakeholders in the data-driven research world – researchers themselves, libraries, the UK research councils, other research funders, IT people and last, but not least by any means, institutions. At present none of these stakeholders has taken a leading position in directing progress. There is a big question – which side of the dual-support system should shoulder the responsibility for data science?

The answer is that neither side can shrug off the responsibility. 'Big research', funded naturally by the research councils, is where data scientists are already at work, but where a system for professionalisation, proper career structures and due recognition are not yet fully in place. 'Small research', the more modest research programmes that produce so much data at present and are going to produce so much more, has to be the responsibility of the institutions in which it takes place for there is no other entity to own small research and its data. Data scientist skills will be needed here, too, at institutional level, to ensure that researchers in all disciplines are provided with the expertise they need to help them create and manage their data optimally.

Resolving the future needs of the research community will therefore be highly dependent upon research institutions taking data seriously. The library has a role, as discussed, in supporting research by supplying data archiving and preservation services to the institution, but the supply of data science skills will need to be planned, appropriated and managed at institutional research management level.

That such arrangements are in place only in a small minority of institutions is not surprising. Data-related matters have moved so fast in the last few years that institutions and funders have had little chance to orientate. Structures, processes and policies have not kept up with data-related activity.

The overall lesson from this study is that UK data science remains, if not in embryonic stage, then at least in its infancy. There are examples of best practice in terms of skills – such as the data centres – and there are examples of best practice in terms of professionalisation of the role – such as the Association of Clinical Data Managers[17]. The main problems concern a lack of career structure for data professionals, a loose definition both of roles and of the skillsets required in the various circumstances that the research community presents and the absence of a clear picture of how data science can be progressed. We hope the recommendations in the next section may help the process of resolving some of these issues.

[17] http://www.acdm.org.uk/resources.aspx

7. RECOMMENDATIONS

The main recommendations from the study are as follows:

1. Recommendations regarding data skills development in research domains (RD):

Recommendation RD1**:** Major research funders in the UK should work with universities and research institutes to define properly and to formalise the role of data scientists, and to develop the means by which the work of data scientists can be recognised and remunerated.

Recommendation RD2**:** These same bodies should work together to create the conditions that support data science, foster its study and encourage professionalisation of the role.

Recommendation RD3**:** The JISC and other organisations that commission original research should take forward a study (or studies) that cover the following issues:

- A description of the role played by data scientists and the value of the contribution they make to research
- Examples of data science careers
- The development of a set of practices that represent good practice in data science

Recommendation RD4**:** The relevant bodies (HEFCE and the research councils) should consider the establishment and funding of a network of trainers with the skills to deliver short postgraduate training courses to researchers covering the fundamentals of data management, thus building basic data science skills into the research process. Some of the research councils have laid the foundations for this with their requirements for a data plan in grant applications.

Recommendation RD5**:** The research councils and other research funders should consider whether, as part of the grant application and award process, they should require at least one member of the project team to be nominated as the project's data scientist. This person should be required to attend a short course covering the fundamentals of data science and management. Research councils should consider the extent to which accrediting valid courses and proof of attendance is necessary.

2. Recommendations regarding data skills development in research libraries (RL):

Recommendation RL1**:** The research library community in the UK should work with universities and research institutes to define properly and to formalise the role of data librarians, and to develop a curriculum that ensures a suitable supply of librarians skilled in data handling.

Recommendation RL2**:** The JISC should consider supporting the development of the International Data curation Education Action (IDEA) working group. This group is well-placed to play an important advisory role in the development of appropriate curricula for future data librarians, particularly those coming through the library and information science route.

3. Recommendations regarding data skills development in general (RG):

***Recommendation RG1*:** Because there are already a number of players active in the data area there is potential for exploiting synergies in respect of data skills training. It is recommended that a study scopes this potential, looking in particular at the activities of the UK Data Archive, universities or research groups where data science is advanced, library schools, the Digital Curation Centre and IDEA (the International Data curation Education Alliance). The study might also look internationally at initiatives in the US, Canada and Australia.

References

Brown S and Swan A (2007) Researchers' Use of Academic Libraries and their Services (2007). Published by RIN in association with CURL. http://www.rin.ac.uk/researchers-use-libraries

Canadian Digital Information Strategy (2007). Library and Archives Canada. http://www.collectionscanada.gc.ca/cdis/index-e.html

Carlson S (2006) Lost in a sea of science data. *Chronicle of Higher Education*, 23 June. http://chronicle.com/free/v52/i42/42a03501.htm

Greer C (2007) A vision for the digital data universe. Presentation. https://www.nanohub.org/resources/2291/

Henty M, Weaver B, Bradbury S and Porter S (2008) Investigating data management practices in Australian universities. Australian Partnership for Sustainable Repositories (APSR). http://www.apsr.edu.au/investigating_data_management

Lyon, Liz (2007) Dealing with Data: Roles, Rights, Responsibilities and Relationships: http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dealing_with_data_report-final.pdf

National Science Foundation, National Science Board (2005) Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century. http://www.nsf.gov/pubs/2005/nsb0540/

Martinez L (2007) The e-research needs analysis survey report. CURL/SCONUL Joint Task Force on e-Research. www.rluk.ac.uk/files/E-Research**NeedsAnalysis**Revised.pdf

Steinhart G, Saylor J, Albert Paul, Alpi K, Baxter P, Brown E, Chiang K, Corson-Rikert J, Hirtle P, Jenkins K, Lowe B, McCue J, Ruddy D, Silterra R, Solla L, Stewart-Marshall Z, Westbrooks EL (2008) Digital Research Data Curation: Overview of Issues, Current Activities, and Opportunities for the Cornell University Library. A report of the Cornell University Library Data Working Group. May 2008. http://ecommons.library.cornell.edu/handle/1813/10903