Jasny *EPJ Data Science* 2013, **2**:1
http://www.epjdatascience.com/content/2/1/1

EPJ Data Science
a SpringerOpen Journal

REVIEW Open Access

Realities of data sharing using the genome wars as case study - an historical perspective and commentary

Barbara R Jasny*

*Correspondence:
brjasny@gmail.com
Science/AAAS, Washington, DC 20005, USA

Abstract

The importance of data sharing has become a mantra within the science research community. However, sharing has not been as easy (or as readily adopted) as advocates have suggested. Questions of privacy, individual scientist's rights to their research, and industry-academia divides have been significant hurdles. This article looks at the history of the debates and problems associated with data access that occurred during the 'human genome wars' and their aftermath as a way to explore some of the challenges facing diverse research communities.

Keywords: data sharing; data access; human genome project

Introduction

In 2009 Tim Berners-Lee, the inventor of the World Wide Web, described the power of linked data, the crucial need to get it 'unlocked from individual silos,' and the imperative to stop people, governments, and enterprises from 'database hugging' [1]. The ability to pool datasets generated by different researchers (or organizations or governments) has been considered crucial to achieving progress in areas as diverse as personalized medicine and sustainability. Those who see data sharing as providing a public good often insist that even the investigators who generate data have an obligation to their funders and the general public to release it - thus superseding their individual rights to explore the data first. As described by Alan Guttmacher *et al.* (2009), 'The model of the investigator owning data has been increasingly replaced by one in which society owns data' [2, p.1].

The reality, however, is that most researchers don't currently share their data voluntarily even when it is made easy [3–5]. Recently researchers from the National Science Foundation-funded DataONE project (5) conducted a survey of more than 1,000 scientists and concluded that 'Barriers to effective data sharing and preservation are deeply rooted in the practices and culture of the research process as well as the researchers themselves.'

The purpose of this article is to examine one process of data sharing in detail in the hope that the problems and successes can inform the system more generally. One of the arenas in which debates over this issue were especially visible and acrimonious concerned the data that represent the foundation of our genetic heritage. It is certainly true that success at delivering the first human genome sequence would have major impacts on the egos and careers of an array of aggressive, colorful people. This includes, but is not limited to

leaders of the private (commercial) and public efforts, - J. Craig Venter on one side and Francis Collins and Sir John Sulston (plus a cast of associates) on the other. However, the primary focus of this article will be on the debates and discussions regarding the value of information and information access.

The so-called 'genome wars' did not occur in isolation, as events related to information access were occurring in a variety of disciplines. This focus is of much more than historical interest. As I will discuss in the concluding section, the realities of data sharing, the challenges of simultaneously fostering research and preserving individual privacy, and debates over the use of commercial data continue to pose challenges for scientists from a variety of disciplines, ranging from genomics to the social sciences.

The human genome

A genome is the sum total of all of our genetic information, encoded in an ordered sequence of 3 billion nucleotides. The information locked in that sequence has enormous symbolic value as our human inheritance. Such data also has potential value as a jumping off point to learn about the genetic pathways involved in human disease and identify potential inroads for diagnosis and therapeutics. There was a lot of money at stake for pharmaceutical and biotechnology industries in getting the first look at the human genome sequence. Money was also under the surface for the academic community as it was crucial to the publicly funded effort not to be rendered redundant in front of Congress, which would fuel the next decades of research.

The competition to present the first complete sequence of a human genome was widely seen as a battle between advocates of free access and advocates of proprietary information, but it was not that simple. At the time, the version published in *Science* magazine [6] in 2001 represented an experiment in providing the general community access to data generated by a private company.

Early history

To set the stage, it is important to go back to the Bayh-Dole Act. In an effort to spur innovation and promote public-private partnerships, Congress in 1980 declared that academic institutions and small businesses could hold on to intellectual property rights for inventions made under federally funded research programs. Universities and research institutions were being encouraged to commercialize discoveries [7]. Later, there would be debates over whether the Act had fostered a harmful anticommons [8] and the National Research Council in 2003 [9, p.23] would conclude that it had 'in some cases, impeded the unrestricted sharing of publication-related data.'

Partly as a result of the Bayh-Dole Act, the National Institutes of Health (NIH) and Venter were at the center of a controversy over publication of genetic information well before the 2001 publication. Venter, working at NIH, had developed a rapid gene-finding approach based on markers called expressed sequence tags (ests). The head of NIH's technology transfer office, Reid Adler, was contacted by an attorney at the biotechnology company Genentech Inc. because of concerns that publication of ests would result in patents being denied to the biotechnology industry 'even if the company had invested in the hard work of reading out the gene's whole sequence, identifying its protein, and figuring out what role it played in the body' [10, p.84]. Adler proposed that NIH patent the ests first; academics could use the ests without cost and companies would be able to license them

for a fee. Bernadine Healy, who was NIH director at the time, supported this idea as well, and in 1991 and 1992 NIH applied for patents on 2,700 of these gene fragments.

The prospect of est patenting by NIH set off loud cries of opposition from the Pharmaceutical Manufacturers Association (PMA), biotechnology trade groups [11], and some in academia. Concern over the possibility that ests could be patented was one factor that led, in 1996, to the 'Bermuda accord,' an international agreement among the major sequencing groups mandating that sequence data be deposited in a public database within 24 hours of generation. This created prior art that could effectively render the sequence material unpatentable. It was also supposed to prevent duplication of efforts and level the playing field so that large sequencing centers would not have a monopoly on the information [12, 13], although, as I will describe, it is arguable whether a truly level playing field exists. Initially there was not an international consensus. The German research ministry had wanted to allow companies 3 months of advance access to data generated with their public funds, and only reversed themselves because of threats by US and British scientists that their German counterparts would be left out of the international community and denied access. Similar discussions occurred in France as well [14].

The beginnings of the genome wars

The Human Genome Project had originally been proposed by Charles DeLisi of the US Department of Energy (DOE) in 1986, and fit in with the DOE goals of understanding human mutation. NIH and the Wellcome Trust in the UK became the lead funders for an international consortium of publicly funded sequencing groups (which will be called the PFG here). The official start date was 1990 and, at a projected cost of $3 billion, the project was expected to take 15 years. The project was divided into mapping and sequencing phases, with the idea that maps of markers ordered along human chromosomes would guide the sequencing strategy. On May 10, 1998 at a meeting of the genome community at Cold Spring Harbor, New York, Michael Hunkapiller (representing Perkin-Elmer Corporation, which was innovating fast sequencing machinery) and Venter announced that they were forming a company (Celera Genomics) to sequence the human genome faster and cheaper than the PFG. They would use the 'shotgun sequencing' approach pioneered by Venter, which skipped the mapping step. It involved breaking the genome up into random pieces and, via algorithms, reassembling the whole thing in one step. The cost was projected to be $200-$250 million and they said they would be finished 4 years earlier than the PFG [10]. Venter announced that he would release data quarterly, which immediately raised the outcry from Robert Waterston (head of the sequencing group at Washington University, St. Louis) 'Quarterly! That's a lot different from overnight.' To which Venter reportedly replied, 'We're a company, Bob. We don't have to release the data at all. But if you think about it, quarterly is a lot closer to nightly than it is to never' [10, p.52].

The announcement at Cold Spring Harbor was the equivalent of stirring up an ant's nest with a stick. The PFG had originally envisioned producing a polished, error-free sequence but now felt pushed to generate a quick-and-dirty product to be competitive with Celera. I heard Phil Green (University of Washington, Seattle) sum up the situation graphically: 'We don't want to be the pooperscoopers for the Venterpillar' and the Welcome Trust announced that it would double the amount of money it was pouring into to the cause.

Venter was clear from the start that he planned to use the PFG data to help Celera's project. Access to public data gave Celera an advantage in assembling the sequence. Francis Collins of the NIH called it a breach of scientific ethics [15]. However, Celera received

surprising support from no less than David Lipman, director of the National Center for Biotechnology Information (NCBI) who said, 'These groups understood when they deposited their data that the whole goal was to make it available without restrictions immediately' [15].

Venter was walking a tightrope between achieving the academic goal of recognition and the commercial goal of making money. Celera's initial business plan was for the data to be available by subscription. The idea was that the raw list of ordered nucleotides that make up DNA needed to be annotated, *i.e.* translated by means of computer programs into the genes and regulatory elements that could then be studied and harvested. The financial return for investors would be from the ability, for a fee, to access user-friendly tools with which to analyze the information. Pharmaceutical companies paid $5-$15 million a year and academic groups such as the Howard Hughes Medical Institute (HHMI) paid $7,500-$15,000 per lab for access to the Celera data [16].

From Celera's point of view, the genome sequence they generated could not just be released because of the concern that competing companies would repackage the data and sell it with their own tools. This concern was justified by actions of companies like Incyte Pharmaceuticals which, as early as May 1996, was selling a version of its database that included publicly available sequence data that had been analyzed with Incyte's proprietary software.

Publishing principles and database protections

The war between the public and private efforts would effectively be over if Venter couldn't publish and Venter couldn't publish if journals determined that his approach to data release violated journalistic policies.

There is a long tradition that a scientific publication must be complete enough so that other scientists can reproduce the work. Before the advent of large datasets, this meant showing the underlying data as part of the journal publication, but this became impractical as datasets grew. *Science* began making explicit statements regarding database deposition in an August 5, 1988 editorial in which Daniel E. Koshland, Jr. (1988) wrote that *Science* and other journals had developed procedures to help ensure that data would be sent to appropriate repositories [17].

For DNA sequence, the repository was the public database GenBank and its mirror sites. While data in GenBank has always been freely accessible, as of September 2012 the website still included the disclaimer that there might be patent, copyright, or other intellectual property encumbrances on the data.

In 1999 there were discussions about whether GenBank could cordon off some part of itself so that the Celera data could be viewable but not downloadable [18]. David Lipman indicated that NCBI did not have the resources to provide legal protection from copying [15] and could not even delay submission of a joint analysis to GenBank by 12 months. At one point, the PFG was willing to offer Celera a year's worth of some form of intellectual property protection as a database provider for data generated as part of a collaboration, but Celera held out for 3-5 years [19]. In 1999, no less proponents of openness than John Sulston and Tim Hubbard of the UK Sanger Center briefly considered such a model because of their concern that restrictions would be placed on reuse of the data [18].

Had Celera been operating in Europe, there would have been legal protections for its sequence database, as the European Union had enacted a database protection directive

in 1996 allowing database creators 15 years of protection for the time, money, and effort they had invested. However Celera was vulnerable in the United States. In 1991 the US Supreme Court made an important decision in *Feist v. Rural Telephone*. Feist incorporated names, addresses, and phone numbers from Rural Telephone's directory into their own directory. Rural Telephone sued for copyright infringement, but the Supreme Court ruled that the information in their compilation was factual, and that facts are not protected by copyright. The reality that Rural Telephone had done the work of creating the database did not entitle it to copyright protection. The database industry responded by seeking help from Congress [19], but no legislation was passed. It was thus clear in the period between 1998 and 2001 that copyright law was not going to solve any of the issues being raised about protection for genomic databases. There was nothing to prevent a competing company from repackaging (with their own annotations) the Celera sequence and selling it.

A battle of media leaks and press announcements ensued, making the project seem less like noble science than a horse race, and splattering both sides with mud. To counter this, President Bill Clinton and Prime Minister Tony Blair issued a joint statement on March 14, 2000 in which they called the sequencing of the genome 'one of the most significant scientific projects of all time' and praised the PFG for its open release of data [20]. The statements were misinterpreted in the press as meaning that the leaders were going to come out against the patenting of genes. The result was a slide in stock prices, with Celera stock falling by 21% [21].

In another effort to shore up the public image of the field, the race to the genome was officially declared a draw. On June 26, 2000 it was announced that a draft sequence of the human genome had been completed by both Celera and the PFG. Pictures of Venter and Collins were taken with President Bill Clinton. One result was the sense that it would be an embarrassment if the two sides didn't publish side-by-side.

What was *Science* thinking?

Floyd Bloom, Editor-in-Chief of *Science* from 1995 until June 2000, focused on the lack of existing database protections. Donald Kennedy took over as Editor-in-Chief in June 2000 and felt that there could be flexibility in approaches to data access.

It was easy in 1999-2000 to find examples where guidelines for data access were being massaged to deal with realities of economics or to promote public-private interaction. One situation that formed a foundation for *Science*'s approach to commercial and academic access to data was the two-tier access existing at the time to the protein sequence archive called SWISSPROT. In 1998, SWISSPROT announced that, because its public funding could not keep up with the growing number of sequences, it would charge a yearly license fee to commercial users.[a]

Another example that was very much in Kennedy's mind was a paper that *Science* had published in 1987 that was based on a large archive of proprietary geophysical exploration data generated by EXXON Corporation [22]. The paper included summaries and not raw data; however, Kennedy felt that the research provided important insights into the history of sea-level change.

On May 25, 2000 Bloom and Kennedy, in an e-mail memorandum to both the PFG and Celera, enunciated *Science*'s position as follows: 'Access to the sequence is an essential pre-requisite for publication - However we believe that the research community is open

to more than one model for database access - Restrictions that provide protection against commercial redistribution of the data can be consonant with publication in *Science.*' (personal communication, Donald Kennedy).

License negotiations

The first draft sent to *Science* by Celera in June, 2000 was not acceptable to the journal. There was a single click-through agreement for academic users, but commercial users and bioinformatics 'customers' would have to pay a fee to see any of the data. Such an agreement would have allowed users to examine the sequence only via browsing at the Celera site and to only use the analytical tool BLAST at the site to search for particular genes of interest. In addition, the first draft did not contain a provision for *Science* to keep an escrow copy, which *Science* considered vital for the archive of record. *Science* kept pushing for better access provisions and drafts were exchanged between Celera and *Science* through the fall and early winter of 2000.

A broad range of researchers (academic and commercial), bioethicists, lawyers, representatives of other journals, and experts in technology transfer from NIH and Howard Hughes Medical Institute (HHMI) were consulted. The ethics of differential access and the meaning of verification of published findings were at the center of many of the discussions. Celera was convinced to change the license terms so that for-profit users could get access to the data if they signed a material transfer agreement (MTA) that committed them to use the data for research purposes and not for commercialization. MTAs are contracts that regulate the transfer of materials (such as cell lines, plasmids, chemicals, or even software) between academic organizations and between academic organizations and industry. Some contained restrictions as to how the material could be used, and how and whether publications or commercial applications based on the material would be allowed. MTAs might require that the researcher receiving the material allow the provider a form of 'droit de Seigneur' in that the provider could review the draft of the researcher's publication or poster for a presentation, usually for 30 to 45 days.

In the agreements negotiated with Celera [23], noncommercial users would not have to sign an MTA; they would only have to execute a click-thru license and could then use the sequence to make discoveries and commercialize them without any reach-through provisions. A scientist working for a commercial organization could only get access to the sequence for purposes of basic research and to verify conclusions presented in the *Science* paper. In theory, that would make the data available to companies. However, even to confirm a result or to do basic research, it could be toxic for a company researcher to look at the data as it might be difficult to prove later that commercial advances had been made independently. Realistically, the only way commercial users would get access to the data was by paying for a subscription or getting a license from Celera.

Science received objections on philosophical grounds that any discrimination among readers was unethical. However, others commented that after facing restrictions in MTAs for clones and other materials they'd prefer to just pay for use of the data and be done with it. In addition, it was not hard for some academics, with their more limited resources, to accept the notion that the companies of the world would have to pay. Another reaction was that it was the lesser of two evils, with the greater evil being that the Celera data would not be released at all. We were told that neither NIH nor the HHMI would block their researchers from signing the agreements.

The public-private divide wasn't the only problem. For the burgeoning field of bioinformatics, which was analyzing and working out how to draw meaning from the raw sequence, verification was not enough; this community needed to be able to repost and recombine data. The ability to include chunks of data or a whole genome, not as a link to someone else's database, but as part of the actual paper was seen as crucial. The bioinformatics community did not believe that the restrictions on their ability to pool data fit a reasonable definition of 'accessible.'

The PFG decided in December 2000 to 'vote with their feet' and publish their paper in *Nature* instead of *Science.* Between December and February, when the publications appeared [6, 24] we received intense reactions - ranging from a hostile letter writing campaign to then National Academy of Sciences president Bruce Alberts' appraisal, 'If successful, this model agreement could greatly increase public access to the large amount of genome sequence data now held secret by private companies' [25].

Aftermath of the human genome publication

Celera and the Celera Genome. There was no question that the value of the Celera sequence, however well annotated, would decrease with time. Venter had said that he would put the data into GenBank in 3 years and that happened in 2005 [26]. Venter left Celera and, like other genomic data-based companies, Celera turned to pharmaceuticals and other avenues of profit [27]. As of September 7, 2012 each of the two human genome publications had received more than 10,000 citations according to Google Scholars.

Rice. In 2002 *Science* published the sequence of the rice genome produced by Syngenta's Torrey Mesa Research Institute, under essentially the same provisions as the human genome paper. Rather than downloading the sequence, academics would be sent a CD with the whole genome once they had signed a letter stating that they were using the sequence for basic research. Kennedy's rationale was that he considered rice to be the most important food crop in the world, but *Science* again received an outpouring of critical letters.

The Cech committee. In response, the National Academies of Science formed the Committee on Responsibilities of Authorship in the Biological Sciences in October 2001, headed by Thomas R. Cech. Their year-long investigation showed the existence of a general consensus about the importance of sharing but implementation varied, leading them to note 'as in many human activities, the devil is in the details' [9, p.35]. They came down hard against the idea of any exceptions to sharing upon publication, creating the acronym UPSIDE (the Uniform Principle for Sharing Integral Data and materials Expeditiously). They concluded that the access via *Science* to the human genome and rice sequence was static rather than the kind of dynamic access that would facilitate further research. However, they also recognized that lack of protection for companies who were generating databases was an issue that could impede the publication process.

Science magazine. The Information to Contributors was changed to explicitly state 'All data necessary to understand, assess, and extend the conclusions of the manuscript must be available to any reader of *Science*' [28]. It took years for the desire on the part of some of the genome community to punish *Science* to fade, but that has mostly disappeared.

Information may be free in heaven, but it's not so easy down here

During the genome wars, researchers who had to contend with the realities of the Bermuda accord were not finding it as straightforward as envisioned. If Venter was trying to 'have

his cake and eat it too' by publishing in *Science* but still protecting corporate interests, some of the large scale sequencers were trying to do the same. They were advocating the high moral ground of immediate release in public while working in private to secure their publication advantage. Partial genomes, posted before the genome sequence was actually finished, were treasure-troves to scientists waiting to parse raw sequence into genes and regulatory elements and answer specific questions of biological and medical interest. As early as 1999, this tension was clear in disputes between major sequencing centers working on the genomes of agents that cause sleeping sickness or malaria and individual researchers eager to annotate genomes and use them in medical research [29].

The sequencers developed the culture of the gentlemen's agreement, which basically said that although anyone could look at and do research with the data, they could not publish until certain conditions were met. Sometimes this meant getting permission from the original sequencing group. In other cases researchers would have to refrain from publishing whole genome analyses until the originating group had published. There came to be a clear distinction between release of data to public access and the time when it was really free of encumbrances [30].

This did, however, leave a lot of grey areas. For example, the editor of *Immunology Today* asked if the sequencers had rights to sit on the data forever or if there was to be a time limit (such as a year) beyond which publication could be open to anyone [31]. Rowen and Hood [32] replied, saying that the time limit would have to start once all the data had been gathered to be fair to data producers, which could take years. Richard Hyman of the Stanford Genome Technology Center went further, responding that in the absence of written consent a third party publishing an analysis would be committing plagiarism and fraud [33].

The Fort Lauderdale meeting of January 2003 attempted to sort out some of the issues [34]. The 40 participants reaffirmed the rapid release principles but presented the case that it was in everyone's interest to nurture data producers who were involved in community resource projects. Funding agencies were called on to provide support for curation, maintenance, and distribution of the data and resources to the producers so they could analyze their data. Resource producers were asked ('when feasible') to generate and publish a project description or marker paper at the beginning of the community resource project, that would act as a citable reference and lay out the analyses for which they expected to have priority. Resource users were told that they should respect the legitimate rights of the producers by citing the marker paper, and respect the need of the producers to publish. No time limit for the prerogative of sequence producers to publish first was set. And, in a statement that sent a chill through the hearts of a number of journal editors, the Wellcome Trust report of the meeting specifically said it was the role of the journals to ensure that the system worked fairly for everyone.

This has left the journals in the middle. Sometimes publications would be moving along smoothly, with peace reigning among the analysts and junior members of the sequencing teams as co-authors, only for the journal to discover that the heads of the sequencing centers had not been consulted and were not happy with the deals that had been made. Nor is this kind of dispute old history. In August of 2009, a paper was published online at the *Proceedings of the National Academy of Science (PNAS)* that dealt with a gene associated with substance dependence in women of European descent. One of the authors had signed a Data Use Certification in which he agreed not to submit publications based on the Study

of the Addiction, Genetics, and Environment dataset until September 23, 2009 but the dataset was used in the PNAS paper. The authors had to retract their paper on September 9, 2009 [35].

The National Human Genome Research Institute (NHGRI) established a 9-month moratorium on publication by third parties of data from the Encyclopedia of DNA Elements (ENCODE). An update to the policies was issued in November of 2009, extending that moratorium to 'all forms of public disclosure, including meeting abstracts, oral presentations, and formal electronic submissions to publicly accessible sites (*e.g.*, public websites, web blogs)' [36].

The Alzheimer's Disease Neuroimaging Initiative (ADNI) is an example of a public-private partnership that has been notably successful in efforts to collaborate and to have precompetitive data generated by a company made available to academic researchers. However, access to the data generated requires researchers to sign a data use agreement in which they affirm not to redistribute the data and to cite the ADNI in a manner deemed appropriate by an ADNI Data and Publications Committee [37].

The genome community has recently bumped into another problem that could affect the future of efforts to reap the fruits of the human genome project. To understand human variation, especially with regard to inherited disease, it is necessary to accumulate not just one but many genome sequences. To protect the privacy of participants in research studies, efforts have been made to deidentify the information or to present only data from pooled samples. However, the community has been shocked to learn that these approaches don't suffice - by algorithmic means or by combining anonymized sequence data with other freely available information it has been possible to reidentify certain individuals [38, 39]. Concern about this situation forced NIH and the Wellcome Trust to block public access to certain information in the databases dbGAP (The Database of Genotypes and Phenotypes) and CGEMS (a site for cancer genetics work) [40]. Certainly there are laws (*e.g.* the Health Insurance Portability and Accountability Act of 1996 (HIPAA) and the Genetic Information Nondiscrimination Act of 2008 (GINA)) that have been passed in order to protect privacy of medical data and to prevent discrimination in health insurance and employment on the basis of DNA information. However, informed consent is made more difficult by the impossibility of knowing the possible uses to which research data may be put in the future or the effects of aggregation of different kinds of personal data on privacy. Without volunteers allowing their information to be used, medical research studies that could benefit society as a whole will be impossible. Possible resolutions, such as controlled access to data that might otherwise be more freely shared, await future resolution.

Conclusions: a bigger picture

While the focus has been on the history of the 'genome wars', they did not occur in isolation, as discussions regarding information access have been occurring simultaneously in a variety of disciplines (Figure 1). Two forces are currently impacting the research community - the need to protect individual privacy regarding information and the push towards open access to data - and the outcomes are not yet clear. Data available from such sources as mobile phone calls, social media information, consumer information collected by companies, and government data have been described as revolutionizing the social sciences to an extent comparable to the effect that the microscope had on the biological sciences

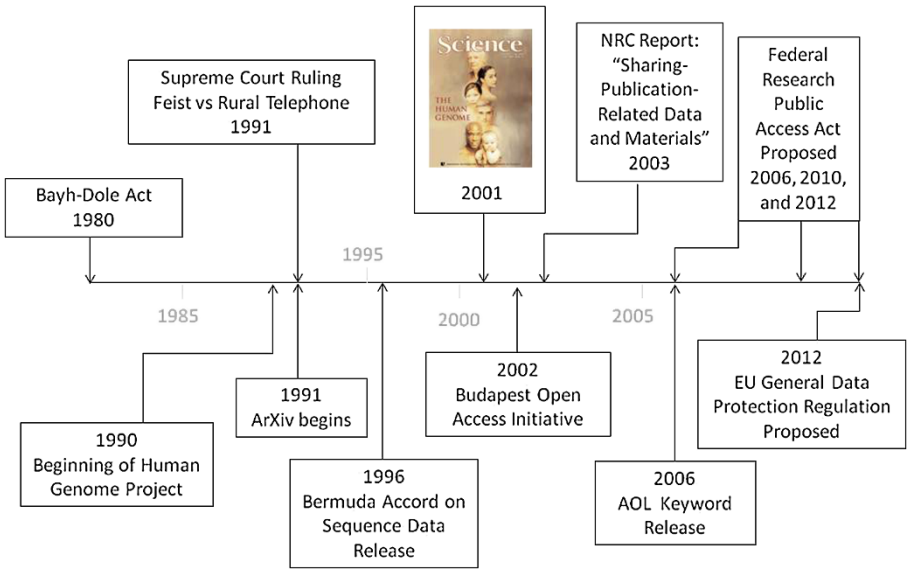

Figure 1 Key events in the publication of a draft sequence of the human genome against a background of information access history.

[41]. However such data individually (and, more importantly, in combination) could render current notions of privacy obsolete.

Large proprietary datasets have been made available to particular researchers because of the relationships or contractual agreements they have forged with the data-generating companies, which raises questions for reproducibility and review for the social science community [42]. For example, Wang *et al*. in 2009 [43] published research relating to the way that mobile viruses can spread, which has implications for the telecommunications infrastructure. Their findings were based on the anonymized billing records provided by 6.2 million mobile phone subscribers. As the privacy of the records were protected by law, the authors noted in the Supplemental Online Material that they would provide further information on request and told *Science* that as long as the researchers were willing to observe the same privacy, technological, security, and legal limitations that they were subject to at the time of the request, they would be glad to facilitate data access at their center (personal communication, A.-L. Barabasi). The motivations for some of these companies to release data to investigators are to learn more about their own clients and operations or to forge deals in which the researchers look at questions of interest to them. However, commentators have complained that such 'private' data threaten the capacity for independent replication on which science is based [44]. (As Bernardo Huberman said 'if an independent set of data fails to validate results derived from privately owned data, how do we know whether it is because those data are not universal or because the authors made a mistake?' [45].) Whether privately arranged access to data that form the basis of scientific publications or release only of aggregated data that protect company or individual privacy will continue remains to be seen. Certainly further investment in technology that can ensure control over the anonymization of data is warranted.

Nor is it easy for private companies to release such data without strings attached. Understanding how people search for information is an active area of research in the social sciences. In 2006 researchers employed by America Online, Inc. (AOL) published log files

encompassing 36,389,567 searches done by users of AOL's proprietary client software on the Internet. This was done without the consent of the individuals involved and, even though screen names had been removed, New York Times reporters and others rapidly showed that it was possible to infer the identities of some of the searchers from the information that was released [46]. Even with these problems, some scientists said that the data release was a service to the research community [47]. AOL was widely criticized in the blogosphere and in mainstream media, some individuals at AOL lost their jobs, and the company was the target of a class-action lawsuit.

There is a long history of efforts to protect individual privacy in the United States and Europe covering a wide range of human activity. Some recent events should be noted as they could affect researchers' ability to gather data. In the European Union, a General Data Protection Regulation was proposed in January of 2012 and is expected to take effect by 2016 [48]. Two provisions could be especially challenging for the research community. The regulation stipulates that data can only be saved for as long as a need can be demonstrated. It also codifies a 'right to be forgotten' - that any individual has the right to have his/her data removed from a database at any time. The European Commission has stated that personal data 'is any information relating to an individual, whether it relates to his or her private, professional or public life. It can be anything from a name, a photo, an email address, bank details, your posts on social networking websites, your medical information, or your computer's IP address' [48]. In the United States, the Federal Trade Commission recently released recommendations for protecting consumer privacy, which included provisions for controls over how much data companies can collect on individuals and how long they can retain it as well as a recommendation that companies establish a 'do-not-track' mechanism for consumers who do not wish to have their information gathered [49]. There need to be consistent, transparent regulations that will safeguard the public but allow research to move forward.

While concerns about privacy could restrict data dissemination, another force is acting to promote access. The idea that sequence information should be freely available was a reflection of a much broader effort by diverse parts of the scientific community to make access to the results of scientific research faster and cost-free. The open source movement started between the 1970s and 1980s with the goal of creating, developing, and disseminating free computer operating systems that would break the hold of corporate entities [50]. While journals that are freely available to readers began to appear by the early 1980s [51] the open access movement gained momentum at about the same time as the human genome project was officially starting. In 1991, a year after the official start to the human genome project, Paul Ginsparg created ArXiv, a pre-publication server whose purpose was to facilitate rapid dissemination of scientific information, without subscription fees, page charges, or peer review [52]. It is currently possible to access more than 800,000 e-prints on ArXiv in physics, mathematics, computer science, quantitative biology, quantitative finance, and statistics.[b] The National Library of Medicine formed PubMedCentral in 2000 in order to have a free repository for NIH-funded research. Initially deposition was voluntary, but Congressional action (the Consolidated Appropriations Act of 2008 (H.R. 2764)) made deposition mandatory with release no later than a year post-publication, after statistics showed poor compliance by the fundees.

At roughly the same time as the report of the Cech committee came down hard on the idea of restrictions to data access in published papers two major manifestos of the open

access movement, the Budapest open access initiative [53] and the Berlin declaration [54] appeared. Certainly some of the statements of groups and individuals who were most vociferous in their opposition to *Science*'s efforts to look for alternatives in publishing data reflect the philosophy of the open access movement. The Wellcome Trust, which has been a prime mover in supporting open access declared recently that 'Our support for open access publishing was a natural progression of our involvement in the international Human Genome Project during the 1990s and early 2000s, where the decision to place the human genetic sequence in the public domain immediately as it was generated helped to ensure this key research resource could be used by scientists the world over' [55].

There have been several attempts, via proposals in 2006, 2010, and 2012 to legislate a Federal Research Public Access Act to require that research funded by 11 federal agencies be made freely and publicly available in government sponsored repositories within 6 months of publication. While the advantages and disadvantages for researchers, funding agencies, and publishers are still being debated, the open access movement has a great deal of energy and backers willing to finance it (at least in the short term).

Does publication, whether in a repository or a journal, mean that enough information is released to form a solid foundation for future research? Despite high-minded principles, the published literature reflects the fact that many researchers do not, left to their own devices, rush to share data. Even when the *Journal of Biostatistics* offered to give formal recognition to authors who provided enough data and methods in their papers to allow an editor to replicate the findings, only a small percent complied [56]. Although the extent of sharing varies from field to field, common reasons given for withholding data are similar: it is too much work, it removes the competitive advantage from the scientists who generated the data and who require publications for their careers, or the raw data was received under confidentiality agreements. The Dataverse Network is a repository for social science data that allows depositors to note concerns and restrictions. In a 2010 survey of the conditions for use posted by more than 30,000 users, Stodden [57] found the most common were 'maintaining subject confidentiality, preventing further sharing, making a specific citation form a condition of use, restricting access by commercial or profit-making entities, and restricting use to a specific community, such as that of the researcher's home institution.'

Certainly the granting and funding and tenure cultures need to enforce good behavior, which people have been saying for years. NIH now mandates that provisions for data sharing be included in research applications for $500,000 or more of direct costs in any single year and several other agencies have similar provisions. The NSF states that to apply for a grant as of January 18, 2011 'All proposals must describe plans for data management and sharing of the products of research, or assert the absence of the need for such plans' [58]. The two-page data management plan submitted as part of the application may include 'policies and provisions for re-use, re-distribution, and the production of derivatives' and 'plans for archiving data, samples, and other research products, and for preservation of access to them.' Although the NSF-wide mandate takes precedence, there are variations within the directorates [59]. For example, the NSF Division of Earth Sciences allows up to 2 years of exclusive data use for selected principal investigators. The directorate for Social, Behavioral, and Economic Sciences allows for the possibility of ethical and legal restrictions regarding access to non-aggregated data.

Community service, whether through generation of shared data or sharing of knowledge and communication with the public needs to be formally recognized. A Data and

Informatics Working Group has recommended that NIH provide incentives for data sharing by providing information on the number of times datasets in its central repository are accessed or downloaded [60].

Even when academic communities are willing to share data, public repositories do not always exist and those that do are under siege in unstable economic environments. It can be easy to drum up funds to create databases, but not so easy to find federal or other moneys to sustain them. Data repositories need continuing support.

The momentum in academia is clearly that releasing and not hoarding data is a virtue. However, the history and examples cited in this paper show that while data sharing may become second nature, it is not an easy, seamless process and is not happening without challenges and compromises.

Competing interests
The author declares that she has no competing interests.

Additional information
This review represents the opinion of the author only and is not an official statement for *Science* magazine or the American Association for the Advancement of Science.

Author's information
The author has been on the staff of *Science* magazine for 27 years and was the manuscript editor handling the human genome paper published by *Science* and described in this review.

Acknowledgements
I would like to thank Don Kennedy for his support and unwavering sense of humor, and for comments on this manuscript.

Endnotes
[a] Important announcement concerning SWISS-PROT www.bio.net/bionet/mm/bionews/1998-July/004469.html.
[b] www.arXiv.org.

Received: 27 December 2012 Accepted: 7 February 2013 Published: 12 February 2013

References
1. Berners-Lee T (2009) Tim Berners-Lee on the next web. www.ted.com/talks/tim_berners_lee_on_the_next_web.html. Accessed 23 Dec 2012
2. Guttmacher AE, Nabel EG, Collins FS (2009) Why data-sharing policies matter. Proc Natl Acad Sci USA 106:16894
3. Louis K, Jones L, Campbell E (2002) Sharing in science. Am Sci 90:304-309
4. Savage CJ, Vickers AJ (2009) Empirical study of data sharing by authors publishing in PLoS journals. PLoS ONE 4(9):e7078
5. Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, Read E, Manoff M, Frame M (2011) Data sharing by scientists: practices and perceptions. PLoS ONE 6(6):e21101
6. Venter JC et al (2001) The sequence of the human genome. Science 291:1304-1351
7. Healy B (1997) BRCA genes - bookmaking, fortunetelling, and medical care. N Engl J Med 337:788-789
8. Heller M, Eisenberg R (1985) Can patents deter innovation: the anticommons in biomedical research. Science 280:698-701
9. National Research Council (2003) Sharing publication-related data and materials: responsibilities of authorship in the life science. National Academies Press, Washington
10. Shreeve J (2004) The genome war: how Craig Venter tried to capture the code of life and save the world. Ballantine Books, New York
11. Eisenberg RS (1993) Technology transfer and the genome project: problems with patenting research pools. Risk Health Saf Environ 5:163-174
12. Dickson D (1996) NIH seeks rapid sequence release. Nature 380:279
13. Bentley DR (1996) Genomic sequence information should be released immediately and freely in the public domain. Science 274:533-534
14. Abbott A (1997) Germany rejects genome data isolation. Nature 387:536
15. Wade N (2000) Rivals on offensive as they near wire in genome race. New York Times, 7 May 2000. http://www.nytimes.com/2000/05/07/us/rivals-on-offensive-as-they-near-wire-in-genome-race.html?pagewanted=all&src=pm. Accessed 23 Dec 2012
16. Service R (2001) Can data banks tally profits? Science 291:1203
17. Koshland DE Jr (1988) The price of progress. Science 241:637
18. Sulston J, Ferry G (2002) The common thread. Joseph Henry Press, New York
19. Pike GH (2003) Database protection legislation introduced in Congress. Infotoday. http://newsbreaks.infotoday.com/nbreader.asp?ArticleID=16598. Posted 20 Oct 2003. Accessed 1 Jan 2012
20. White House Office of the Press Secretary (2000) Joint statement by president Clinton and prime minister Tony Blair of the UK, 14 March 2000. http://clinton4.nara.gov/WH/EOP/OSTP/html/00314.html

21. Schehr R, Fox J (2000) Human genome bombshell. Nat Biotechnol 18:365
22. Haq BU, Hardenbol J, Vail PR (1987) Chronology of fluctuating sea levels since the Triassic (250 million years ago to present). Science 235:1156-1167
23. Accessing the Celera human genome sequence data. http://www.sciencemag.org/site/feature/data/announcement/gsp.xhtml. Accessed 23 Dec 2012
24. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860-921
25. Chemistry and Industry (2000) Celera makes genome data freely available. http://findarticles.com/p/articles/mi_hb5255/is_2000_Dec_18/ai_n28808580/
26. Kaiser J (2005) Celera to end subscriptions and give data to public GenBank. Science 308:5723
27. Kling J (2005) Where the future went. EMBO Rep 6:1012-1014
28. General information for authors. Science. http://www.sciencemag.org/site/feature/contribinfo/prep/gen_info.xhtml#dataavail. Accessed 3 Dec 2012
29. Macilwain C (2000) Biologists challenge sequencers on parasite genome publication. Nature 405:601-602
30. Contreras JL (2010) Prepublication data release, latency, and genome commons. Science 329:393-394
31. Bell E (2000) Publication rights for sequence data producers. Science 290:1696-1698
32. Rowen L, Hood L (2000) Response. Science 290:1696-1698
33. Hyman R (2001) Sequence data: posted vs. published. Science 291:827
34. Wellcome Trust (2003) Sharing data from large-scale biological research projects: a system of tripartite responsibility. Report of a meeting organized by the Wellcome Trust and held on 14-15 January 2003 at Fort Lauderdale, USA. http://www.genome.gov/Pages/Research/WellcomeReport0303.pdf. Accessed 23 Dec 2012
35. Schekman R (2009) PNAS takes action regarding breach of NIH embargo policy on a PNAS paper. Proc Natl Acad Sci USA 106:16893
36. ENCODE consortium data release policy summary. http://genome.ucsc.edu/ENCODE/terms.html. Accessed 8 Apr 2012
37. Alzheimer's Disease Neuroimaging Initiative (ADNI) data use agreement. http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Data_Use_Agreement.pdf. Accessed 8 Apr 2012
38. Gitschier J (2009) Inferential genotyping of Y chromosomes in latter-day saints founders and comparison to Utah samples in the HapMap project. Am J Hum Genet 84:251-258
39. Homer N et al (2008) Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet 4(8):e1000167. doi:10.1371/journal.pgen.1000167
40. P^3G Consortium, Church G, Heeney C, Hawkins N, de Vries J et al (2009) Public access to genome-wide data: five views on balancing research with privacy and protection. PLoS Genet 5(10):e1000665. doi:10.1371/journal.pgen.1000665
41. King G (2012) Ensuring the data-rich future of the social sciences. Science 331:719-721
42. Lazer D, Pentland A, Adamic L, Aral S, Barabási A-L, Brewer D, Christakis N, Contractor N, Fowler J, Gutmann M, Jebara T, King G, Macy M, Roy D, Van Alstyne M (2009) Life in the network: the coming age of computational social science. Science 323:721-723
43. Wang P, González MC, Hidalgo CA, Barabási A-L (2009) Understanding the spreading patterns of mobile phone viruses. Science 324:1071-1076
44. Markoff J (2012) Troves of personal data, forbidden to researchers. http://www.nytimes.com/2012/05/22/science/big-data-troves-stay-forbidden-to-social-scientists.html. Accessed 23 Dec 2012
45. Huberman B (2012) Sociology of science: big data deserve a bigger audience. Nature 482:308
46. Barbaro M, Zeller T Jr (2006) A face is exposed for AOL searcher No. 4417749. http://www.nytimes.com/2006/08/09/technology/09aol.html?pagewanted=all. Accessed 3 Dec 2012
47. Wiggins RW (2006) AOL is caught in its own long tail. http://newsbreaks.infotoday.com/nbreader.asp?ArticleID=17374. Accessed 3 Dec 2012
48. Europa press release (2012) Commission proposes a comprehensive reform of data protection rules to increase users' control of their data and to cut costs for businesses. http://europa.eu/rapid/press-release_IP-12-46_en.htm. Accessed 23 Dec 2012
49. Federal Trade Commission press release (2012) FTC issues final Commission report on protecting consumer privacy, 26 March 2012. http://www.ftc.gov/opa/2012/03/privacyframework.shtm. Accessed 23 Dec 2012
50. Wayner P (2000) Free for all: how Linux and the free software movement undercut the high-tech giants. Harper Collins, New York. http://www.jus.uio.no/sisu/free_for_all.peter_wayner/portrait.pdf. Accessed 23 Dec 2012
51. Suber P (2009) Timeline of the open access movement. www.earlham.edu/~peters/fos/timeline.htm. Accessed 23 Dec 2012
52. Ginsparg P (2001) Creating a global knowledge network. Invited contribution for conference held at UNESCO HQ, Paris, 19-23 Feb 2001, second joint ICSU press - UNESCO expert conference on electronic publishing in science, during session responses from the scientific community, Tue 20 Feb 2001. http://people.ccmr.cornell.edu/~ginsparg/blurb/pg01unesco.html. Accessed 21 Dec 2012
53. Budapest open access initiative. http://www.opensocietyfoundations.org/openaccess/read. Accessed 23 Dec 2012
54. Berlin initiative. http://oa.mpg.de/files/2010/04/berlin_declaration.pdf. Accessed 23 Dec 2012
55. Carr D, Kiley R (2012) Open access to science helps us all. http://www.newstatesman.com/blogs/economics/2012/04/open-access-science-helps-us-all. Accessed 23 Dec 2012
56. Peng RD (2011) Reproducible research in computational science. Science 334:1226-1227
57. Stodden V (2010) Data sharing in social science repositories: facilitating reproducible computational research. http://www.stanford.edu/~vcs/papers/nips2010Stodden12062010.pdf. Accessed 23 Dec 2012
58. National Science Foundation (2011) Grant proposal guide. http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp#dmp. Accessed 23 Dec 2012

59. University of Michigan Library (2012) NSF data management plans. http://www.lib.umich.edu/research-data-management-and-publishing-support/nsf-data-management-plans#directorate_guide. Accessed 3 Dec 2012
60. Data and Informatics Working Group (2012) Draft report to the Advisory Committee to the Director, 15 June 2012. acd.od.nih.gov/Data%20and%20Informatics%20Working%20Group%20Report.PDF. Accessed 23 Dec 2012

doi:10.1140/epjds13
Cite this article as: Jasny: Realities of data sharing using the genome wars as case study - an historical perspective and commentary. *EPJ Data Science* 2013 **2**:1.