5. Institutions in the data ecosystem

Actors in the public knowledge domain and in private data companies

Merel Noorman

Introduction

The role institutions play in mobilising open data and their position in that mobilisation are key aspects in the development of open data. As the discussion in Chapter Four shows, institutions act in response to the drivers that are pushing for open data and they are key actors in overseeing and implementing a range of changes to facilitate the development of open data. Institutions are also situated within other broader changes that affect how they operate, which include the significant changes in systems for generating, sharing, and disputing human knowledge using digital technologies (Edwards et al. 2011). As outlined in Chapter Two, the way knowledge is generated has changed over historical time and the shift to Mode 2 Knowledge Production means that it is now becoming an increasingly large-scale, global, interdisciplinary, collaborative effort organised around open data sharing. Part of that change is the way digital technologies, including formal university and research centre systems as well as open public systems and platforms such as social media, crowd-based knowledge phenomenon such as Wikipedia as well as big data, have enabled new ways of producing and circulating ‘knowledge’. In general terms, this includes shifting to less-centralised forms of knowledge production and circulation. This has had an effect on the ecosystem and the way that institutions interrelate.

However, the move towards open data, in particular, is changing wider organisational processes of data production, dissemination and use. This change is key in defining and shaping any broader social change towards a knowledge society. The possibility of unrestricted availability of data makes possible new kinds of education, services, business models and scientific and scholarly communication practices. Institutions in the scientific and scholarly domain, such as libraries, data centres, national archives and

---

universities, are expected to play a leading role in enabling these changes, by developing and maintaining technical and organisational infrastructures for the production, preservation and sharing of open data. At the same time, these institutions are confronted with a rapidly-changing environment, in which relations between institutions within the data ecosystem are being continuously redefined. For many institutions, this is a relatively new and unchartered area, and most institutions have only taken initial steps in exploring and giving shape to their new roles and responsibilities.

This chapter moves beyond broad discussions about societal changes to focus on institutions that are involved in scientific and scholarly practices. It explores some of the changes that these institutions are undergoing and the resulting challenges they face. The next section examines the changes in the landscape, which is followed by a discussion of three institutional challenges in particular: securing funding for open access, controlling the quality of data, and providing training and education. The final section concludes by discussing how many institutions in this domain are renegotiating and reinventing their roles, responsibilities and relationships to position themselves as coordinating actors within the evolving data ecosystems. However, in order to successfully make data open, they need to engage meaningfully with research communities.

Institutions and their changing role in data ecosystems

As the previous chapters have shown, the move to making research data open is partly the result of a drive from a range of open data advocates and some level of top-down push for open access from policymakers. Institutions are therefore facing increasing demands from civil society organisations, policymakers and from scientific communities for access to data as well as data services and support. This section examines the ways that various actors envision how institutions and organisations may adjust to open data. In particular, it uses documentary evidence, including policy reports, strategy documents, and other relevant literature, from the institutional issues grand challenge research within the RECODE project to illuminate current negotiations about the changing role of institutions within the open access to research data ecosystem.

All the different activities required to make and keep data openly accessible take place within a network of institutions. The discussion in Chapter Three points to the different requirements that need to be put in place to make data open and that this requires considerable work to make
data easy to access, use, and evaluate. Data must be digitally generated or converted into standardised and machine-readable formats and metadata must be added; the data and metadata have to be reviewed and checked for inconsistencies, noise or errors and, if possible, linked to other data sets. To store data and make them accessible, infrastructures must be funded, built, and maintained. As discussed in more detail in Chapters Six and Seven, tools must be developed to make the data searchable and reusable. To preserve the quality of data, multiple versions of data sets have to be managed and occasionally migrated to other, new technological platforms. Furthermore, the general requirements of making data open as outlined in Chapter Three and as the issues of doing so at the research practice level as discussed in Chapters Six, Seven and Eight, including copyright issues and informed consent, must also be clarified and managed. Additionally, making research data available for unrestricted use means that researchers and data managers need to be trained and persuaded to publish their data. Strategies have to be developed to evaluate the quality of data sets, models or code, and to measure their impact. Once the data is online, different levels of access have to be managed and the data's security has to be maintained.

In the networks in which all these activities take place, institutions perform varying and multiple roles and functions, often in collaboration with other institutions, and this open data ecosystem is still evolving. For many institutions, open access to research data is a new development that they are just coming to grips with, and they are not well integrated into the open data movement. Moreover, it is a new ambition or requirement to add to their many existing priorities. As various reports, roadmaps and policy documents have pointed out, enabling open access to research data and successfully exploiting the various approaches, thus requires changes in research cultures, infrastructures and funding models (see, for example, The Royal Society 2012 and Higher Level Expert Group 2010). For many institutions, the expectations expressed in such policy guidelines and roadmaps mean that they are obliged to take on new responsibilities as well as exploring and developing new practices.

Policy reports such as the influential Royal Society’s report on open data (2010) generally expect universities, research institutes, libraries and funding bodies to take a leading role in enabling these changes in culture and infrastructure creation. For instance, The Royal Society recommends that:

Universities and research institutes should play a major role in supporting an open data culture [...] Learned societies, academies, and professional bodies should promote the priorities of open science amongst their members and seek to secure financially sustainable open access to
journal articles. They should explore how enhanced data management could benefit their constituency and how habits might need to change to achieve this (p. 10).

The Royal Society sees a role for universities and research institutes in: adopting open data as a default position; evaluating researchers on their data sharing; developing strategies and policies for taking care of their own knowledge resources; and offering services to support researchers with their data management. They are also expected to provide funds to encourage culture change and to ensure sustainability of any resources and infrastructure that are developed.

Libraries have also become increasingly important in research data management. From a review of the literature, Cox and Pinfield (2013) find that libraries may provide services regarding research data management such as: offering advice on funding sources; promoting open access; data analysis advice; guidance on research data citation and copyright issues; and technical advice on data formats and metadata. For many libraries, this is uncharted territory, so providing these services requires the adoption of new processes and skills (The Royal Society 2011, p. 63). Furthermore, libraries are often seen as the executors of data strategies and policies, as many of the responsibilities assigned to a research institution or university are realised through their library or data repository.

In practice, it is often unclear which stakeholder is responsible for what. Many issues cross stakeholder groups, due to the complexity of the data journey (from collection through to making the data open access for reuse). For example, training and skills development are seen to be the responsibility of various stakeholders: governments should adapt new policies for data management skills to be taught at university and secondary school level; funders should educate their grant-holders about data management and institutions, with the help of libraries; and IT departments should provide training for their researchers and other staff on data management (Noorman et al. 2014). Researchers should also serve as mentors to early investigators and students who are interested in pursuing data sciences (National Science Board 2005). Such diffuse and, sometimes, conflicting roles and responsibilities can complicate the process of making research data freely accessible. Moreover, each research project may allocate responsibility in different ways.

Nevertheless, funders and policymakers are placing an increasing emphasis on the coordinating role of institutions, particularly regarding what they perceive as a revolutionary shift in science towards more large-scale
collaborative and global research projects, which are facilitated by extensive data infrastructures (Noorman et al. forthcoming).

Challenges

Research for the RECODE project indicated that some institutions have already made considerable progress in sharing data and providing open access to data. Several large international consortia have, for example, created institutions to manage huge amounts of research data, such as the CERN data centre and the European Bioinformatics Institute. Other institutions have embraced the idea of open access and begun to take steps towards transforming their data-sharing practices. Yet, developments are still at an early stage and making research data publicly available has proven to be a considerable challenge in most disciplines, with approaches that work in one field not necessarily working in another (Tsoukala et al. 2015).

The values and norms of research practices within different disciplines shape the ways that researchers produce their data and make them openly accessible (Leonelli 2014). As discussed in more detail in Chapter Six, research practices encompass discipline-specific knowledge bases, technical tools, research teams, laboratories and culture as well as subject ethics and moral positions. These practices vary widely across disciplines and sub-disciplines, as science is made up of multiple epistemic communities producing very different products (Knorr Cetina 1999), and each community has different ideas about the significance of open data. For instance, whereas archaeology has relatively limited experience of sharing data and open data, bioengineering has a longer tradition of openly sharing models and methods, but not raw data. Established ethical and moral frameworks constrain the volume of data that is openly shared in health and clinical studies. The need to share expensive and large-scale equipment, such as telescopes and particle accelerators, has encouraged the particle physics community to establish effective data-sharing practices within large-scale global collaborations. Open data are, thus, embedded and enacted differently in different research practices, which are discussed in more detail in Chapter Six. The awareness, knowledge and infrastructure required to make data openly available is not, however, equally distributed across the scientific landscape.

This section focuses on the challenges involved in navigating between the competing interests of heterogeneous stakeholders, entrenched institutional cultures and wide-ranging and sometimes contradictory ideas about
open data. Institutions often have different audiences, multiple funders and diverse ambitions. This can lead to conflicting requirements and expectations, for instance, about the quality or preservation of data, or the allocation of responsibility. It can also result in additional demands being made on an institution’s often already-stretched capacity. The organisational culture within an institution can also be a barrier to making data open. Institutions generally have multiple functions and ambitions that have shaped the roles, skills and practices within them, and these do not always coincide with what is needed to make data openly accessible. Finally, it is difficult – if not impossible – to define what constitutes data, because data can take different forms in different disciplines.

The following section explores challenges in three different areas – financial support, data curation and quality control, and training and education – and discusses how various institutions have taken steps to address them.

**Financial support**

One of the key challenges institutions face is securing funds for open research data and, as noted in Chapter Three, the level of financial and other resources impacts on the development of open data. Efficiency gains through the reuse of data and the avoidance of duplication of data collecting and producing efforts are an important driver for open access to research data (OECD 2007; High Level Expert Group 2010). However, despite the potential cost savings that open access to research may bring in the future, preparing, archiving and making data freely accessible can be expensive, depending on the characteristics of the data to be stored, searched and used. Costs can be particularly high in data-intensive research fields, where there are extremely large data sets and high-tech computing equipment is required to process and interpret the data. In the RECODE particle physics case study, for instance, making and keeping data available to a wider public is an expensive process, because of the sheer amount of effort and resources needed to produce and store the data. Moreover, users need extensive computational resources and specialised knowledge in order to access and interpret the data. Yet, the costs of making data freely available and easily accessible may also press heavily on the available research budgets for smaller individual projects (Sveinsdottir 2013).

To make data openly accessible, funds must be secured for various phases in the data lifecycle, including the preparation, processing, sharing and archiving of data (Parse.Insight 2010). Researchers need to spend time formatting data, adding metadata and making them accessible. Archives,
data centres and repositories incur significant expenses for this acquisition, processing and access, such as personnel wages, training costs for researchers and (data) librarians and outreach programmes (Beagrie et al. 2010). They also have to invest in the development and use of the technical infrastructure required, including the hardware needed to store the data and the software tools to use them. Moreover, increasing volumes of open research data may change existing practices and introduce new ones that will require funding and resourcing. For instance, monitoring access to data or maintaining the integrity of data may generate new costs. It may also require establishing an ethics board or developing and implementing additional administrative and editorial procedures. These requirements show that resources are needed by institutions to translate the vision of open data advocates, as described in Chapter Three, into a working institutionally supported ecosystem.

As the open access movement grows, governments, funding agencies, universities and libraries have allocated increasing funding and have developed policies to stimulate open data resources and sharing (Mosink et al. 2013; DCC 2012). Funding agencies have used two main funding strategies. First, through funding researchers and their projects, they have contributed to the development of open data infrastructures. Subject data repositories developed as part of research projects tend to be financed in this way. However, there are limited and uneven numbers of subject-specific data centres and repositories (Cox and Pinfield 2013) and institutions are often expected to maintain outputs in the long term for research that falls outside the remit of these centres and repositories. Second, funding bodies have attempted to stimulate open access by directly investing in developing and maintaining data centres and repositories, which offer data services to researchers and research groups, often at no cost. Science and medical funders frequently contribute to this kind of joint initiative, for example, at the European Bioinformatics Institute.

It is not immediately clear which institution should bear the responsibility for funding the many tasks involved in open data. Institutions tend to consider researchers responsible for obtaining financing for the publication and curation of their data. As data producers, scientists are viewed as the starting point of the data journey, so they are deemed responsible for ensuring data quality, ethical data collection and clear communication of data, e.g. the writing of metadata and context. Nevertheless, the top-down push for open access has also created responsibilities in terms of funding the activities required. Research communities look towards national and transnational funding bodies as well as research institutes to provide
resources that will enable them to implement the various mandates and policies. Funding bodies in particular are expected to make funds available to support open data. They are seen to be responsible for providing – or at least investing in – infrastructure, in the form of data repositories, which will store data from research they have funded (OKF 2013).

The heterogeneity of data sets and associated requirements in the wide variety of disciplines adds to the challenge of financing the various phases of open research data. Institutions are now obliged to find ways to provide various services for different kinds of research, often on a limited budget. For instance, data from psychology experiments require different kinds of data management than field notes from archaeological excavations. Universities, national data centres and libraries have all stepped in to provide infrastructure and financial resources to preserve data sets that fall outside the remits of existing repositories and data centres but, as the number of such data sets grows, this may not be a sustainable option for all of them. The curation of open data continues to require resources, i.e. staff or volunteers with suitable skills and expertise are needed to keep the data up-to-date and accessible. They will have to make decisions about issues, such as what data to keep and how to transfer them to new technologies or formats. To stimulate their use, an effort has to be made to bring the data to the attention of relevant audiences. Universities and libraries will need ‘larger budgets and highly skilled staff if the roles that are suggested are to be fulfilled by institution, such as universities’ (The Royal Society 2012, p. 67). This raises the question of what should be funded. Some disciplines produce petabytes of data, which cannot all be stored, while some data sets might not appear interesting or useful enough to keep – although it is difficult to predict what will become valuable data in the future. As the volume and number of data sets grows, institutions will have to start making decisions about what data to keep and how to store it. They will have to develop strategies for choosing what to invest in, and these decisions may be affected by public demands for outcomes and results.

Several initiatives offer potential solutions to some of these challenges. Based on its analysis of the costs of data preservation, Jisc recommends that institutions should ‘take advantage of economies of scale, using multi-institutional collaboration and outsourcing as appropriate’ (Beagrie no date). In an effort to achieve this, some institutions have started to collaborate in offering data services. Several universities around the world have also established data repositories, sometimes as collaborative initiatives between multiple universities. One example of this is the collaboration between the Dutch Data Archiving and Networked Services institute
(DANS) and several archives and libraries in a federated data infrastructure, which is based on a front-office/back-office model (Dillo et al. 2013). This federated system comprises a network of local data stewards who are close to scientific practices, combined with centralised data services. DANS, jointly funded by the Royal Netherlands Academy for the Arts and Sciences (KNAW) and the Netherlands Organisation for Scientific Research (NWO), provides free storage and preservation for data sets in the humanities, social sciences and other disciplines (DANS, no date). DANS also runs the Dutch Dataverse Network, ‘an open source application to publish, share, reference, extract and analyse research data,’ which was first developed at Harvard University (The Dataverse Network project, no date). A number of databases or disciplinary repositories have arisen following collaboration between multiple universities, research institutes, funding agencies and governments, and some of these consortia have successfully obtained funding for data preservation for longer periods of time. For instance, the International Nucleotide Sequence Database Collaboration has developed and maintained three databases – DNA Data Bank of Japan, European Nucleotide Archive, and GenBank – for over 18 years. These databases receive funding from member institutions, project grants, funding bodies and governments.

Some institutions have begun exploring new ways of recovering the costs of open data. For example, several data centres have started charging for access to their larger data sets. The Dryad Digital Repository, which provides open access to research data underlying scientific publications, has developed a business and sustainability plan based on a combination of membership fees, data publishing charges (DPCs) and project grants (Dryad 2013). A diverse range of stakeholders, including journals, research institutions, publishers and scientific societies, can become a member and pay a fee, in exchange for a say in the governance of the organisation and discounts on submission fees. There have also been initiatives to tackle the challenge of long-term curation. University College London (UCL) has attempted to address part of this challenge by offering three different services: data storage services for the run-time of the project; data preservation services; and access services. By offering storage for the run-time of the experiment, UCL aims to encourage researchers to think about what will happen to the data after the project ends. Several funding bodies also now require applicants to specify how their data will be preserved longer-term as part of their applications for funding.

Although financial support for open data remains a challenge, this section has shown that more funding is becoming available from the traditional
channels of research funding – for both researchers and data infrastructure. At the same time, various institutions have also started experimenting with alternative ways to cover their open research data costs. Institutions are building new alliances and exploring new organisational structures that will address the challenge of funding open data.

**Data curation and quality control**

An important aspect of curating open data is ensuring and maintaining their quality, with a view to enabling their reuse. To ensure that open research data is of value to research communities, researchers need to have some level of confidence in the trustworthiness and integrity of data sets, and in data repositories. Open data sets and metadata that contain significant inconsistencies, inaccuracies, flaws or that are incomplete are unsatisfactory and hard to work with.

In many disciplines, some formal and informal mechanisms are already in place to assess the quality of data at various stages of the data lifecycle. Research communities may perform several review processes, manually and automatically validating data. Data may be checked as part of the automated processes that control scientific instruments, for instance, through completeness or consistency checks, file format validation, metadata checks, storage integrity verification and tools for annotating the quality information (APARSEN 2012). Such automated procedures can quickly process data and identify and problems in real time. Nevertheless, expert knowledge may still be needed to make appropriate decisions on how to treat data which is flagged as problematic (Campbell *et al.* 2013), because automatic quality checks usually focus on just the technical quality, e.g. the completeness of the metadata or the consistency of the data. Assessing its scientific quality means evaluating data and metadata content, by considering aspects such as whether appropriate methods were used to collect the data, or if the data accurately reflect actual observations or responses. Evaluating data on that level usually requires expert knowledge and can only be achieved through peer review or appraisal by a dedicated subject specialist. Scientific practices also tend to have built-in self-correcting mechanisms, such as replicating experiments or using publicly-available sources, which encourages researchers to produce high-quality data (KNAW 2013).

Open access to high-quality research data requires additional efforts and expertise to ensure, for instance, that data are interpretable, assessable and reusable (Swan *et al.* 2008; The Royal Society 2012). A complicating factor is that ideas about what comprises a sufficient level of quality will differ,
depending on who produces, manages or uses the data. Institutions providing access to public sector data for researchers, commercial organisations or citizens may have to deal with competing interests in setting their quality standards. The varying forms of data also raise questions about which, and when, data should be published – for example, should raw data be made available as early as possible, or should the data be processed first, losing some information, but making it easier for others to interpret and reuse?

One of the most important barriers that institutions face in evaluating and maintaining the quality of open data lies in the blurred distribution of responsibilities among stakeholders (Pearlman et al. 2013). Data quality issues entail the involvement of a variety of stakeholders in the data ecosystem, such as research funders, universities, data centres, repositories and researchers, at different stages of the data lifecycle. The role of researchers is central for data quality, as responsibility rests with them during the first stage of the data lifecycle, as part of their overall responsibility for undertaking research that is valid, accurate and ethical. However, engaging researchers in developing quality assurance practices poses a challenge in many disciplines. In particular, issues related to data management seem counter-productive to many researchers, who feel that it will require significant work to make their data accessible and reusable, but they will not be rewarded for doing this work (Kuipers and Van der Hoeven 2009). This is because, often, current institutional structures are not set up to reward researchers for their added work in terms of funding, promotion or knowledge gain. At the same time, data centres and libraries are the main actors being assigned responsibility for ensuring the quality of data, but they often lack the time or expertise to determine the level of quality. Data centres, institutional repositories and publishers serving diverse research groups have to make decisions about the extent to which they invest in ensuring the quality and integrity of data sets from various disciplines. Many institutions consulted within RECODE reported that it was too expensive to employ several data librarians or data scientists who are specialists in particular subjects and therefore capable of quality assessment. Moreover, as reviewing practices are community-specific and dependent on the form of data, it is difficult for repositories or publishers to formulate recommendations about data quality for every discipline and data type.

Various institutions have taken up the challenge of the quality of open data and have assumed a coordinating role in this evolving landscape. For example, the RECODE project found that some academic journals are contributing to quality assuring research data by developing standards,
methods and criteria for reviewing data effectively. They formulate requirements regarding data documentation and incorporate these in their editorial policies. Furthermore, increasing numbers of journals demand that the research data supporting articles they publish should be openly accessible. An example is the PLOS policy, which requires such research data to be openly available through an appropriate repository (Bloom 2013). This policy mandates that the research data should be recorded and deposited according to disciplinary standards, and it provides extensive references and links to discipline-specific bodies’ data documentation requirements to support this. Several publishers and journals also compel their reviewers to check the underlying data before they will publish submitted research (Penev et al. 2011). Another example where institutions are working to ensure the quality of open data is the increasing use of, and demand for data management plans (DMPs). DMPs have become a commonly-used means to encourage researchers and research groups to ensure the integrity and quality of their data. Funders, universities and data centres are increasingly encouraging or, even mandating researchers to develop a DMP at the beginning of their research projects. Such a plan should specify things such as how the researchers intend to ensure the quality of their research data (see, for example, University of Edinburgh 2014). Through such initiatives, institutions position themselves as having leading or coordinating roles in open data, setting standards for what is considered to be good-quality data.

The demand for high-quality open data has not only generated new practices and activities for institutions, it has also contributed to the development of new kinds of scientific communication. A relatively new type of publication is data journals, which publish articles that discuss data sets that are openly available in (certified) repositories, in terms of acquisition, methods, processing, etc. These articles describe the data acquisition process and discuss the considerations around experimental design (Gorgolewski et al. 2014), but they do not provide any analysis or results. Nevertheless, the articles undergo peer review, as do the underlying data. These types of publications draw attention to the significance of research data as independent publication objects as well as considering their quality and potential for reuse (Mayernik et al. 2014). They also help to establish good practices, such as referencing data and making them available through accredited repositories. Another example is the emergence of new mechanisms to enhance data quality, for instance through providing

---

2 Ubiquity Press is a well-known publisher of data journals in the humanities: http://www.ubiquitypress.com/.
platforms where researchers can discuss data sets, or tools that can be used for altmetrics. Altmetrics is the study and use of scholarly impact measures, based on activity in online tools and environments. According to the Altmetrics manifesto, ‘altmetrics expands our view of what impact looks like, but also of what’s making the impact’ (Roemer et al. 2013).

However, RECODE has also found that evaluating and maintaining the quality of data, both in terms of bits and bytes as well as scientific value, requires considerable work and significant changes in organisational and cultural practices. Peer review strategies have to be developed; data citation practices have to be actively encouraged and made part of institutional evaluation and reward systems. The initiatives described above illustrate that some institutions, often in collaboration with research communities, have developed a range of tools, practices and relationships to ensure the quality of open data. In order to accommodate the heterogeneous demands for data quality from a wide variety of stakeholders, there is a need for institutions to develop new roles and relationships. Publishers, data centres and libraries are therefore collaborating to encourage changes in existing cultures and to offer incentives to make research communities prioritise their data management and quality assurance.

Training and education

Various reports and roadmaps assign institutions a key role in encouraging researchers to develop open data practices. ‘Scientists need to take action to make their data available but it is up to supporting institutions to clear barriers and facilitate this process, by offering incentives and infrastructure’ (The Royal Society 2012, p. 10). Institutions are expected to encourage researchers to open up and share their data as well as using existing open data, by building support infrastructures, offering support and education.

Encouraging and enabling researchers to publish and share their data is posing a significant challenge for institutions. Many repositories that were created to encourage open access publications and data sharing remain almost empty. Borgman (2012) notes that, despite significant investments in, and promotion of data sharing, the ‘dirty little secret’ is that very little data sharing is actually taking place (Borgman 2012, p. 1060). She notes that relatively few studies show any consistent data release, and data sharing seems to be concentrated in a few fields. ‘[L]ittle research data is [sic] circulated beyond the research teams that produce them, and few requests are made for these data’ (Ibid.). Studies indicate that researchers are reluctant to share their data because of various concerns, ranging from being scooped,
to being unable to protect the privacy of their research subjects (Kuipers and Van der Hoeven, 2009). Moreover, open access to research data requires specific skills and knowledge that have to be developed and maintained. Yet, researchers are reluctant to participate in training courses, unless they are directly relevant for their research. Another study showed that researchers are unfamiliar with terms like ‘digital curation’ and ‘digital repository’, and suspicious of policies that issue various requirements and mandates. They prefer advice that conveys a sense of purpose and assistance (Freimna et al. 2010).

In addition, RECODE also found that institutions are also facing several barriers. Multiple conceptions of open data make it difficult for librarians and data centre professionals to help researchers across the board. Every discipline and sub-discipline requires a different kind of data expertise. Moreover, it can be difficult for librarians and data curators to make the connection between the relatively new and rapidly-evolving field of digital data management and the everyday practice of doing research. Cox and Pinfield argue that it is ‘like any area of specialist activity – complex and jargon ridden,’ and assert that a ‘whole social world of organisations, projects, thought-leaders and key influencers, technologies, discourses, concepts and terminology has to be mastered in order to be ‘taken seriously’” (2013, p. 4). Librarians may find it difficult to acquire sufficient technical knowledge of digital data management to position themselves as key players, while also maintaining domain-specific expertise and knowledge of doing research.

The established institutional culture may also represent a barrier, because open access to research data is just one of many competing priorities. Moreover, various parts of an organisation each have their own set of ambitions, so it can be a considerable challenge to raise making more research data open up the agenda, when other concerns take priority. Further progress in the area of training and skills development can also be hampered by the balance of power. Librarians may establish and administer the institutional repository with significant knowledge about scholarly communication issues but, since they do not bring any funding into the university, the library is generally perceived as a service-based unit that lacks much strategic influence. Thus, the distribution of responsibility often remains unclear. One study on research data management notes the lack of professional preparation. ‘[A]lmost no one within the academic community receives systematic professional training and certification in the management of research data. Still worse [...] virtually no one in academia perceives that they have a professional responsibility or mandate for research data management functions’ (Halbert 2013).
In some disciplines, such as particle physics, genetics or social geography, digital data management training is already an integral part of the (post-graduate) curriculum, but it remains a relatively new area in many subject areas. There are very few disciplinary-focused data curation training programmes at universities (Creamer et al. 2012; Walters 2009; Lyon 2012). Universities, departments, research groups and research institutes are only just beginning to gain experience of providing appropriate courses, workshops and tools to support researchers, librarians, information specialists and other staff in their data management activities. Important developments in this respect are the increasing number of training programmes and materials that data centres, libraries and research consortia are offering for researchers as well as the establishment of professional training programmes for data curators and information specialists. The training available for data management and curation is mainly provided by dedicated national bodies, libraries, information science schools or data centres. The UK Digital Curation Centre (DCC) plays a leading role in offering training for practitioners in need of resources on data management. The DCC offers workshops in data management as well as short, intensive half-day or three-day courses for absolute beginners (Keralis 2012). They also offer information and a range of tools to help researchers prepare their data management plans. The DANS institute provides various workshops, training courses and guest lectures for researchers and students in the humanities and social sciences at various Dutch universities and research institutions (DANS, no date).

In addition to these training initiatives from dedicated projects and organisations, some promising practices are emerging within universities. For example, the University of Southampton has, through collaboration with the UK Research Data Service and involvement in projects like the Institutional Data Management Blueprint Project (IDMB), worked to improve and formalise initiatives to support their university researchers to manage their research data (Brown and White 2013). The university aims to develop an understanding of different disciplinary needs through partnership and cooperation, to implement simple, low-cost technical solutions and applications, and to focus on training and support. Another example is the UK Orbital project at the University of Lincoln’s School of Engineering. The project has proposed a set of recommendations to support further development of their research data management structure (Stainthorp 2012). This project underlines the fact that researchers are heterogeneous not only in terms of discipline, but also between individuals in the same team. It is, therefore, important to gain an understanding of the culture
within any given set of researchers before considering how to influence their research data management behaviour.

Libraries are also considered well suited to playing a greater role in guiding researchers’ data management practices. Libraries have a long tradition of subject liaison staff, who work closely with researchers to understand their needs, so they could comprise the ‘last mile’ of research data infrastructures – ‘the part of the network that will provide connections between the systems and the researchers, and ultimately, to new users of the data’ (Gabridge 2009). Librarians could take on the role of data stewards for various stakeholders, especially researchers, by activities such as organising conferences, distributing literature, devising training courses, web tutorials and advocacy programmes tailored for specific research communities. Indeed, several physical and digital libraries, such as the Edinburgh University Library and the California Digital library, have started to develop this new intermediary role. They liaise with researchers and help them to deposit their data at the point of creation, provide advice about data standards, and create curation plans for the whole data lifecycle, in compliance with funder mandates. They also provide seminars and workshops or individual tuition for research and professional staff.

Open access to research data requires specific skills and knowledge that have to be developed and maintained. As this section shows, several institutions have taken up the challenge of educating and training researchers, librarians, information and data scientists and other professionals, building on existing and emerging digital data management practices. Libraries, data repositories, data centres and dedicated organisations, such as the DCC, have become advocates of data sharing and open access, and position themselves as valuable resource-providers for knowledge and expertise about open data.

Conclusion

Institutions such as universities, libraries, data centres, publishers, professional associations and funding bodies, are all playing an important role in making research data open. They support researchers, provide infrastructure and funding, and set best practice guidelines. Open access to research data offers many benefits, but some challenges still need to be overcome if these gains are to be realised.

One of these is the fact that, although open access may produce significant cost savings in the long term, it generates considerable costs in the short term. Open access requires significant and continuous effort
to make sure that data can be found, interpreted, evaluated and used. Another challenge is the relatively low level of data management skills and awareness about the opportunities and limitations of open research data that exists in institutions and many research communities. Cutting across these challenges are issues concerning the heterogeneity of the stakeholders, the multiple conceptions of open data and entrenched institutional cultures. Importantly, none of these challenges are specific to open access to research data, and need to be addressed by stakeholders within big data or open government data as well.

Developments in the data ecosystem that address these challenges are still in the early stages, but they indicate major shifts in roles, responsibilities and relationships within the research landscape. Whereas researchers in many disciplines used to be responsible for their own data, even after their projects had finished, this responsibility is now partly delegated to data centres and repositories. Institutions such as universities, funders, publishers and libraries have taken on coordinating roles in training, education and setting standards for data quality. They have begun to offer data services, establish infrastructures and issue policies. In terms of financial resources, knowledge and expertise, several institutions have engaged in collaborative efforts to develop data repositories, data management services, training programmes, etc. Libraries are working with data centres and other libraries to offer long-term preservation of, and access to, research data, as well as skills development programmes. Universities, research institutes, and funding bodies are participating in international collaborations. Indeed, through building alliances and (trans)national collaboration, some institutions have already become important actors in the data ecosystem, giving shape to a diversity of access arrangements and governance structures. In order to achieve a knowledge society, these stakeholders must continue this work, and stakeholders within other open data ecosystems need to be encouraged to take such leadership and coordination roles. This is happening in many contexts, but it needs to be significantly expanded, especially in relation to the citizen participation aspirations of many open data advocates.

Within the scholarly landscape, institutions are playing an increasingly important role in opening up knowledge to a broader public. Nevertheless, as this chapter has shown, opening up data in any context raises significant challenges, and making data available will not necessarily result in ease of use or accessibility across the board. In order for institutions within all of the data ecosystems to contribute to a data-capable and well-informed society, they will have to invest more into a support structure that will enable a diverse range of people to find, access, interpret and use data.