CHAPTER 1

Why manage research data?

Graham Pryor

Consider this: every year around £3.5 billion of taxpayers' money is spent on research undertaken by UK universities. That is a considerable level of investment and members of the public might reasonably assume that all due care will be taken to ensure this national endowment is applied wisely. Justifiably, they might also expect that the fruits from such rich endeavours will be afforded the attention necessary to ensure an optimum return on investment. So what are these 'fruits' and how indeed in practice is their value observed?

A challenge for the information professional

Way back in 1979, Dennis Lewis, then head of ASLIB, the UK's Association for Information Management, wrote what came to be known as the Doomsday Scenario for librarians. His main claim was that information professionals wouldn't be around by the year 2000, meaning that the *types* of information professionals he saw working in 1979 (mainly librarians and information scientists) would be long gone, swept away by a new information age that would by implication belong to computing scientists. In the event, it didn't happen quite like that and the traditional custodians of documented knowledge, with their armoury of skills in appraising, classifying, preserving, storing and retrieving information somehow managed to reinvent themselves as the purveyors and stewards of digitally encoded knowledge, albeit they have confined themselves in the main to handling published materials. But the challenge has not gone away. Inexorably it has continued to increase in scale and complexity, re-presenting itself in a multiplicity of dimensions until, most significantly today, the accessible output from scholarly research is no longer to be considered exclusively through its documented or published form. We are talking here about digital data, a component of the knowledge enterprise that is as urgently in need of effective stewardship as any of the more traditional products of scholarly research.

This book is an attempt to explain to the library and information community what has to be done at a local, national or international level to engage with this fresh challenge, and why it is important that the traditional role of the information

professional should undergo some urgent re-tooling, not only to sustain the profession but also to ensure that the research community can benefit more completely from its centuries-old reserve of knowledge management acumen.

This introductory chapter will survey the key changes that have taken place in the information landscape during the past decade, within the research community and at the level of national and international policy. Each of the themes it introduces will be explored in greater detail in the succeeding chapters. For the information professional it is anticipated that such an approach will offer the prospect for familiarization with a new arena of activity, to enable gaps in understanding to be filled and for fresh workplace or career opportunities to be revealed. These are the fruits on offer within this volume. To resume that initial metaphor more appositely, our key questions must be: what should we recognize today as the fruits from scholarly research and what needs to be done to preserve and enjoy them?

The data deluge

Overwhelmingly, the output from research in the 21st century is data, produced chiefly in electronic form and having a scale of generation that is rapid, vast and particularly remarkable for its exponential rate of increase. Although this condition is to be found in all disciplines it is at its most dramatic in the sciences, where the annual rate of increase is in the region of 30%. Consider the biosciences, where the raw image files for a single human genome have been estimated at 28.8 terabytes, which is approaching 30,000 gigabytes (MacArthur, 2008). Or the high energy physics community, where the Large Hadron Collider (LHC) experiment at the European Organization for Nuclear Research (CERN), in Geneva, to which 19 UK universities have contributed, is expected to produce around 15 petabytes (15 million gigabytes) of data annually. A private individual attempting to store that quantity of data would require in excess of 1.7 million dual-layer DVDs! Yet the LHC is not unique among the global research community in generating massive data volumes.

Research programmes are funded and undertaken nationally and internationally. Expenditure on research can attain colossal proportions. In the USA, research spending on science and engineering alone reached almost $55 billion in 2009 (Britt, 2010), while between 2007 and 2013 the European Commission is spending €50 billion (£42.4/$61.5 billion) on its framework programme for research. Weighing the anticipated output from all of these programmes in a global context, it is easy to comprehend the source of the now-familiar *digital deluge*, a term that describes not only the data directly generated by these programmes but includes the further proliferation that occurs when they are shared or accessed by interested communities around the world.

At this point it is pertinent to be reminded of observations by research colleagues in the humanities that they don't actually work with data, an assertion based on the mistaken belief that data is exclusively the stuff of science, whereas as humanists they might claim instead to work with information and knowledge. Yet data is the primary

building block of all information, comprising the lowest level of abstraction in any field of knowledge, where it is identifiable as collections of numbers, characters, images or other symbols that when contextualized in a certain way represent facts, figures or ideas as communicable information. Moreover, in the digital age, the information and knowledge to which humanists will steadfastly lay claim can only be communicated to another person, whether across campus networks or via the internet, after they have been encoded as transmittable data.

In the specific arena of academic research, data is the output from any systematic investigation involving a process of observation, experiment or the testing of a hypothesis, which when assembled in context and interpreted expertly will produce new knowledge. So we all 'do data', whether we are humanists, scientists or social scientists. That data is a serious business for humanists too was underlined by the outcry in 2008 when funding was withdrawn from the UK's Arts and Humanities Data Service (AHDS), a national service established to enable the discovery, creation and preservation of digital resources across the arts and humanities research, teaching and learning community.

But while data may be the principal output from scholarly research, whatever the discipline, like the tip of an iceberg only a small proportion will be made visible. As the most conspicuous and probably the most familiar intellectual product of research that is conducted in a university, the scholarly article or paper has long been established as the means to deliver the results of experiments or the proof to a new hypothesis. For acceptance by a reputable journal, the organ through which the research paper will normally be assessed, published and delivered, the output from an often lengthy and laborious research process will have to be massively and selectively compressed. To be peer reviewed, to be selected from among and inserted alongside a host of competing articles, as well as to function both informatively and accessibly, published research can only represent those particular aspects of the experimental or investigative process that are essential to making the case and providing the necessary evidence to prove a hypothesis. Hence, a severe routine of selection, reduction and distillation from the greater expanse of experimental and evidential data generated and collected within a research programme will eventually reduce down to a publishable document, finely tuned to deliver a measured and measurable argument, with the greater volume of data from which the paper has been produced remaining hidden and largely inaccessible.

When there is such a focus on pruning research output to meet the strictures of the publishing process, should we expect that all possible value has been wrung from the broader wealth of data that was gathered or produced during the lifetime of a project? It is unlikely. And can we reasonably anticipate that it will have been left in a condition that will facilitate further use by the original researcher, or by others? Probably not.

The wealth of data and the merits of planning

From these initial observations it is evident that the data deluge of the 21st century is

a phenomenon that, if left unchallenged and unmanaged, is likely to result in considerable financial waste as well as opportunity loss. When considering the massive investment of time, intellectual effort and hard cash that goes into a research programme, should we not be expecting to draw more from the data generated than can be extracted from a well honed paper or series of papers? Surely data produced so expensively should not be treated as spoil, put aside like the waste materials from an intellectual mine? Neither is this a simple monetary argument, for without due attention, without the systematic shaping of datasets for subsequent reuse or re-purposing, such careless disdain for source data is likely to spawn a host of missed opportunities in economic, social and scientific advancement. The value of research data is not to be measured simply by the accountant's abacus.

That the research output from our universities has a direct impact on our lives and our state of physical and mental well-being is no mere accident but the consequence of strategic decisions taken at a national level. Take, for example, the UK's Engineering and Physical Sciences Research Council (EPSRC), which funds research across a broad range of disciplines including information technology, structural engineering and materials science and which seeks to align research with outcomes having relevance to society and business. In its 2008–2011 plan the EPSRC has identified £1.9 billion of research themes that will sustain advances in energy, the digital economy and next generation healthcare (EPSRC, 2009). It is public investment on this scale, coupled with a determination to produce results of benefit to the common good, which provides a meaningful indicator of the potential and critical value inherent in the data generated from research.

The expectation by the major funders that research data as a recognized asset will be afforded due care and attention has become more overt in recent years, confirmed by an emerging requirement for the inclusion of data management plans within research grant proposals. Their message is clear: data should no longer be abandoned on the workbench like wood shavings in a carpenter's shop; increasingly it is expected to join the finished assembly of scholarly output as a valued and managed component with an extended life and sustained usability.

To emphasize their collective solidarity behind that message, Research Councils UK (RCUK) have published seven *Common Principles on Data Policy* (RCUK, 2011), with the intention of providing an overarching framework for individual research council data policies. While recognizing 'that there are legal, ethical and commercial constraints on release of research data', the Principles also state emphatically that 'publicly funded research data are a public good, produced in the public interest, which should be made openly available with as few restrictions as possible in a timely and responsible manner that does not harm intellectual property'. The Principles take care to support caution against the inappropriate or premature release of data, as well as to urge the proper acknowledgement of sources. They also point out that 'it is appropriate to use public funds to support the management and sharing of publicly-funded research data', which neatly makes the connection with the councils' role as funders of research.

In the UK today, most domain-based research funders expect grant applicants to submit a statement describing their plans for access, management and the long-term curation of research outputs, although the approach varies according to individual funder, as described in a critical analysis of individual funder requirements available from the Digital Curation Centre (DCC) website (www.dcc.ac.uk/webfm_send/339). Although the Arts and Humanities Research Council (AHRC), Economic and Social Research Council (ESRC) and Natural Environment Research Council (NERC) each focus on the long-term sustainability of digital resources, the biomedical funders – Biotechnology and Biological Sciences Research Council (BBSRC), Medical Research Council (MRC) and the Wellcome Trust – are more concerned with the data-sharing potential of research resources (DCC, 2011). Such heterogeneity of approach perhaps reveals differences in individual research cultures and goals; in practical terms it may also produce additional challenges for the authors of cross-disciplinary research proposals, who will need to satisfy the content and formatting requirements of very different data management plans as defined by different funding agencies. Measures to address this diversity of demands are discussed in some detail in Chapter 5.

When the funding bodies also provide infrastructure services in the shape of national data centres (whose features are described more fully in Chapter 8), to which data can be offered for deposit and through which researchers will enjoy the support of a structured curation management programme, data management plans serve as an effective instrument for the eventual delivery of data to those centres. But elsewhere there exist neither carrots nor sticks to ensure that, once funds are released, there will be any rigorous adherence to the agreed plan. This lack of any monitoring function is not peculiar to UK funders; in the USA the National Science Foundation (NSF) announced that from 18 January 2011 all research proposals submitted to the NSF must include a supplementary two page data management plan, which will describe 'how the proposal will conform to NSF policy on the dissemination and sharing of research results' (NSF, 2010). That may convince scrutineers in the NSF that a proposal is on-message but it is a long way from ensuring that the data produced will be properly prepared, managed and preserved for long-term access and reuse. It certainly carries no sanctions to ensure compliance with any front-loaded statement of conformance. So if there is no rigour being applied by the funding agencies and with the majority of disciplines lacking the services of a national data centre, who is providing due care and attention to our research data output – the academic research community itself?

The research lifecycle

The research lifecycle has been described variously as a linear or cyclical model but in practice it consists of multiple investigative sub-cycles including tools, methodologies and a series of reiterative steps that serve to reinforce surges of forward progress in the overall sequence of activity. This process may be described more simply as following

the pattern illustrated in Figure 1.1. It has six sequential phases, commencing usually with an idea or hypothesis and concluding with the delivery of a product, most often in the shape of a published report or other scholarly text. To an ever-growing extent the published report is accompanied by supplementary data, in the shape of files containing datasets produced during the research programme, or links to databases with supporting information or protocols, which themselves may provide the spark to animate further investigation.

In each and every phase of the lifecycle the researcher will gather and use data and/or generate new data. In the initial phase, typically, when structuring a hypothesis and planning the research programme, existing published data will be gathered and reviewed; some will be used to set the scene, other data will be selected as the raw material for new research. Later, the research process itself will necessitate pulling in further data from published sources or from collaborators in the field, either for speculative reanalysis within new contexts revealed by the research programme or to provide authoritative benchmarks for comparing or measuring the output from new research. This phase will also be the principal source of new data arising from investigation or experimentation. As depicted by Figure 1.1, all of the six phases will be datacentric to some degree in that they will each depend on the use or generation of data, not least at the point of reuse, where data produced, filtered and synthesized may present opportunities for new research. This last function is perhaps the most crucial to developing the body of knowledge and deserves further explanation.

For certain groups of researchers, such as systems biologists, the availability of others' research data is no less than fundamental to their own work process. Systems biologists, who will often have an interdisciplinary background, work in complex teams

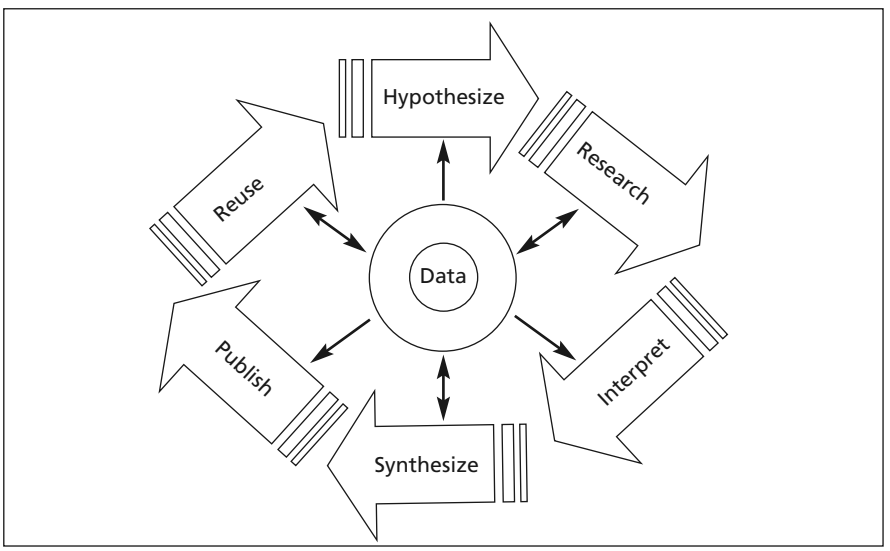

Figure 1.1 *The six datacentric phases of the research lifecycle*

that include an array of complementary experts, for example bioinformaticians, biophysicists and mathematicians. Their modus operandi is to produce new knowledge by modelling existing knowledge taken from large datasets generated within the global experimental and theoretical research community. At the risk of stretching a point, it is worth also referring here to a more popularly cited instance of effective data reuse, where new research has been enabled by the digitization of weather records extracted from a previous century of ships' logs, with data not originally gathered for that purpose now being used in research into climate change.

Despite its obvious 'datacentricity' the diagram at Figure 1.1 presents the researcher as a user or producer of data rather than as a data manager or custodian. And why should we expect anything more? As observed in the 2009 Research Information Network (RIN) study of researchers in the life sciences, 'data curation is only one element in the research lifecycle [and there is] little evidence that planned data management has yet been adopted as standard practice' (RIN, 2009, 49). Data can be difficult and costly to acquire or produce; it may take years to gather, often depending on the careful establishment of intricate relationships with collaborators or target study groups. Furthermore, in highly innovative fields of research it is likely that new techniques of data production and manipulation may first need to be developed. These are the issues that will absorb researchers and in which they will exercise skill: the getting of data in order to conduct research, often within a limited timescale, not the management or curation of data over the long term.

For researchers, career rewards are secured from the quality of their research output rather than from any efficacy in data curation, and while there may be a guarded acknowledgement of the value that may be achieved from sharing data, the effort necessary to preparing data for sharing is generally regarded as an unwelcome burden, requiring skills that are not necessarily present in a researcher's basic toolkit for conducting research.

A data curation lifecycle

But the needs of a full and effective data management programme are not so different from those depicted in the research lifecycle. The DCC's curation lifecycle model (Figure 1.2; DCC, 2009), which like the research lifecycle has its critical starting point at the research conceptualization stage, is designed to ensure that all necessary phases of curation are planned and undertaken in the correct sequence.

Data, meaning any digital information recorded in a binary as opposed to decimal form, where binary is the numerical format that computing devices employ to store and manage information, is shown at the centre of the curation lifecycle. It includes both digital objects (e.g. text files, image files or sound files) and databases (structured collections of records or data stored in a computer system). Notes provided to support the model describe the full lifecycle actions necessary for the preservation and curation of data, such as the assignment of metadata. These actions, together with an

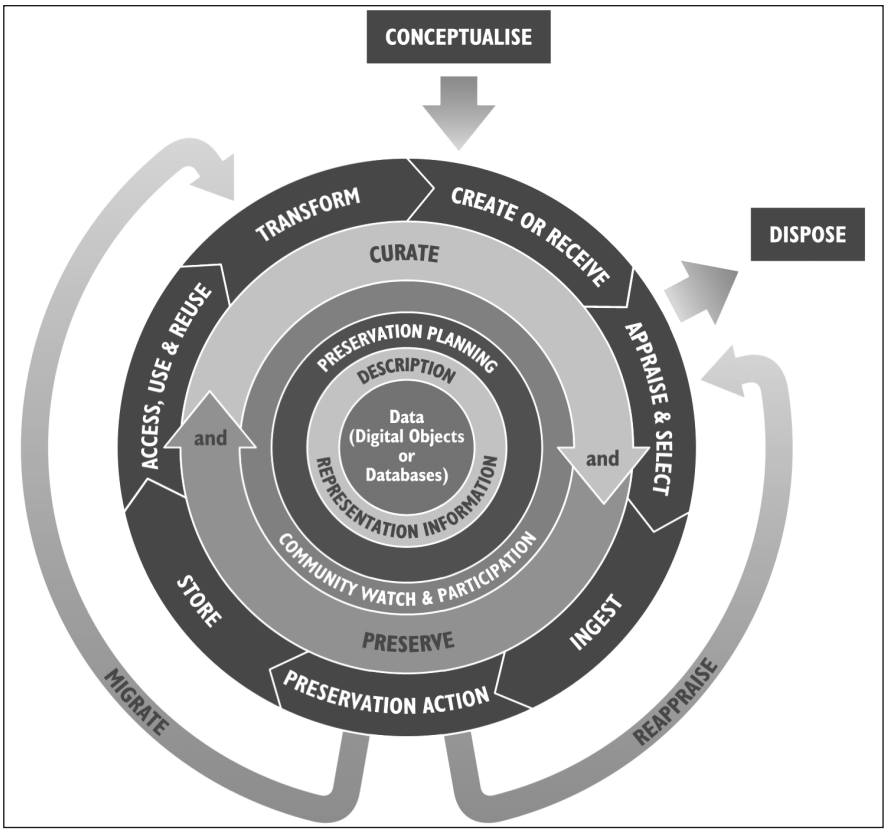

Figure 1.2 *The DCC curation lifecycle model [reproduced by permission of the DCC]*

explanation of other sequential and occasional actions, are described in more detail in Chapter 2. In the outermost ring, in an implicit reflection of the research lifecycle, the series of sequential actions starts with planning for the creation, capture and storage of data, eventually concluding with some predicted transformation of the curated data – although transformation is itself but a new beginning! Additionally, three occasional actions are also shown as pivotal to the concept of data curation, since disposal, reappraisal and migration are key to decisions informing a process that will enable a view over the longer term.

A routine of decisions about data disposal is needed to take account of not only changes in the potential long-term value of datasets but also any legislation governing the length of time that certain types of data must be preserved. The nature of some data, where for instance confidentiality is an issue, may even dictate the use of secure destruction methods. In all cases, the cost of curating data over the long term will require serious consideration and periodic review presents one means of achieving cost containment. Reappraisal is also necessary where data has failed to meet formal

validation procedures, since there is little point in retaining data that is neither reliable nor robust. Finally, migration of data may be undertaken following reappraisal or decisions about disposal and usually involves transformation to a different format, an undertaking that is essential if data is to continue to function within a changed storage environment or where it is necessary to ensure the data's immunity from hardware or software obsolescence.

The intention of the lifecycle model is unambiguous: to explain how maintaining, preserving and, most crucially, adding value to or extracting value from research data should be achieved throughout an optimized lifecycle. It does this by prompting us to ask what are the essential ingredients of an effective digital curation architecture. Notice too how the language used in the model is all about data handling and the needs of the data itself. It is a step beyond the view provided in Figure 1.1, where data is shown as an anonymous factor of production, feeding and enabling the research process. Here in the DCC model, the emphasis is on the changing aspects of the data rather than the research. In the DCC model data is to be captured and matured according to a plan, with a structure that is independent of the individual idiosyncrasies of research programmes. In this model the tone is redolent of care, where data is nurtured, massaged and preserved according to a dynamic and continuous process. There is one inherent omission: the sequences and activities are explained in careful detail yet who should be responsible for enabling or pursuing these preservation actions is less clear.

The sustaining professional: a longer view

Digital curation involves the active management of research datasets in order to preserve their long-term research value, yet this is a concept with limited appeal to the majority of a research community that receives short-term funding and is composed of a highly mobile workforce. Typically, within or across disciplines, members of that workforce will over time combine, disperse and recombine with seeming fluidity; the research they undertake will rarely follow an exclusive and linear path and as a community they will exhibit changing patterns of allegiance and interests. The dedication and opportunity to plan and work with data over the longer term must therefore belong with a different kind of community, one that is organizationally stable, sustainable and with the freedom and capacity to make plans and projections that will exceed the kind of short-term goals and funding allocations common to research themes and projects. The terminology used in the DCC curation lifecycle model also suggests a different kind of skill set to that traditionally associated with researchers, one that instead implies the stewardship and husbandry of data rather than its active use. It was not necessarily the intention of its author, the DCC's Sarah Higgins, who gives further insight into the curation lifecycle in the next chapter, yet the greatest resonance of this model is with the information practitioner, the archivist perhaps, or the librarian.

Those researchers who have recognized the need better to manage their data are faced by a dilemma. While they may eschew responsibility for acquiring and applying

skills in data management (or curation) beyond the basics necessary to enable their research, in surveys sponsored by the Joint Information Systems Committee (JISC) and RIN, researchers have consistently remarked that in order properly to discharge their function information professionals employed to provide support to any research group will require a substantial level of discipline knowledge. If interpreted unsympathetically, the inference here is that a level of discipline expertise would be expected that would put data curators almost on a par with the researchers themselves, thereby ruling out most professionally trained information practitioners from providing data management support to a university research team. It should not be the case, although such pejorative attitudes are heavily reinforced by a culture of self-sufficiency among the research community, in which the tendency to rely on oneself or one's trusted colleagues rather than central services is endemic.

Cultural barriers

Project StORe, an initiative from the JISC repositories programme, confirmed this culture as a significant barrier to change. One of the earliest of recent studies of research information behaviours, the 2006 StORe survey of seven scientific disciplines found researchers claiming undisputed rights to the management of their data with uncompromising declarations such as 'it's my responsibility' and, more dismissively, 'the university has assigned a librarian to our department . . . but I have not used her services' (Pryor, 2006). Such ingrained attitudes do pose a serious challenge to enthusiastic information professionals with a mission to engage with the research data agenda and it would appear that, before they can make headway, information professionals have a two-pronged challenge to overcome. The first involves regaining a greater parity of esteem with the research community, without which they will lack credibility; the second requires them to persuade and demonstrate that they have a material contribution to make, one that is likely to be of tangible benefit to researchers and the research programme. The former should follow the latter, but information professionals must be active in taking the lead; they can't wait for the researchers to come knocking. True, researchers may admit they are concerned about such issues as accessibility and barriers to acquiring information, but such is the pace of research that they will not set aside the time to seek assistance from beyond their own research group or to indulge in much more than a quick session with Google. It is, therefore, up to the information professional to learn about these concerns and to use them as a pretext for offering genuine advice and assistance, which has to be the first step in proactively reconnecting with the research community.

The re-purposed librarian

In recent years the traditional role of professional information intermediaries has been largely replaced by services that give direct access to ubiquitous online resources. Not

by design but from reliance on the availability of search engines, many researchers have effectively removed themselves from the mainstream library user population. Of course there is potential risk to them from this behaviour, given the limitations of generic search engines such as Google, and if steps could be taken to reconnect researchers with information professionals the resulting benefits to information discovery would, in turn, enhance the research process. Nonetheless, as a consequence of this change in dependency, libraries and librarians have become associated primarily with serving the needs of the undergraduate population, whereas for centuries they occupied a more august role as the recognized exponents of skill in classifying and organizing information and knowledge, including its appraisal, selection and annotation. In that role they have long been the natural source of expertise in storing and preserving information and, until the recent growth in online tools for the discovery and download of information, they had been unrivalled in their ability to retrieve information, to distribute and to share it, as well as manage access to it. This veritable catalogue of qualifications will become a repeated motif throughout this chapter, since it remains highly appropriate to the management of research data and obligingly describes a toolkit waiting to be opened and deployed!

The situation in the USA is somewhat more encouraging than in the UK. In spring 2010, for example, the library at the University of Virginia opted to pursue a new strategic direction and focus more on providing structured support to data management, with the primary aims of building data literacy among library staff, developing knowledge of how researchers at the university actually manage their data and creating opportunities for active consultation and collaboration. In the face of universally shrinking budgets, this strategy required the identification of services most likely to produce the greatest value to the institution, as well as hard and radical decisions over which existing services to drop or change in favour of the new regime. What emerged was a new Scientific Data Consulting Group (University of Virginia, 2011) consisting mainly of existing library staff who had been 're-purposed'. One must assume that Virginia's research community, like that also at the University of Minnesota, is proving receptive to this initiative.

At Minnesota a programme of assistance in the creation of data management plans was launched proactively in advance of the NSF declaration, building on the results from studies of researcher needs conducted at the university in 2006 and 2007 (University of Minnesota, 2007). The response to those studies was generally positive, attracting such statements from faculty staff as 'If there were a workshop on organization and file management, I would go. The Libraries do this so well.' The University of Virginia initiative too appears to be flourishing. Guidance to help Virginia's researchers comply with the National Science Foundation's requirements for data management plans has been produced and the university has joined a group of major research institutions working to develop a flexible online tool that will help researchers generate data management plans. This group includes the UK's DCC, which developed the first such online tool (downloadable from www.dcc.ac.uk/dmponline).

Risky business

In the UK the needs of the research community are no less pressing, where of particular concern is how bare and basic are some of these unsatisfied needs. Witness the practices recorded by a scoping study undertaken by the project Incremental, a collaboration between Cambridge University Library and the Humanities Advanced Technology and Information Institute (HATII) at the University of Glasgow. From a series of in-depth interviews Incremental documented the extent to which some research datasets exist in a state of relentless jeopardy, perhaps the most serious revelation being that researchers at both institutions were having difficulty in finding even their own data. This was principally because of their use of inconsistent file structures and naming conventions, the extensive and risky practice of storing critical research data on cheap and flimsy media such as flash drives, and the scant deployment of networked storage facilities in some areas (Incremental, 2010). An already difficult situation was found to have been compounded by the creation of only minimal documentation to describe the nature and condition of the data that had been stored and, most surprisingly, a severely limited awareness of the opportunities and routines for data back-up. Incremental's response has been immediate and pragmatic: to meet researchers' demands for simple, clear, engaging and available help and support by producing accessible visual guidance on the creation, storage and management of data, supported by discipline-specific training in data curation principles and techniques. But no explanation for the prevalent dearth of basic data management practice was given in the study report other than an acknowledgement that the technological and human infrastructures currently provided by institutions are often insufficient to meet researchers' data management needs, as a consequence of which they are forced to do the best they can with the limited time, skills and resources available.

National centres, services and strategies

As an exemplar for building bridges between the information and research communities Incremental has proved to be a success. But it is nevertheless a pilot project with limited scope and a very specific locus. In the UK generally there remains an uphill struggle to identify resources sufficient to bridge an apparent gulf between the actual capabilities of information professionals and their perceived inadequacies in the mind's eye of researchers. Under such circumstances is there perhaps an alternative body better positioned for the challenge? Incremental is one of eight projects funded under the JISC's research data management infrastructure programme (JISC, 2010), with whom its findings resonate. This is a generously resourced programme with a strategic ambition to provide the UK higher education sector with examples of good practice in research data management. But beyond this developmental community there are several national organizations already well established in the data curation field, not least the data centres previously mentioned. One, the UK Data Archive (UKDA), has

been operational for over 40 years, curating the largest collection of digital data in the social sciences and humanities in the UK.

As respected centres of expertise, these data centres provide not only guidance on data management practice but also the costly infrastructure necessary for the storage, preservation and access management of data deposited with them. They also reflect and influence the development of data management policy by the research councils that, typically, are their principal source of funding. In the case of the UKDA, for example, involvement in drafting the ESRC's Research Data Policy (ESRC, 2010) allowed UKDA staff to draw on their skills and enduring practical experience as well as to consult with other expert bodies such as the DCC, leading to the publication of a well informed and practical document that identifies the responsibilities of research grant holders, the ESRC and the data service providers. This was an exercise in the construction of policy as a tool for support and assistance rather than the composition of a political decree, an object rarely afforded much regard within academia! Similarly, the Natural Environment Research Centre's (NERC) network of data centres supports an integrated Data Discovery Service, covering the several strands of environmental research funded by the NERC, and providing an authoritative interface between the broader body of data users and the NERC research community.

Other centres hosted in the UK include the Archaeology Data Service and the European Bioinformatics Institute, with further domain-specific services under development, such as those being designed to support projects funded by the Medical Research Council. But notwithstanding the value and success of these organizations as assured custodians of the knowledge produced in their individual subject domains, they do not represent or serve all the fields of active research; nor are their services necessarily inclusive across their own domains, where they may adopt a selective approach to the preferred coverage and range of data that they will accept. Neither should we be complacent in regarding their custodianship as assured and sustainable: witness the demise of the much-appreciated AHDS in 2008, which has been reported as a direct consequence of unsympathetic financial pruning by the AHRC.

A case for the proper management of research data can be advanced on financial or ethical grounds but agreeing the roles and responsibilities for managing the research data deluge, as well as enabling a coherence and consistency of approach, remains a complex question requiring active participation and commitment from a range of stakeholders. In the UK the initiative to inject coherence to the research data community has been taken by the JISC, through the DCC, which it funds and which was launched in 2004 as a key component of the JISC's Continuing Access and Digital Preservation Strategy (www.dcc.ac.uk/about-us/history-dcc). In late 2010 the DCC's role was extended to accommodate aspects of a further initiative to create a UK Research Data Service, with funding from the Higher Education Funding Council for England (HEFCE), whereby it will from 2011 provide data management services in support of a new national cloud computing and storage infrastructure for research.

At the other end of the digitally connected world the Australian National Data

Service (ANDS) has embarked on a ten year programme to transform collections of Australian research data into a cohesive network of research repositories, at the same time taking steps to equip Australian research data managers with the skills to become expert in creating, managing and sharing research data under well formed and well maintained data management policies. The concept of research data manager is here an inclusive notion, in which the ANDS programme seeks to address the broad issues of research data ownership and the roles and responsibilities associated with ownership and maintenance. An ambitious platform engineered for the nationwide promotion of best practice in the curation of experimental, research and published data.

ANDS is a top-down government-sponsored programme, initially proposed in 2007 by the Department of Education, Science and Training (Australian Government, 2007) and introduced in 2008 by the Federal Department of Industry, Innovation, Science and Research (DIISR), which entered into an agreement with Monash University to establish ANDS under the National Collaborative Research Infrastructure Strategy (NCRIS). Funding of A$48 million (£30.4/$47 million) over two years was agreed in 2009 'to create and develop an Australian Research Data Commons (ARDC) infrastructure' (ANDS, 2011). A more detailed discussion of national strategies for research data management in Australia and in the USA is provided in Chapter 9 of this book.

But however forward-looking and well intentioned, can such national strategies expect to be successful in coaxing or cajoling the traditionally independent researcher to participate in and support them? A good deal of positive advocacy will have to be rolled out before that bond can be secured, coupled with a sound demonstration of the benefits to all potential stakeholders, particularly those for whom signing up to the concept of systematically managed data may represent a new and burdensome workload. For some this new interest in their research data may even be perceived as a threat to traditional rights and practices, for notwithstanding the intellectual property rights asserted by employing institutions, the data produced and assembled by university researchers is regarded as their intellectual capital, the basis of their credentials as effective researchers and the stuff on which career progression is built. Whether real or imagined, any fears that their perceived ownership of that data is in peril will have to be assuaged before progress can be made.

After Doomsday

Such a conundrum returns us to the role of the modern data practitioner or information professional; still to be born perhaps from the ashes of an outmoded perception of the librarian or information scientist but the most likely candidate for the role of standard-bearer when national or institutional strategies are to be rolled out. There is no question that the Doomsday Scenario for librarians painted in 1979 has proved to be wrong. The profession has continued to adapt and change as it always had, finding and developing new roles as the digital age advanced, and we began to

hear about media librarians and systems librarians, witnessed the introduction of e-libraries and more recently watched the implementation of information repositories. There has been a succession of changed and changing roles.

Yet nearly all of this flux and change has been seen in the context of published information, not data. True, there is a handful of data librarians employed in our more iconoclastic institutions – it was estimated not so long ago that there were five in the UK, principally individuals 'originating from the library community, trained and specializing in the curation, preservation and archiving of data' (Key Perspectives, 2008) – but the library world has yet to commit wholeheartedly to the transition. The library schools in our universities may provide a sound education in what is broadly described as knowledge management or information management, but training in the intricacies of web search engines, information systems and database design does not properly equip the new professionals with an outlook that will fit them for a role as data manager in a research intensive university. This is despite the profession having unrivalled occupation of the high ground when it comes to owning a long list of fundamentally appropriate skills in classifying, organizing, appraising, selecting, annotating, preserving, storing, retrieving, distributing, sharing and managing access to information – some list indeed, worth repeating here, and one that closely reflects the activities implicit to the DCC's data curation lifecycle model!

The following chapters in this book address in greater detail many of the issues raised in this introduction, and more, adding practical advice on such topical themes as data management planning and the sustainability of digital curation, with analyses of national policies and strategies in the Old and the New World. All start from the premise that we have answered the question 'Why manage research data?' We hope readers of this book will as quickly be convinced and that, on good argument, it will inspire them also to become committed advocates of the research data management cause.

References

ANDS (2011) *Overview of Funding Processes*, Australian National Data Service, http://ands.org.au/funded/funding-overview.html.

Australian Government (2007) *Towards the Australian Data Commons*, Department of Education, Science and Training, www.pfc.org.au/pub/Main/Data/TowardstheAustralianDataCommons.pdf.

Britt, R. (2010) *NSF Science Resources Statistics*, National Science Foundation, www.nsf.gov/statistics/infbrief/nsf10329/.

DCC (2009) *Curation Lifecycle Model*, Digital Curation Centre, www.dcc.ac.uk/resources/curation-lifecycle-model.

DCC (2011) *Overview of Funders' Data Policies*, Digital Curation Centre, www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies.

EPSRC (2009) *EPSRC Landscapes*, Engineering and Physical Sciences Research Council,

www.epsrc.ac.uk/research/landscapes/Documents/LandscapeIntro.pdf.

ESRC (2010) *Research Data Policy*, Economic and Social Research Council, www.esrc.ac.uk/about-esrc/information/data-policy.aspx.

Incremental (2010) *Scoping Study Report and Implementation Plan*, www.lib.cam.ac.uk/preservation/incremental/docs.html.

JISC (2010) *Research Data Management Infrastructure Projects (RDMI)*, Joint Information Systems Committee, www.jisc.ac.uk/whatwedo/programmes/mrd/rdmi.aspx.

Key Perspectives (2008) *Skills, Role and Career Structure of Data Scientists and Curators: assessment of current practice and future needs*, Joint Information Systems Committee, www.jisc.ac.uk/publications/publications/dataskillscareersfinalreport.aspx.

MacArthur, D. (2008) *How Much Data is a Human Genome?*, www.genetic-future.com/2008/06/how-much-data-is-human-genome-it.html.

NSF (2010) *Dissemination and Sharing of Research Results*, National Science Foundation, www.nsf.gov/bfa/dias/policy/dmp.jsp.

Pryor, G. (2006) *Project StORe Survey Report Part 1: cross-discipline report*, http://hdl.handle.net/1842/1419.

RCUK (2011) *Common Principles on Data Policy*, Research Councils UK, www.rcuk.ac.uk/research/Pages/DataPolicy.aspx.

RIN (2009) *Patterns of Information Use and Exchange*, Research Information Network.

University of Minnesota (2007) *Understanding Researcher Behaviors, Information Resources and Service Needs of Scientists at the University of Minnesota*, www1.lib.umn.edu/about/scieval/documents.html.

University of Virginia Library (2011) *Scientific Data Consulting*, www2.lib.virginia.edu/brown/data.