Chapter 2

The Case for Open Scientific Data: Theory, Benefits, Costs and Opportunities

Vera J. Lipton

This chapter provides the theoretical, historical and economic background for the study of open scientific data. It consists of five key sections:

1. The emergence of open scientific data
2. Science and scientific data in the evolving knowledge economy
3. The envisaged benefits of open scientific data
4. The costs of developing open data infrastructures
5. Open data and commercialisation of public research

Introduction

This chapter reviews the theories underpinning open scientific data and explains when and why the concept of open scientific data has been adopted by the research community, funders, policymakers and broader civil society. The author first considers the historical developments and the role of scientific data in the evolving knowledge-based economy. This chapter then outlines the theories that advocate open science and the open flow of data in the economy and the role of open access in fostering the dissemination and development of science.

The chapter concludes by analysing the economic arguments put forward for open scientific data—including the economic, social and scientific benefits that open sharing of scientific data is likely to generate into the future and the tensions between open science and commercialisation of public research. These benefits and tensions are illustrated in three case studies showcasing the application of open scientific data—the Human Genome Project, the *E. coli* epidemic in Germany in 2011 and the Global Positioning System.

2.1 The emergence of open scientific data

Scientists were instrumental in developing the Internet and other communication technologies, and now scientists are leading the way in applying these technologies to the creation, communication and dissemination online of the results of their work. Across many disciplines, and from many locations, researchers are using

digital technologies to share tasks, knowledge and research outcomes. Working in real time, digital technologies help to speed the creation of knowledge and its dissemination. This revolution in scientific communications and these new means for collaboration open the way to a dramatic increase in the social value of science.

Good data enables good science, and digital technologies provide the means for acquiring, transmitting, storing, analysing and reusing massive volumes of data. In embracing these technologies, research organisations and researchers are extending the frontiers of science. This is open innovation in science, or open science.

Open science is part of the broader access to knowledge movement that advocates the distribution of educational, intellectual, scientific, creative and government works online through permissive licences by the right holders (open access) [45–51]. More specifically, open science refers to online scientific resources, whether data or a publication, that anyone can access, use, reuse and distribute without permission from any other party. It may be that those resources have been placed in the public domain and so are not subject to any legal control. Or it may be that permission has already been granted to use, reuse and distribute the resources. Whichever case applies, the ability to use and build upon such resources simply requires access to them.[1]

While the definition and licencing mechanisms for open access are new, the sharing of scientific data in digital formats predates the emergence of the open access to knowledge movement and also predates the World Wide Web [54].

Open access publishing has roots in electronic publishing experiments that began in the 1970s [55] and led to the adoption of the Budapest Open Access Declaration in 2002 [56]. The origins of open access to data go back even further. The World Data Center was established in 1955 to collect and to distribute data generated by the observational programs of the 1957–1958 International Geophysical Year. Scientists from 67 countries participated in the data collection that year and agreed to share data generated from cosmic ray, climatology, oceanography, Earth's atmosphere and magnetic research, with a view to making the data available in machine-readable formats [57, 58]. One year later, in 1959, representatives of 13 governments agreed on scientific collaboration enabled by a free sharing of scientific observations and results from Antarctica [59]. In 1966, the Committee on Data in Science and Technology (CODATA) was founded by the International Council for Science to promote cooperation in data management and use [60].

In 1982, the Internet era began and the open research data made a giant leap forward shortly thereafter. In the 1990s, several Internet-based open research data initiatives were introduced—The Committee on Earth Observation Satellites (1990), International Geosphere-Biosphere Programme (1990), US Global Change Research Program [61], Inter-American Institute for Global Change Research (1992), United Nations Framework Convention on Climate Change (1992), Intergovernmental Oceanographic Commission of UNESCO (1993), Global Climate Observing System (1993), International Social Science Council (1994), World Meteorological Organization (1994), University Corporation for Atmospheric Research (1995), Human Genome Project (1996) and American Geophysical Union (1997) [62].

The success of the Human Genome Project has also drawn the attention of governments, funding agencies and scientific organisations. In 2003, the Human Genome Project was completed. The same year marked the adoption of the Berlin

[1] See Lessig [52]. There are many ways to enable open access, including through publications and data repositories, dedicated websites, journal websites, etc. The two prevalent methods of delivery are (i) publication in open access journals and (ii) self-archiving in open access repositories (Green Road). See Harnad et al. [53].

Declaration on Open Access to Knowledge in the Sciences and Humanities, which called for open access 'to original scientific research results, raw data and metadata, source materials and digital representations of pictorial and graphical and scholarly multimedia materials' [63]. The open science story continued to unfold.

An important earlier milestone was the launch of arXiv, originally developed as a repository of preprint publications in high energy physics, in 1991. The arXiv model was developed on an existing infrastructure that supported the flow of information within networks of close colleagues, known as invisible colleges [64]. Hosted by Cornell University, the model later expanded to other scientific fields and established the culture of exchange of preprint publications that later inspired the principle of open access to publications as we know it today.

Previously, many journals refused to consider papers posted online on the grounds that such posting constituted 'prior publication'. Over the years, the practice became more readily accepted by publishers, and arXiv has expanded to other fields to science. Competing platforms have emerged, such as the Sponsoring Consortium for Open Access Publishing in Particle Physics[2] developed by CERN, the European Organization for Nuclear Research and Inspire [65], an overarching High Energy Physics information system also developed by CERN and interconnected with arXiv.

The movement for open access to research data is built on those early foundations with open access to scholarly publications. In the 1990s, calls for recognition of open scientific data by key international organisations came, such as the OECD. In the Declaration on Access to Research Data from Public Funding adopted on 30 January 2004, the OECD recognised 'that open access to and unrestricted use of data promotes scientific progress and facilitates the training of researchers' [66, 67]. Three years later, the OECD codified the principles for access to research data from public funding [68].

Similar declarations and statements came later from the European Commission (2008) and were followed by mandates for open access to scientific data introduced by research funders, with the National Institutes of Health being the first in 2003. In 2010 the National Science Foundation announced that all future grant proposals would require a two-page data management plan and that the plan would be subject to peer review [69]. The policy was a tipping point in stimulating similar mandates outside the United States. In the following years, similar policies mushroomed in other parts of the world. Some countries have even legislated to require open access to research data. The policies and mandates of research funders are examined more fully in the next chapter.

Another driver for the movement in support of open access to data is the increasing use of 'e-research'[3]—the application of digital technologies to research and to research practice, whether in current or new forms. E-research is forcing some rethinking of the means for producing and sharing scientific and scholarly knowledge. The development of digital communications and novel ways of creating and handling data are creating an interest in 'data-led science' [71]. Generally, e-research refers to large-scale science involving global collaborations. Enabled by

[2] After several pilot projects, SCOAP was formally launched in January 2014 and was extended at least until 2019. From January 2018 SCOAP will also support HEP publications in three journals of the American Physical Society.

[3] The term e-science is most popular in the United Kingdom, continental Europe, Australia, and some other parts of Asia. In the United States, other parts of Asia, and other parts of the Americas, the concept of cyberinfrastructure for research is more common. The difference between these terms is interesting. One stresses the practice of research, the other the infrastructural condition for that practice, but both concepts are understood to refer to a shared view of computationally intensive research as a qualitatively novel way of doing research. See Jankowski [70].

the Internet, these typically require access to very large data collections and high-performance computing and visualisation capabilities that the originating scientists can access. Dominant in e-science are the disciplines of physicists, computer scientists, life scientists and some computational social sciences [72]. Yet the prospect of e-research has spread out across the entire scholarly community, including the interpretative social sciences and humanities.

Clearly, open access to publicly funded research data is a mainstream development associated with broader moves towards open science or Science 2.0, an approach that attempts to open up the process of scientific research for review and broader uptake. Open science advocates the sharing of scientific knowledge over the Internet by challenging the application of exclusive property rights over scholarly outputs. This movement further promotes 'digital openness' in the conduct of science and scientific communication, facilitating online access to scientific publications and research data. The resulting open scientific content and emerging online communities are reshaping the fundamental processes of science creation and dissemination.

In general terms, policymakers and open data advocates are seeking to drive changes in the ways data is created, managed, shared and reused. Adoption of these practices will require a shift in the behaviour of researchers and in established practices of research, including data sharing and data preservation practices. These processes are taking place alongside and are intimately connected with the evolution of digital technologies and interactive communications. Open science is developing and building upon the body of digital knowledge, data and infrastructure that it inherently generates. Central to these changes is a developing ecosystem, with novel means for creating and using scientific data to generate knowledge and to use this knowledge for social and economic benefits.

2.2 Open scientific data in the evolving knowledge economy

To understand the role of data and science in the changing knowledge economy, it is necessary to look at the approaches for producing and disseminating scientific knowledge and how these have developed over time and also at how science and scientific knowledge relate to other areas of society. This social role of science is important, because the characteristics of scientific data are shaped in the context of the production and use of scientific knowledge. Indeed, data, information and knowledge have become the central features of an evolving knowledge economy in which innovation plays a central role [36, 73–75]. The terms data, information and knowledge are often used interchangeably, even though many scholars have studied the evolution of these expressions and the differences among them, resulting in many books devoted to this subject [76–87].

2.2.1 Defining and differentiating the terms

Broadly speaking, data consists of figures without any interpretation or analysis [88]. Information captures data at a single point—in other words, the data has been interpreted to provide meaning for the user. And information can lead to knowledge, by combining it with experience and insight.[4] As such, data involves a lower level of abstraction from which information and then knowledge are derived [89]. A detailed discussion of concepts of 'data' and 'open data' is provided in Chapter 4 of this book.[5]

[4] *Ibid.*

[5] See Chapter 4, especially Sections 4.2 and 4.5.

In reality, however, the boundaries between data, information and knowledge are not always clear. What is data to one person can be information to someone else. What seems to matter, though, is the capacity of humans to use data and information to develop meaningful knowledge. Also important is the capacity of humans to interpret data as well as to process and absorb knowledge developed by others. These attributes have been identified to be crucial to knowledge and technology transfer [90–92]. They are also crucial factors in the development and adaptation of computer-assisted data processing and artificial intelligence.

In the operational context, the Reference Model for an Open Archival Information System defines information as 'any type of knowledge that can be exchanged. In an exchange, the knowledge is represented by data' [93, 94]. However, data can change over time. The mistake people often make is to think that the information presented is always an accurate reflection of data. Yet that information can only be as accurate as the data underpinning it, and as data changes so can the information derived from it. Buckland [86] looked at the subtle differences in further detail and distinguished between information as process, as knowledge, or as a thing [86].

The differentiation between 'information' and 'knowledge' is also apparent in the theories underpinning the creation of knowledge in society. There are subtle differences between the information society and the knowledge society, the two terms most commonly used. Although the term 'knowledge society' was coined by Drucker back in 1969 [95], further development occurred only in the 1990s by Mansell & Wehn [96] and Stehr [97]. According to Wessels et al. [89] and Castelfranchi [32], knowledge society produces, extracts value from, and makes data available to all its members. The key objective of the knowledge society is to improve the human condition [32].

However, a knowledge society cannot be achieved simply by providing universal education, nor can it be achieved by making information technologies available to everyone, or by making information previously accessible only to selected circles available freely to all.

Castelfranchi argues that the driving force of a knowledge society is its 'cognitive capital'—that knowledge has become an actively productive factor of economic development.[6] He also notes that knowledge itself is 'intrinsically motivated' and that a real knowledge society would be a society guided by this value—meaning that the motivation to engage in knowledge consumption and production would arise from within the society and its members because knowledge is naturally satisfying to them. But this is exactly what is not happening, Castelfranchi notes. The proposed vision for a knowledge society is one of an instrumental and subordinated activity; it is a society in which knowledge has to justify its utility and in which science is no longer a curiosity-driven activity. He goes on to say that today even virtues have to demonstrate they are 'useful'.[7]

In this sense, the definition of a knowledge society is similar to the theory of an 'information society' that treats information as the key commodity in production, consumption, and innovation. Information can be used to create knowledge to fuel innovation and economic growth. However, knowledge in an 'information society' circulates within selected economic, political and social networks and has a more limited social agenda of inclusion than a knowledge society.[8]

A knowledge society is distinct from an information society and a knowledge economy because

---

[6] *Ibid*, 1.

[7] *Ibid.*

[8] Wessels, at point 24.

> *... it sees information and knowledge as open to all. Its central value is openness, which means that data, information and knowledge are seen as a 'commons' or shared asset in society. This has the potential to allow any member of society to use data to engage and participate in economic, social, political and cultural projects.*[9]

Both theories—a knowledge society and an information society—posit that the creation and accumulation of knowledge can lead to economic growth. As such, the vision of both theories is the creation of a knowledge economy.

To understand the role of data and science in this evolving knowledge economy, it is necessary to briefly look at the ways of production and dissemination of scientific knowledge and how these have developed over time and how science and scientific knowledge relate to other areas of the society. The social role of science is important, because the characteristics of scientific data are shaped in the context of the production and the use of scientific knowledge. This approach was also advocated by American sociologist Robert K. Merton, who argued that the production and role of knowledge need to be understood through the 'modes of interplay between society, culture and science' [98]. Specifically, he studied the relationship between science and religion.

2.2.2 Mertonian science

Following on from the claim by Max Weber that the Protestant work ethic drove the emergence of the capitalist economy, Merton argued that the ascendance of Protestantism and the arrival of experimental science were similarly interwoven [98, 99]. Merton held that science became popular in seventeenth century England and was taken up by the Royal Society, which at that time was dominated by Puritans and other Protestants, because Protestant values corresponded with the emergence of new scientific values, resulting in 'modern science'. Merton separated science from religion, which was a major shift in understanding the position science has in society. In particular, Merton differentiated between science as 'handmaiden' to theology during the Middle Ages and the 'modern science' emerging from the seventeenth century onwards. This shift from science as an adjunct of theology to 'modern science' is also known as the Scientific Revolution.[10]

Merton also defined the four sets of norms of modern science as the following:

- *Communalism*—the common ownership of scientific discoveries, according to which scientists give up intellectual property in exchange for recognition and esteem.

- *Universalism*—according to which claims to truth are evaluated in terms of universal or impersonal criteria, and not on the basis of race, class, gender, religion, or nationality.

---

[9] *Ibid.*

[10] The transformation of science into an autonomous discipline began in Europe towards the end of the Renaissance period and continued through the late eighteenth century. This scientific turn also influenced the Enlightenment. The 'modern science' included mathematics, physics, astronomy, biology (including human anatomy), and chemistry. The institutionalisation of modern science was marked by the establishment of the Royal Society in England in the 1660s and the Academy of Sciences in France in 1666.

- *Disinterestedness*—according to which scientists are rewarded for acting in ways that outwardly appear to be selfless.

- *Organised scepticism*—all ideas must be tested and be subject to rigorous, structured community scrutiny.[11]

These four characteristics are often referred to as the principles of the Mertonian sociology of science and are often put forward for the development of open science.

2.2.3 Modern science

Thomas Kuhn elaborated on the concept of scientific revolutions in 1962. In his seminal work *The Structure of Scientific Revolutions,* commonly viewed as one of the most influential books of the twentieth century, Kuhn challenged the Mertonian view of progress with what he called 'normal science'. In Kuhn's view, scientific change occurs as a process with a number of stages, leading to paradigm change. He argued that 'normal' scientific progress occurs through the accumulation of generally agreed facts and the theories built on them. Kuhn argued that progress occurs episodically—periods of conceptual continuity, or 'normal science', are disrupted by episodes of 'revolutionary science'. During such revolutionary periods, anomalies are discovered that challenge established theories and lead to new paradigms requiring that old data be questioned in new ways. Consequently, a new paradigm moves from the 'puzzle-solving' function of its precursor and so changes the rules of the game by motivating renewed research activity.[12]

As in any community, Kuhn argued that some scientists are bolder than their colleagues. Whether because they see that a problem exists or for some other purposes, these pursue 'revolutionary science' to explore alternatives to established assumptions. From time to time, such activity creates a rival framework of scientific thought. The candidate paradigm, being new and incomplete, may appear to have numerous anomalies and will meet opposition from the general scientific community. However, at some point, the attitudes of scientists will change, and the anomalies will finally be resolved. Those with the ability to recognise a new theory's potential will be the early adopters of the challenging paradigm, Kuhn said. Over time, as the challenger paradigm is tested and as views unify, it will replace the old model. Thus, a 'paradigm shift' occurs.[13]

2.2.4 Digital science

One of the paradigm shifts in science envisaged by Kuhn was the emergence of the Internet and communication technologies. Toffler [102] coined the term the 'Third Wave', which he saw as the Information Age that succeeded Industrial Age society (the 'Second Wave') in developed countries [102, 103]. The new society characterised the combination of knowledge and information as the principal factor in the exercise and distribution of power, replacing wealth. The era is further

---

[11] Merton, at point 33.

[12] See Kuhn [100]. Asking new questions of old data on pages 139, 159. Moving beyond 'puzzle-solving' on pages 37, 144. Change in rule sets on pages 40, 41, 52, 175. A similar view was expressed by Bronowski in [101], that is, all fundamental scientific discovery 'opens the system again' and 'to some extent are errors with respect to the norm' on pages 108 and 111.

[13] *Ibid,* Kuhn [100].

characterised by the emergence of novel technologies and scientific fields such as global communications networks, DNA analysis and nanotechnology.

Toffler also predicted that the rise of the Internet will transform the very nature of democracy. In an interview discussing his book, Toffler said that the centralised, top-down management and planning used in industry would be replaced by a style that he called anticipatory democracy—more open, democratic and decentralised.[14] However, Toffler was aware of the limitations of the Information Wave. He was convinced that a society needs more than just cognitive skills: it needs skills that are emotional and affectional. He said that a society cannot be run on data and computers alone [104].

At the same time, the emergence of the Information Wave has ushered in a new era in scientific communications and impacted the production and practice of science. Previously, modern science had become the exclusive system in society for knowledge production, with few opportunities for lay people and amateur scientists to participate in science utilisation and production [105].

This has changed with the evolution of the Internet, which is:

> *... shaping the move away from traditional science and research while, at the same time, developing further ... not least influenced by the development it has originally initiated [106].*

In the second stage, the changes induced by new communication technologies have led to digital science, sometimes also referred to as cyberscience, or open science, as qualitatively distinct from 'modern science'. Nentwich [106] has argued that this model leads to a qualitative 'trend extrapolation'[15] and to the more or less complete replacement of old ways of practising science as the result of enabling by new cybertools.

Similarly, in the book *The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies* published in 1994, a team of authors proposed that there had been a shift in the production of scientific knowledge from what they termed 'Mode 1' to 'Mode 2'.

Mode 1 was *characterised by the hegemony of theoretical or, at any rate, experimental science, by an internally driven taxonomy of disciplines and by the autonomy of scientists and their host institutions, the universities* [33].

Mode 2 is a new approach to knowledge production that is *socially distributed, application-oriented, transdisciplinary, and subject to multiple accountabilities*.[16]

As a result of the emergence of digital communications, the production of scientific knowledge in *Mode 2* is more centrally located within social relations. This also means that data is viewed in a different way to that found in *Mode 1*. The key difference is that if data is produced through publicly funded research, then the broader public should have a right to access it. Furthermore, according to *Mode 2* knowledge production, data is seen as having value through its reuse by a broader range of users than just the research community that produced the data.[17] Data users include researchers, policymakers, businesses and citizens, and there is a belief that each of these users will advance within their own domains through the democratisation of science. This, in turn, will lead to a more informed public that is better able to participate in social debates and the development of the society [27].

---

[14] *Ibid*, Toffler [103].

[15] *Ibid,* 48.

[16] *Ibid.*

[17] Wessels, at point 24, 56.

At the same time, research practices are also becoming more complex and include input from societal actors such as business and research funders (whatever their motives), among others.[18] So it can be argued that the users of scientific data are likely to include two types of external actors—transdisciplinary users applying scientific data to advance their own fields of endeavour, such as new technology development, and scientific users applying the data to advance science. However, the increasing number of stakeholders and groups also means that they may hold differing views of what data is and what it means for knowledge production, and this may complicate the process of making data open.[19]

Yet another issue arising in *Mode 2* of scientific knowledge production is in the form of the market forces that influence society and the production of science. Science is rarely characterised by the open paradigm, Fuller argued [107]. Researchers have the tendency to organise themselves in competitive networks, each seeking to control funding, academic appointments and the conduct of associations and journals. Much of the history of science documents those struggles and the displacement of one network by another.[20] These pressures are likely to continue into the future, even though digital science makes research more transparent and enables checking of research quality.

In *Mode 2*, the criteria that determine quality are indicative of a broadening social composition in the system of review. The implication is that 'good science' becomes more complex to assess; it is no longer confined to the judgements of peers within the discipline. Broadening the review system does not, however, necessarily mean that the research becomes of lesser value. Instead, it gains complexity ([33], p. 8).

Other market influences on the mode of scientific production are the increasing linkages between industry and academia and the push to commoditise and commercialise research. These are further discussed in the section below dealing with the inherent tensions between open data and commercialisation of scientific knowledge.

For now, I will outline the issues raised by Nowotny et al. [33] who observed that specialised knowledge plays a crucial role in many dynamic markets.[21] Specialised knowledge holds a vital place as a source of created advantage—both for its producers and users of all kinds. As a result, the demand for specialist knowledge is increasing. The core of the thesis of science production introduced by Nowotny et al. is that the expanding numbers of potential knowledge producers run in parallel with the expanding demands for specialist knowledge. The effect is to create the settings for the evolution to a new model for scientific knowledge production.

This and the push to commoditise and commercialise research have implications for all institutions—whether in the academic world or as private sector research stakeholders—that have an interest in the production of scientific knowledge. As markets for specialised knowledge emerge, so must the game change for all these institutions, albeit not necessarily all at the same pace.[22] The economic aspect of these developments is that knowledge-based innovation enables companies to generate market power and monopoly rents because, even though knowledge is non-rivalrous (can be used simultaneously by many agents without detracting from its

[18] Wessel at point 24, 119.

[19] *Ibid.*

[20] *Ibid.*

[21] *Ibid*, 12.

[22] *Ibid*, 13.

utility), it is at least partially excludable (innovating firms can restrict access to the novel features of their inventions) [108]. This causes problems in that public knowledge can be easily appropriated as a private good, the economic returns of which may not later return to the broader society. For these reasons, the cultivation and preservation of the scientific commons are of outmost importance.

From the discussion above, it is clear that the changing modes of both science production and science utilisation in the Internet era have had a significant impact on the ways scientific knowledge and data are created and utilised. It is clear that the combination of a proliferation of data and the open data movement is a significant feature in advancing a knowledge society and the attendant knowledge-based economy.[23] The actions of many actors, who organise themselves in networks and interact with a range of public and private level stakeholders with whom they exchange digital data, are a key aspect of this process. To increase these linkages and knowledge utilisation, it is necessary to make the data widely available. Open scientific data can bridge this gap while contributing economic and social benefits.

However, to achieve this goal, the data will have to be provided in a manner that permits not just sharing but also reuse across society.[24] This aspect is not well covered in the theories of knowledge society, which for the most part envisage that merely releasing scientific data into the public domain is sufficient for the benefits of open data to accrue. The model proposed in this book rebuts this argument, positing that simply providing *access* to data in the public domain is useless to the society and that only data *reuse* can realise the envisaged benefits. These aspects are canvassed in Chapter 8.

2.3 The envisaged benefits of open scientific data

Of the many benefits put forward for the adoption of open science in general and open scientific data in particular, some can already be seen in practice, while others will only become apparent as open data collection increases.

2.3.1 Solving great problems facing humanity

Open scientific data is important because the need to share scientific outcomes has perhaps never been greater. As nearly every region feels the effects of climate change, as conflict and food insecurity are rising, and as the demand for natural resources increases, the world looks to science for solutions. This world is interconnected—over 40% of its population was able to access the Internet in 2013, and the number of users online is growing exponentially ([109], p. 3). In this global digital village, open science offers hope—hope for those living in prosperous societies and hope for the remaining half of the globe, over 3 billion people, who live on less than US$2.50 a day [110].

Ease of access to scientific data, knowledge and application will play an enormously significant role in the planet's future well-being. Yet it may not be science alone, but rather the knowledge and discipline that it imparts and the learning that it yields when shared broadly and applied wisely. For science to deliver its full value to the society, it must be easily and freely accessible [5].

[23] Wessels at point 24, 14.

[24] *Ibid.*

2.3.2 Increased dissemination and impact of research

At present the majority of science is not easily accessible, and only a fraction of it is freely accessible despite the fact that scientific knowledge is plentiful and growing rapidly—doubling, on average, every 15 years.[25] Indeed, the current system of science generates massive volumes of knowledge and data. Yet much of the knowledge and data stays locked in institutional repositories, costly scientific journals, or patent applications. Locking up knowledge does not contribute to the greater good. Statistics confirm this—90% of scientific publications are *never cited*, and up to half of the world's scientific papers are *never read* by anyone other than their authors, referees, or editors [111, 112], while 98.5% of patents are *never asserted* [113]. Many scientific outcomes are lost because of the failure to make them available to those who could use them and add value. This gap between the capacity for science creation and its dissemination is a 'dual tragedy'—a tragedy of science and a tragedy of society—as Australian science commentator Julian Cribb put it.[26]

Open science can help bridge this gap. The Internet, Web and social networking have created new opportunities for disseminating scientific research, by sharing research data sooner and more widely. Much science is publicly funded, and the society increasingly expects that the outcomes of public science will be freely available. In the United Kingdom, Australia and in many other countries, universities constitute the primary recipients of government funding for research. In recent years, governments in these countries have taken considerable steps to develop mechanisms to increase the economic, social and environmental impact of science. Releasing research data is a logical step.

The 2009 study of the economic effects of open access to Australian public research found that a one-off increase in accessibility to public sector research and development produces an estimated return to the national economy of A$9 billion over 20 years.[27] The potential economic benefits of open research data are immense, indeed. In addition, there is an increased research impact realised from investing in curated research data activity. Early evidence shows that when researchers make their well-managed and curated data accessible along with publications, they can expect an increase of up to 69% in the number of citations.[28]

2.3.3 Reduced duplication of research effort

Open scientific data has the potential for significant savings to be realised through better targeting of scientific effort and reduced duplication of research. Scientists, especially early career scientists, devote a great deal of their time to data collection. Moreover, the cost of collecting data for multiple research projects can be high, especially for clinical trials and drug testing [116]. If projects complement or build on one another, why would it be necessary to provide funding for a research team to generate new datasets when another existing dataset could shed light on the problem? Further, is it really necessary to create a dataset that would be used just by one research team for a single project and then be discarded?

---

[25] See Larsen and von Ins [6]. The rate of doubling of the body of scientific knowledge was calculated as an average number of scientific records included in the following databases: Web of Science (owned by Thomson Reuters), Scopus (owned by LexisNexis), and Google Scholar. Duplicate entries were removed.

[26] Cribb, at point 62.

[27] See Houghton and Sheehan [114]. Public sector R&D was defined as 'the proportion of R&D stock available to firms that will use it' and 'the proportion of R&D stock that generates useful knowledge'.

[28] See Piwowar et al. [115]. Their subsequent research found that cancer clinical trials that share their microarray data are cited about 70% more frequently than clinical trials that do not.

Data that is shared, reused and recycled can achieve savings or free resources for new research. It is important that both scientists and research funders recognise this. Given the wealth of information collected in clinical trials, it is apparent that there is a variety of secondary uses that could enhance scientific advances in ways not foreseen by original authors. Indeed, the ability to access and reuse existing research can enable follow-on research and discoveries faster and more cheaply and can also facilitate the reproducibility of results.

2.3.4 Enhanced quality of scientific outcomes and methods

Openness has been the core principle of scientific enquiry since the early days of modern science. Henry Oldenburg, a German theologian and the first secretary of the Royal Society, pioneered the peer review of scientific publications. In 1655 he referred to the printing press as:

> *... the most proper way to gratify those [who] ... delight in the advancement of Learning and profitable Discoveries [and who are] invited and encouraged to search, try, and find out new things, impart their knowledge to one another, and contribute what they can to ... the Universal Good of Mankind [117].*

Oldenburg's contemporary, Irish scientist Robert Boyle, sets two other precedents that shaped the future of science. Boyle published his results in lively English, making them accessible to those who did not speak Latin or were not trained as scientists. He also described his experiments in great detail so that others could reproduce them. In short, Boyle believed that science belonged to everyone and the principles of science could be tested and repeated by anyone [118].

The vision of open science is to enable scientists and the general public to access and scrutinise scientific results—to 'search for the truth'—as double Nobel Laureate Linus Pauling famously defined science. And the truth is often interpreted to be an evidence. Science is based on the best evidence we have at the time. Evidence identifies what is true and what can be trusted. As science develops, new evidence confirms or rebuts previous evidence, resulting in self-correction. But often circumstances do not allow scientists to be 100% certain that their findings are true. They work with the best evidence available.

Scientific data typically presents the evidence. This needs to be assessed as to its degree of reliability, which then determines the degree of confidence that can be invested in the conclusion. In borderline cases, computer algorithms and replicated computer analyses can be used to probe the results. More often than not, computers can do science faster and more accurately than humans. Increasingly, they can perform computations that humans cannot. Open scientific data can serve as the springboard for computational science, or e-science. Such science brings high integration of modelling and simulations into the methodologies in particle physics, bioinformatics, earth, geospatial and social sciences.

2.3.5 Enhanced education

The long-term stewardship and open availability of research data also present better educational opportunities across all ages, all disciplines and all around the world. At the secondary education levels, students can use open data repositories to further their scientific understanding and skills. University students need open data to experiment with or to learn the latest data management techniques. In the digital era, the development of data science and data management and curation skills, which

require a good educational foundation, is of particular interest to governments. These are growth areas for employment in an era of shrinking job opportunities [119].

2.3.6 Improved governance

Open research data repositories can play a role in supporting good governance. Openness of scientific information empowers non-scientific communities and the wider public to participate in knowledge creation and utilisation. Open datasets also enhance public decision-making [120], and open data policies can broaden the influence of governments [121]. Countries with limited public resources for devoting to science can benefit even more from access to public data resources [122]. In the context of increasing commoditisation of science, the governance of scientific data will become more important. Open scientific data empowers researchers, not markets, to control scientific knowledge into the future. Open scientific data thus leads towards a more transparent and accountable governance of science that, in turn, advances a more open, collaborative and democratic society.

2.3.7 Envisaged economic benefits and costs of open scientific data

Quite rapidly, data is becoming 'the lifeblood of the global economy' and represents 'a new type of economic asset'—the European Commission has recently stated [123].[29] Between 2008 and 2012, worldwide cross-border trade in data increased by 49%, while trade in goods or services rose by just 2.4% [124]. As the world adopts new technologies facilitated by data—technologies such as artificial intelligence, blockchain and robotics—open scientific data presents enormous opportunities to reap economic benefits. Governments and industry recognise that knowledge of the use of these technologies provides a decisive competitive advantage—in better performance; in providing products better tailored to the user, through new services; and in fostering innovation.[30]

Data holds the enormous potential to create jobs and increase our wealth. In the European Union alone, 100,000 new data-related jobs will be created between 2014 and 2020 [125]. Another recent study found that big data analytics solutions have the potential to unlock an additional £241 billion (2015 prices) in economic benefits for the United Kingdom over the period 2015–2020.[31] This is equivalent to an average of 2.0% of that country's gross domestic product (GDP) per year. The global market for data-related hardware, software and professional services is booming at even faster rate and is predicted to reach €43.7 billion by 2019, or 10 times that of 2010 [127].

Such statistics demonstrate the impressive economic value of data in our society. Those data-related services predicted to grow dramatically include data-centre computing, networking, storage, information management and analytics. Public research organisations, including universities, are very well positioned to provide such services. Open scientific data can therefore be a precursor for such organisations seeking to expand into these areas. In the first place, however, the infrastructures for open scientific data need development. In addition to direct economic benefits in terms of employment opportunities for researchers and analysts, such infrastructures will generate additional economic benefits derived from supporting

---

[29] European Commission (2017).

[30] *Ibid.*

[31] From 2015 to 2020, the total benefit to the UK economy of big data analytics is expected to amount to £241 billion or £40 billion on average per year [126].

the goals of research and innovation. It may take some time to identify and to measure those wider benefits, however.

In the context of open data, economic studies that have established the benefits of public sector information (PSI) and, more recently, open research data exist. The studies measuring the economic benefits of PSI [128–133] all concluded that the benefits accrued would exceed the revenue received from charging users for data. In Europe, the direct PSI reuse market was quantified to represent €32 billion in 2010 and was growing at the rate of 7% annually [130].

The experience of the Australian Bureau of Statistics (ABS) provides a specific, documented example from a data-intensive government institution, one of only a few to compare the before-and-after effects of moving from a user-pays model to an open access policy. The study showed that after adopting a CC-BY common-use licence, the ABS saved the costs of sales transactions and sales staffing and experienced far fewer licence inquiries and so less demand on staff resources, as well as broad social uptake. The savings for the ABS amounted to about A$3.5 million per year and for users around A$5 million, among other efficiencies and accrued benefits.[32]

2.4 The costs of developing open data infrastructures

A major shortcoming of the economic studies measuring the impact of PSI is their inability to quantify, or at least to estimate, the level of public investments required to develop the underlying infrastructures for data release. In many cases, such infrastructures would have been well established before open access to data was introduced. In other cases, the infrastructures evolved over time and required modernisation or just a simple upgrade to enable packaging of data products, as was the case with the ABS [134]. However, the costs of developing open data infrastructures should not be underestimated.

With regard to research data repositories, several recent studies in the United Kingdom and Australia combined qualitative and quantitative approaches to measure the value of research data and measure its impact [135–140]. These studies have covered several research fields and organisations—including the Economic and Social Data Service (ESDS), the Archaeology Data Service (ADS), the British Atmospheric Data Centre (BADC) and the European Bioinformatics Institute (EBI). All the studies are based on the economic evaluation framework, incorporating both quantitative and qualitative methods, developed by Beagrie and Houghton. The economic methods are based on estimating a range of values—from those focusing on minimum values to methods that measure some wider impacts.

They incorporate two ways of expressing return on investment in the data centres—firstly, the ratio of users' value to investment in the centres, and secondly, the ratio of value of the additional reuse of the data hosted to investment in the centres, as depicted in **Figure 2**.[33,34] The proposed model is interesting and useful because it captures not only the user value (economic benefit) but also the investment value (economic costs).

Four interesting findings of the economic studies stand out.

Firstly, the value of research data to users was found to exceed the investment made in data sharing and curation in all the studies.[35] Secondly, research data have had substantial and positive efficiency impacts, not only in terms of reducing the

---

[32] *Ibid.*

[33] Beagrie and Houghton [138] at point 86.

[34] *Ibid*, 9.

[35] *Ibid.*

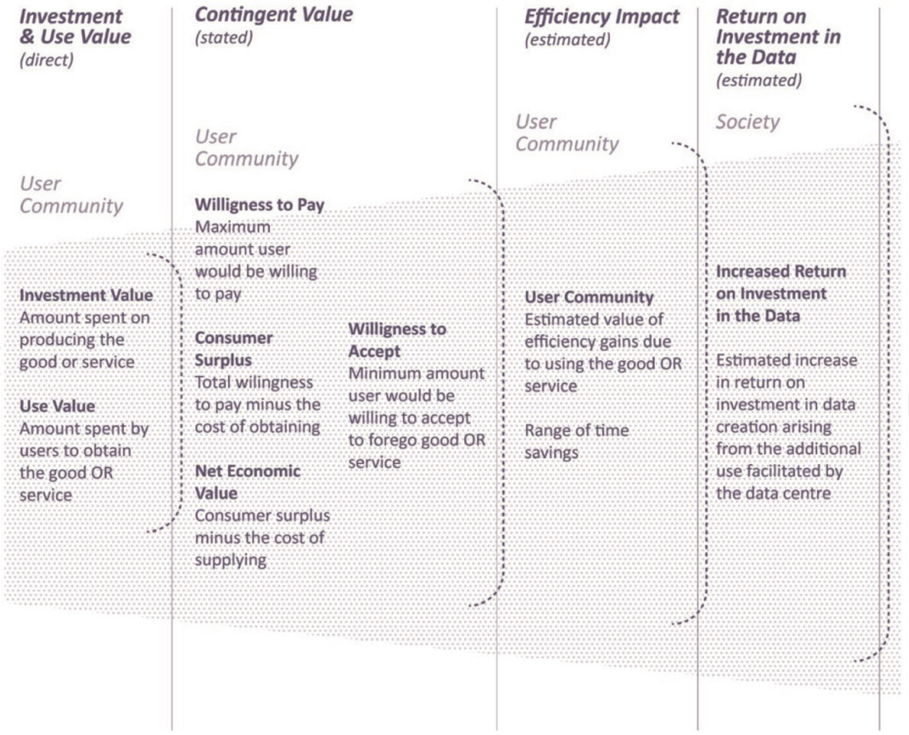

Figure 2.
The model estimating the value of open scientific data.

cost of conducting research but also enabling more research to be done—to the benefit of researchers, research organisations, their funders and society more widely. Thirdly, substantial additional reuse of the stored data was documented, with between 44 and 58% of surveyed users across the studies saying they could neither have created the data themselves nor obtained it from elsewhere.[36] Finally, the evaluation indicated that research data stored in repositories is reused by a wide range of stakeholders. Close to 20% of respondents to the ESDS and EBI user surveys, around 40% of the BADC user survey, and almost 70% of the ADS survey were from the government, non-profit and commercial sectors. Consequently, the value of public research data is being realised well beyond the academic sector.[37]

A unique feature of the ADS Impact Study was the inclusion of an analysis of the evolving, cumulative value of its archive, while other studies only provide a snapshot of the repository's value (which, the authors argue, can be affected by the scale, age and prominence of the data). In this regard, Beagrie and Houghton noted that in most cases data archives are appreciating rather than depreciating assets. Most of the economic impact is cumulative, and it grows in value over time. It will be important to capture this cumulative appreciative effect in future studies. Like libraries, data collections become more valuable as they grow, provided that the data remain accessible, usable and used.[38]

The early evaluations show that the economic benefits of open research data are already felt across many sectors. At the same time, the costs of developing open data infrastructures can be high, too. As Stallman has said, open is more akin to free speech than to free beer [141].[39]

Moreover, the responsibility for developing such infrastructures is not clear, which can lead to tensions. While the economic value of data grows over time, data

---

[36] *Ibid*, 4–5.

[37] *Ibid.*

[38] *Ibid.*

[39] Stallman (2002).

also needs to be curated over time, with significant costs. At present, most research grants only appear to cover data curation in the course of a research project and do not provide for ongoing curation. This aspect is not covered in the methodology developed by Beagrie and Houghton and needs to be explored further.

Berman and Cerf [142] discussed possible ways of funding open data infrastructures and concluded that there is no obvious actor to cover the costs [142]. Their assessment was that public research organisations are unlikely to allocate enough resources to support open data. The costs of infrastructure would absorb a great portion of their research budgets, and this is clearly not a sustainable option. The private sector has the capacity to develop such infrastructures; however, the business case and incentives for that involvement appear to be lacking.

One example of this difficulty is Google, a brand that is synonymous with data access. In early 2008 the company announced the Google Research Datasets program to store and make freely available open source scientific datasets, but by the end of that same year, the company had decided to end the project, diverting those resources elsewhere.[40] University libraries do not have the funds to curate open data, either. The solution, according to Berman and Cerf, might be an increased focus on developing partnerships and linkages[41] between the public and private sectors.

Another model is to develop supranational or national data infrastructures as is the case of the European Open Science Cloud spearheaded by the European Commission [144]. While the Commission is still working with member states on the definition of governance and financing for the initiative, the project is gaining a momentum. It is envisaged that over time, a co-funding mechanism mixing different revenue streams will be set up to increase the accountability, build trust, share resources and build long-term capacity for European research data [145].

One further economic challenge associated with the implementation of open research data is the restriction on data sharing because commercialisation appears to be a greater priority for policymakers, as discussed in the following sections.

2.5 Open data and commercialisation of public research

Open science challenges the application of exclusive property rights over scholarly outputs and calls for free access and reuse of scientific outputs. However, the open science movement has emerged at the time when major governments are decreasing their funding for research[42] and when there is a shift in the private

[40] *Ibid.* See also Google Blogoscoped [143].

[41] *Ibid.*

[42] Spending on R&D in government and higher education institutions in OECD countries fell in 2014 for the first time since the data was first collected in 1981. Countries with declining public R&D budgets include Australia, France, Germany, Israel, the Netherlands, Poland, Sweden, the United Kingdom, and the United States. See OECD [146].
In the United States, for the first time in the post-World War II era, the federal government no longer funds a majority of the basic research carried out in the country. Data from ongoing surveys by the National Science Foundation (NSF) show that federal agencies provided only 44% of the US$86 billion spent on basic research in 2015. The federal share, which topped 70% throughout the 1960s and 1970s, stood at 61% as recently as 2004 before falling below 50% in 2013. Also, in the United States, investments in research and development as a percentage of discretionary public spending have fallen from a 17% high at the height of the space race in 1962 to about 9% today, reflecting a shift in priorities of the government. The biggest decline has taken place in civilian research and development, which has dropped significantly as a proportion of both GDP and federal spending. See Mervis [147].
In the United Kingdom, the research funding slumped below 0.5% GDP in 2015 and has been declining steadily since 2009. See Rohn et al. [148].

sector to increasingly draw on public research.[43] Many governments now require publicly funded research organisations to increase the impact of public research and generate income through the protection and commercialisation of intellectual property, including through the creation of start-up enterprises.[44] The need for commercialisation has affected the goals of government research funding, causing public sector research agencies to justify the success of their research by proving or providing a convincing argument for the future economic value of their science and technology bases [151]. In countries such as Australia and the United Kingdom, universities supplement a vast portion of their income from tuition fees received from international students.[45]

The push towards commoditisation and commercialisation of public research leads to new tensions [38, 40]. Data from publicly funded research can have commercial value and lead to new partnerships with industry. Open access to research data also leads to new business models that will enable significant economic returns to be realised a few years down the track, as was the case with the open genomic data, as illustrated below.

The seemingly opposing trends towards opening research data and increasing the commercial returns from public research appear to be closely connected to the development of new technologies. On the one hand, policymakers are trying to open up research data to speed up innovation and the development of new technologies; on the other hand, they are trying to privatise and protect more and more research and emerging technologies with intellectual property, thus preventing the data and research from being shared in the future. These tensions were already pronounced in the early stages, as demonstrated in the project to map human genes.

2.5.1 Human Genome Project

In 1984, the US government started planning for a grand scientific project looking to map and decipher the entire human genome. But instead of doing it in secret laboratories in one country alone, this project brought together genome sequencing institutions from around the world. In the early 1980s scientists in many

---

[43] For example, in the pharmaceutical sector in the United States alone, roughly 75% of the most innovative drugs, the so-called new molecular entities with priority rating, trace their existence to the National Institutes of Health (NIH). See Angell [149].
Chesbrough has shown that technology companies require timely access to knowledge as they increasingly innovate by combining research outputs from external and internal sources and increasingly draw on research from universities and other public research organisations [11, 12].

[44] The policy measures advocated by the OECD in this regard focus on balancing stable institutional funding with a fair level of pressure from competitive R&D project grants, on encouraging the commercialisation of public research, and on improving science-industry relations and other linkages within the national innovation system and internationally. Increasing public research links with industry and their contribution to innovation is another main policy objective, because there is increasing pressure for public investments in research to be held accountable for their contribution to innovation and growth. Two types of measures are typically used—one is to link public research organisations and universities to other innovation system actors, particularly firms, through collaborative R&D programmes, technology platforms, cluster initiatives, and technology diffusion schemes and another is to better commercialise the results of public research through science and technology parks, technology incubators, and risk capital measures in support of spin-offs, technology transfer offices, and policies on intellectual property of public research. Source: OECD Public Research Policy [150].

[45] In 2016, over 20% of revenue of Australian universities, 6.25 billion AUD was received from fee-paying overseas students. See Department of Education and Training [152].

countries started to use computer technology in researching genetics and DNA sequences—developing processes for generating such data in digital formats.

Encouraged by these early experiments, the Human Genome Project got underway in 1990 and was initially funded by the Department of Energy and the National Institutes of Health in the United States. Their laboratories were joined by over 20 collaborating institutions from across the globe, including from the United Kingdom, France, Germany, Japan and China [153]. In 2003 the International Human Genome Sequencing Consortium announced that its project was complete —2 years ahead of schedule, under budget, and with 99.99% accuracy.[46]

The success of the Human Genome Project resulted from the convergence of science, technology and society in recording one entire human DNA sequence—around 3 billion letters of genetic code. This is the code that opened doors to improved understanding of human health as well as the detection and diagnosis of many diseases.

An important accelerator was a meeting in 1996 of representatives from sequencing centres around the world. At that meeting in Bermuda, scientists committed to make genomic data publicly available prior to publishing their findings in a scientific journal. Agreement on that principle was among the major achievements of the Human Genome Project and has, it is argued, had as much influence as the sequencing outputs themselves.[47] Over the years, sharing of genomic data has become a more established practice and biological research has exploded. The practice of data sharing demonstrated the enormous capacity of the research community to mobilise—a shift in how scientists work together as a global community to create knowledge.

The commitment to data sharing resulted from a fierce battle over the nature and ownership, and ultimately control, of the human genome. Two years after the Bermuda meeting a private gene sequencing company called Celera Genomics was set up in California. Celera owned a sizeable number of genome sequencing machines and aimed to build its own human genomic database, which it would only make available to subscribers. Celera also intended to claim ownership of 300 clinically-important genes and, at some stage, filed over 6000 patent applications to this end.[48] The emergence of this powerful competitor created a fresh impetus for the Human Genome Project. In the United Kingdom, one of the key scientists in the field later suggested that 'it has not been a race but a battle to ensure that the tools to speed biomedical research were available to all' [155].

The battle went on for about 3 years. On 26 June 2000 the White House hosted a press conference that changed the rules. In front of representatives of the International Human Genome Consortium and Celera Genomics, President Bill Clinton announced that both public and private research teams were committed to publishing their genomic data simultaneously, for the benefit of researchers in every corner of the globe.[49] Later that year, the Human Genome Sequencing Consortium published in

---

[46] *Ibid.*

[47] *Ibid.*

[48] Dr. Craig Venter, the Managing Director of Celera, later abandoned most of these applications, in response to promises made at US Congress in 1998. Releasing the entire human genome into the public domain extinguished patentability of all applications filed after the release date. A patent search conducted in 2009 revealed only 4 patents granted to Celera. See Cook-Deegan and Heaney [154].

[49] Office of the Press Secretary, *Remarks Made by the President, Prime Minister Tony Blair of England (via satellite), Dr. Francis Collins, Director of the National Human Genome Research Institute, and Dr. Craig Venter, President and Chief Scientific Officer, Celera Genomics Corporation, on the Completion of the First Survey of the Entire Human Genome Project*, media release, The White House, Washington, 26 June 2000 (http://www.genome.gov/10001356).

Nature while Celera's findings appeared in *Science*. The methodology presented by Celera was criticised by many scientists, who argued that the company's assembly of the genome would not have been possible at the time without the data released by the Human Genome Consortium. In retrospect, Celera may have well been the first commercial user of open genomic data published in GenBank—a distributed database that stores the DNA sequence in various locations around the world.

The sequencing of the human genome, as a single undertaking, had a scale unmatched in the history of biological science. The resulting dataset spearheaded the democratisation of science and has transformed medicine, renewable energy development and food production across the globe [156]. Open genomic data brought together an understanding of the whole of humanity for the benefit of all. And the United States, the key investor in the project, has reaped the majority of the economic benefits of the project.

Today, GenBank supports a multi-million dollar genomics research industry to develop DNA-based products. The initial investment of the US government of approximately US$3.8 billion, or approximately 0.075% of the country's GDP, has developed the critical tools to help identify, treat and prevent the causes of many diseases. The project further created huge growth opportunities for the high-tech American biotechnology industry, which accounted for more than three-quarters of US$1 trillion in economic output, or 5.4% of GDP, in 2010 [157]. The project further created over 300,000 jobs in the United States alone.[50] A single private-sector actor would have never succeeded in creating such a spillover of knowledge and innovation—government funding of open data infrastructures did.

In the Human Genome Project, government-funded research has played an active role in innovation and the creation of new markets for that innovation, with the resultant economic growth. It is this kind of innovation-led, 'smart' growth that requires strategic investments in innovation and mission-oriented projects, as leading innovation economist Mazzucato has argued [158, 159]. However, a common economic narrative regarding market creation positions the private sector as the principal force for innovation, with contributions from the public sector only important in setting the conditions for that private sector activity. Government investments in open data projects and infrastructures can actively shape and create new lucrative markets, while enabling the more equitable and sustainable sharing of the fruits of public research. These types of government investments can spur genuine innovation and create breakthrough technologies, Mazzucato argues.[51]

There are other examples in which public investments in scientific open data infrastructures have generated new and substantial economic returns and business opportunities, as illustrated later in this chapter by the well-known Global Positioning System technology. Open data can also enable enormous savings of public money by facilitating swift responses to public health emergencies. This capacity of open scientific data was powerfully demonstrated during the outbreak of a highly virulent *E. coli*-strain bacterium in Germany in May 2011.

2.5.2 *E coli* epidemic, Germany 2011

In May and June of 2011, almost 4000 people in 16 countries mysteriously fell ill with digestive symptoms. Almost a quarter of them suffered haemolytic uremic syndrome. In many cases, the syndrome led to kidney failure. Of those who were affected, 54 people died. The highly virulent *E. coli*-strain bacterium was found to

[50] *Ibid,* at point 111.
[51] *Ibid.*

be resistant to common antibiotics. There appeared to be no cure and the source of the infections was not known either. These dramatic events brought together scientists from four continents to work on what later became known as the 'world's first open source analysis of a microbial genome' [160].

Researchers from the Beijing Genomics Institute had first analysed the strain, working closely with their colleagues in Hamburg. Three days later a full genomic sequence of the bacterium was published [161]. By enabling free sharing[52] and permanent access to the original results[53] the Chinese microbiologists spurred dynamic international collaboration. Just 1 day later the genome was assembled and within a week over 20 reports were filed on a website dedicated to crowdsourced analysis of the bacterium.[54] The reports were crucial to identifying the strain's virulence, resistance genes and effective treatment. These efforts, along with concentrated measures taken by public authorities and doctors, resulted in the epidemic being averted.

Thanks to open data, the cure for the epidemic became known earlier than the source of the epidemic. Open data revealed that the epidemic was caused by an enteroaggregative *E. coli* strain, not an enterohemorrhagic strain, as originally thought. Open data further revealed that the strain had acquired the genes that produce Shiga toxins present in organic fenugreek sprouts. This hint led to the source of the epidemics being discovered. The agriculture minister of Lower Saxony identified an organic farm in Bienenbuettel that produced a variety of sprouted foods to be a source of the epidemic. The farm was immediately closed.

The value of human lives saved is priceless. The costs associated with the epidemic being averted cannot even be estimated. Every day of waiting would have resulted in more people falling sick and more people dying. The key economic benefits of dealing with public health emergencies like this lies in the swift response. In this case, open data was the key.

2.5.3 The global positioning system

The global positioning system (GPS) is a space-based radio navigation system owned by the US government and operated by the United States Air Force. According to the National Aeronautic Space Agency, GPS originated in the 1950s during the time of the Soviet Union's first Sputnik satellite mission. Scientists in the United States found they could track the satellite by monitoring its radio transmissions and measuring the shifts in those signals, analysing the Doppler effect that an observer experiences as an object moves past. In the mid-1960s, the United States Navy is built on this experience to conduct experiments with satellite navigation for the purpose of tracking US submarines carrying nuclear missiles. The submarines observed the Doppler changes of six satellites that orbited the poles and were able to pinpoint their locations within minutes [162].

When the Department of Defence sought to build a robust, stable, satellite navigation system in the 1970s, it decided to use those satellites to support a navigation system that took on the earlier ideas and experiences of Navy scientists. The result was the launch in 1978 of the first Navigation Satellite Time and Ranging system. Comprising 24 satellites, the system became fully operational in 1993.

---

[52] The original strain was published under the Creative Commons Zero licence.

[53] EHEC Genome with a DOI name [http://datacite.wordpress.com/2011/06/15/ehec-genome-with-a-doi-name/ (15 June 2011)]. The genome data is available at: http://gigadb.org/dataset/100001.

[54] GitHub Inc. is a sharing platform principally used by computer programmers. See the page: ehec-outbreak-crowdsourced/BGI-data-analysis. https://github.com/ehec-outbreak-crowdsourced/BGI-data-analysis/wiki/_pages.

Meanwhile, in 1985, the GPS technology became available to any user as a GPS receiver. That service[55] remains available worldwide to all users with no direct charges.

The GPS technology rapidly became a subject of intensive innovation, especially as it was infused with other applications. Over time, GPS became embedded in virtually every communication device. The economic benefits of GPS accrued to the United States up until 2013 were estimated at about US$56 billion, or 0.3% of national GDP.[56]

From the case studies listed above, it is clear that the development of open technologies and open data infrastructures has an enormous economic potential to harness greater returns from public research. It appears that the benefits reaped from open data and open technologies surpass, by some distance, the economic benefits received from licencing and IP commercialisation of public research[57], especially in the university sector where many patents are 'sleeping patents'—meaning that they remain commercially unexploited, are neither licenced nor used internally, and are not held for purely defensive purposes.[58] The effect is that the patented knowledge cannot be shared but, at the same time, is not generating any economic or other benefits. The result is a net loss associated with the cost of IP protection.

Open data and open technologies can lead to substantial economic benefits, and the value of open data assets and collections will further appreciate over time [166, 167]. For these reasons, scientific open data assets have a far greater potential to generate economic returns from public research than the returns received from commercialisation of public intellectual property.

The public function of research is best upheld in collaborative spaces that are open to all stakeholders—researchers working in the public and private sectors—all around the world. There is also ample evidence showing that open data and open scientific knowledge can effectively spur collaborations with the private sector and lead to the development of new technologies faster.[59]

---

[55] GPS currently provides two levels of service: Standard Positioning Service (SPS) that uses the coarse acquisition (C/A) code on the L1 frequency and Precise Positioning Service (PPS) that uses the P(Y) code on both the L1 and L2 frequencies. Access to the PPS is restricted to the US Armed Forces, US federal agencies, and selected allied armed forces and governments. The SPS is available to any user globally. Source: NASA [162].

[56] Results of a 2015 study commissioned by the National Executive Committee for Space-Based Positioning, Navigation and Timing. See GPS World Staff [163].

[57] According to the OECD, in Australia, Europe, and the United Kingdom, the licencing revenue received from IP hovers around 1% of R&D expenditure and appears to be declining. In the United States, the figure stood at 4% in 2011 and the revenue was also declining. A study by the Brookings Institution found that 84–87% of US universities do not realise enough income to cover the costs of running their technology transfer office. See Valdivia [164].

In the United States, the top 15 universities with the highest income received from intellectual property received only US$1 billion from licencing revenue in 2015. Just over US$400 million of commercial income was received from intellectual property licencing by the top 15 biomedical research institutes. See Hugget [165].

Patents, licencing income, and spin-offs are frequently used indicators to assess an institution's or a country's capabilities to turn public research into innovation. In terms of patent applications filed by universities in the United States, the average annual growth rate fell from 11.8% (2001–2005) to 1.3% (2006–2010), while other public research organisations experienced a negative growth of −1.3% over the latter period, compared with 5.3% growth recorded between 2001 and 2005. Source: Cervantes and Meissner [166].

[58] *Ibid*, 77.

[59] For example, Chesbrough argued that an important element of 'open innovation' is the use of purposive inflows and outflows of knowledge to accelerate internal innovation [12, 168].

Conclusion

The nature, dissemination and use of scientific knowledge are profoundly changing in the context of the digital revolution. Digital technologies have provided the means to collect, process, analyse, store and disseminate vast amounts of data. These technological advances are changing the core processes of science production, with a shift away from the modern science espoused by Thomas Kuhn towards the digital science emerging in collaborative online spaces.

For open science, data has a role that has changed from how it was treated in earlier contexts of science creation. The key among the differences is the principle that data produced in publicly funded research should be openly available. In addition, there is a growing view that the value of data is multiplied when it is shared with a range of stakeholders beyond the research community that initially collected it. In this changing context, as data becomes more accessible, it opens the way to uses that can create new knowledge and fuel innovation and economic growth, thus furthering the aims of a knowledge-based society. Open scientific data aims to encourage, for the first time in history, the participation in science creation, validation and dissemination by both scientists and non-scientists.

As digital formats increasingly become the preferred means for data storage and distribution, computers alone now have the capacity to validate and generate scientific outcomes—a capacity that will grow further with advances in artificial intelligence and quantum computing, along with the development of algorithms that can rapidly process and calculate vast amounts of data to solve problems. Consequently, there is a strengthening argument to the effect that open scientific data challenges established research and science conduct and related communication practices. At the same time, open scientific data promises to beak the monopoly held by researchers over the validation and creation of scientific outcomes. This transformative social role of science is important and needs to be understood in the interplay between evolving technologies, advanced communication, changing culture and increasing education.

However, the calls for sharing of research data in electronic formats came well before modern digital technologies. The World Data Center was established in 1955, and major international scientific data projects emerged in the 1960s. Digital sharing of scientific data builds on these early foundations and took a huge leap forward in 2003 when the Human Genome Project was completed. The year 2003 also loosely marks the emergence of the open access movement, which brought renewed calls for greater availability of scientific data with the adoption of the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities.

The benefits of open data are well covered in the theories of innovation and in economic literature. Clearly, open data in general, and open scientific data in particular, holds an enormous potential to increase the social and economic benefits of public research. The economic benefits are already felt in those fields that adopted open scientific data early—fields such as genomic and geospatial research. As data has rapidly become a commodity in the global economy, scientific data represents a new type of economic asset. There is a decisive competitive advantage for those who know how to use open scientific knowledge. However, the increased demand for scientific knowledge also poses a risk to public science in the form of the increased privatisation of public research. The open science movement counterbalances these developments by placing a renewed emphasis on the broader dissemination and free sharing of scientific outcomes in the public domain.

At the same time, the benefits of open scientific data can only be realised if the infrastructures for open science are developed and if the data is not only openly shared but also gets reused. The reuse aspect is not well covered in the theories of

knowledge society and digital science production. These theories view the release of scientific data into the public domain as sufficient for the economic and social benefits of open data to accrue. This chapter argues that while data sharing is a prerequisite, only data reuse can harness the envisaged returns on investments in open scientific data.

The three parameters identified in this chapter—the changing role of scientific knowledge in society, the possible benefits of scientific data and the necessity to reuse the data to realise the benefits—need to be viewed in relation to one another, Chapter 2 concludes.

In the next chapter, I follow up this thinking by examining the open data policies recently introduced by research funders and the potential of these policies to drive the release and reuse of open scientific data into the future.

Author details

Vera J. Lipton
Zvi Meitar Institute for Legal Implications of Emerging Technologies,
Harry Radzyner Law School, IDC Herzliya, Israel

*Address all correspondence to: vera.lipton@bigpond.com