Published in 2010, The Commodification of Academic Research, Radder H (ed), University of Pittsburgh Press.

Chapter 7

The Commodification of Knowledge Exchange: Governing the Circulation of Biological Data

Sabina Leonelli

ESRC Centre for Genomics in Society

University of Exeter, UK

1. Introduction

Philosophers of science tend to focus their attention on the conditions under which scientific knowledge is produced and applied. This chapter considers instead the conditions under which knowledge is *exchanged* in science, with particular attention to the boom in bioinformatic resources characterising contemporary biology and medicine. I show how the ongoing commodification of the life sciences affects the ways in which data are circulated across research contexts. The necessity for scientists to develop ways to communicate with each other and build on each other's work constitutes a powerful argument against at least some forms of privatisation of data for commercial purposes.

Science exists in its current form thanks largely to the modes of open communication and collaboration elaborated by scientists and their patrons (be they monarchs, churches, states or private institutions) throughout the centuries. As 'big science' research blossoms and expands[1], the traditional modes through which scientific knowledge is shared are replaced by digital communication technologies, such as databases available through the internet, that can cope with the increasing amounts and complexity of the data being exchanged, as well as with the

[1] Scientific research, especially in biology, is increasingly financed and structured around large projects involving overt collaboration and sharing of resources among various institutions. The projects are typically interdisciplinary and, given the specificity of the topics at hand, they include researchers based at different locations, often widely distant from each other (as in the case of American-Japanese collaborations). Fuller (2000) reviews some of the issues involved in the governance of big science.

uncertainty about the value of some types of data as evidence.[2] The regulation of data circulation across geographical locations and disciplines is in the hands of the private and public sponsors of these databases. My analysis focuses on the contrast between the strategies and values hitherto supported by the public and private sectors in governing data circulation. Both sectors have strong reasons to welcome the commodification of biology – more often referred to as 'translation'– as a desirable development. However, they maintain different perspectives on the procedures best suited to achieving a commodified science. Ultimately, public institutions favour the development of tools for making data travel efficiently across the multifaceted community of life scientists, thus fostering the advancement of biological research. By contrast, the values endorsed by the private sector have hitherto proved harmful to the open exchange of knowledge that is vital to the development of future research. Science can only be enriched by the R&D efforts of private sponsors if data produced in that context are made accessible to any biologist that might need to consult them – a reality that biotech and pharmaceutical companies are slowly coming to terms with, but are not yet acting upon.

The structure of the chapter is as follows. I start by highlighting the importance of disseminating data in biology at a time when biological research is characterised by the massive production of data of various types. After introducing the field of bioinformatics and its role in creating tools to store and diffuse data, I consider the contrast between the regulatory policies for

[2] Especially in the case of genomics, it is impossible to determine the value of data as evidence for future discoveries: the more it is known about the complex regulatory role played by the genome, the more it will be possible to link specific genes to traits at other levels of organisation of organisms.

data circulation that are supported by private and public sponsors of databases, such as the corporate giant Monsanto on one hand and the National Science Foundation on the other. I focus particularly on the regulatory tools characterising the public governance of data exchange. In this context, regulation is geared towards what I call 'resource-driven competition': competition is used as a mechanism to create resources through which research methods and procedures can be improved. By contrast, private sponsors are driven by the need to obtain profitable products in the quickest and least collaborative way. Their management of data exchange, which I refer to as 'product-driven competition', is geared towards the fast-track creation of new entities or processes by any means available. This instrumentalist approach is context-specific and short-term, and as a consequence there is no significant investment in tools or techniques that would enhance the usability of data *in the long run*.

With this analysis in mind, I consider the three stages through which data are shared: (1) *disclosure* by scientists who have produced the data; (2) *circulation* through digital databases; and (3) *retrieval* from databases by scientists seeking information relevant to their own research purposes. I discuss how each of these stages is affected by the private and public regulatory approaches to knowledge exchange. I conclude that the values and methodological criteria imposed by privately sponsored research have a disruptive impact on all three stages of data circulation. In the long term, the resulting inability for researchers to build on each other's work could be damaging to both science and society.

2. Disseminating Data in Biomedical Research

Even a committed Kuhnian will find it hard to deny that science is, at its heart, a cumulative process. This is particularly true when we focus not on the concepts and theories that scientists produce and sometimes discard, but on the results that they achieve in the course of their experiments. I am talking about *data*, that ultimate mark of the measurements undertaken in (and often also outside) the laboratory to document features and attributes of a natural process or entity. Bogen and Woodward have pointed to the relative independence of data production from claims about phenomena. As they put it, 'we need to distinguish what theories explain (phenomena or facts about phenomena) from what is uncontroversially observable (data)' (1988, 314). In biology, typical examples of data are the measured positions of gene markers on a chromosome (figure 1) and the scattered colours indicating gene expression levels in a microarray cluster (figure 2).

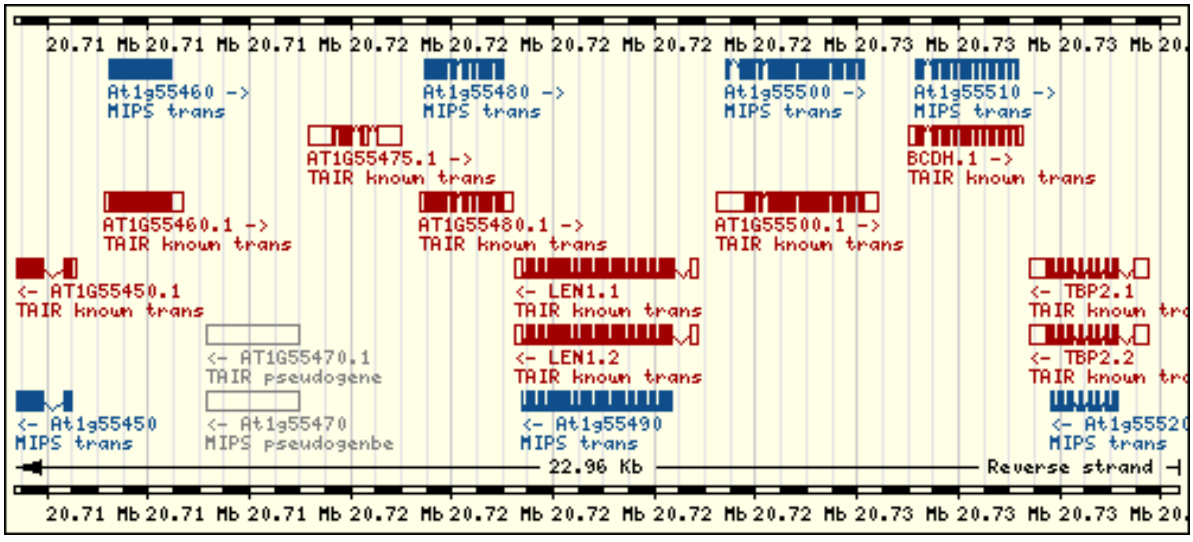

Figure 1. The red, blue and gray marks indicate the position of gene markers on a chromosome (represented by the dashed black lines at the top and bottom margins of the image) as detected

by various investigators (the data of each contributing research group is marked by a different colour). Courtesy of the Munich Information Centre for Protein Sequence.

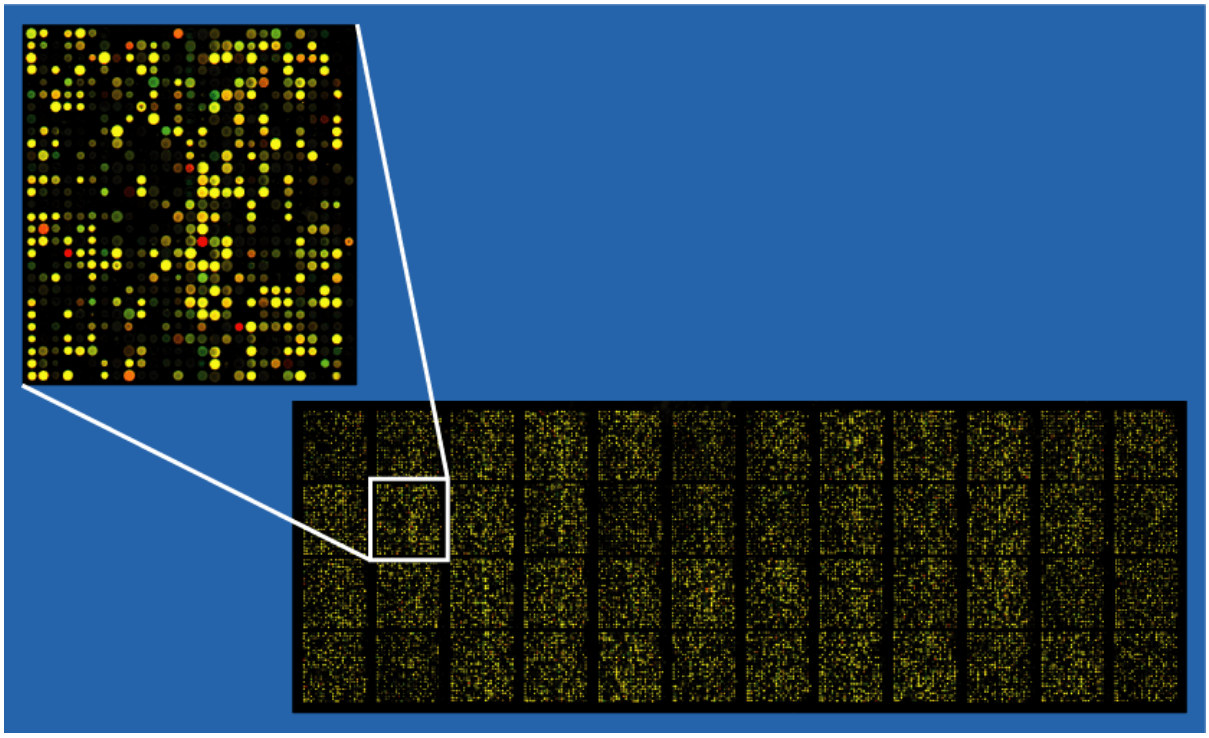

Figure 2. The coloured dots visible in the enlarged section of this microarray cluster represent the expression levels of specific genes in a particular region of a chromosome. Downloaded from the Internet, June 2007.

My epistemological starting point here is the Duhemian intuition underlying Bogen and Woodward's view: data can be used as evidence for a variety of scientific claims, depending on a

scientist's theoretical framework, expertise, commitments and goals. For example, a geneticist working on fruit-fly metabolism can use measurements of the level of expression of specific genes in particular conditions (as in figure 2) to inform claims such as 'gene cluster X is expressed as an enzyme affecting the metabolic cycle of *Drosophila melanogaster*'. Bogen and Woodward focus their discussion on the use of data as evidence for claims about phenomena. They stress the locality of data, that is, the extent to which they are idiosyncratic products of a specific experimental setting at a particular time.[3] While respecting the idea that the experimental context in which data are produced is crucial to their interpretation *as evidence for a new claim* (a point to which I will come back several times below), I wish to emphasise a different property of data that emerges when data are circulated across research contexts. This property is the relative independence of data from specific theoretical or even experimental frameworks and it manifests itself in the context of data circulation, rather than data production or use.

When researchers pass their data to one another, data are taken to speak for themselves. The results of measurements and observations are relied upon as incontrovertible facts, independent of their 'local' origins. The quality and reliability of data, and thus the conditions under which they were produced, are critically scrutinised and possibly disputed only when data have already been appropriated by a new research context: that is, when they are used as evidence for new claims about phenomena. When data travel across scientific communities, it is their neutral value as 'records' of phenomena that counts (and that makes them travel widely, so

[3] 'The characteristics of [data] are heavily dependent on the peculiarities of the particular experimental design, detection device, or data-gathering procedures an investigator employs' (Bogen and Woodward 1988, 317).

to speak[4]). In that context, data are everything but local. They can be and indeed are successfully transferred across different research contexts in biology. Indeed, experimental biologists tend to trust data more than they trust theories and models, and are, as a consequence, deeply concerned with finding ways to facilitate data circulation across disciplinary, institutional and geographical boundaries.

There is no predicting the extent to which each available dataset might contribute to understanding the complex structure of living organisms. It is therefore of paramount importance that existing data can be put to as many uses in as many contexts as researchers deem necessary. Contemporary biologists are gathering massive amounts of data about organisms (including data about all their 'omics': genomics, metabolomics, proteomics, transcriptomics etc.). This is done through increasingly sophisticated instruments and techniques, such as shot-gun sequencing for genomics, which allows the whole sequence of a relatively complex organism to be compiled in a matter or weeks; or microarray experiments, collecting hundreds of thousands data points documenting gene expression levels in a specific cell culture (as in figure 2). Further, the number of organisms studied to this level of detail is getting larger by the day. This richness of data is both the strength and the curse of contemporary life science. It is a strength insofar as it promises to inform hitherto unthinkable levels of understanding and control over living organisms. Biologists are succeeding in producing genetically engineered modifications of plants and

[4] Indeed, it can be argued that data travel across communities precisely thanks to this temporary detachment from information about the local context in which they were produced. See Leonelli (2009a) and my discussion of the procedures through which data are standardised within publicly sponsored databases, below.

animals at an astonishing rate; further, attempts to construct in silico organisms from scratch are under way and no longer look like the material of loony science-fiction. Yet, these developments are only possible if biologists can take advantage of the ocean of data produced by the thousands of laboratories involved. This is where the curse emerges, for assembling tools and procedures through which all produced data can be stored and easily retrieved proves a daunting task.

For a start, there are considerable technical challenges. Consider the sheer size of the datasets being produced by researchers all over the globe about almost any aspect of the biology of organisms – billions of new data points every year. Further, there is the high variability in data types and formats, which makes it difficult to group them all together. And last but not least, there is the high degree of disunification characterising biology as a whole. Philosophers of science have long been aware that biology is fragmented in countless subdisciplines and epistemic cultures, each of which endorses its own, project-specific combination of instruments, models and background knowledge.[5] All these communities study the same small set of organisms, commonly referred to as 'model organisms'[6], so as to understand their complex biology. At the same time, what each community means by 'understanding' depends on the specificity of its research interests and resources. Each group or individual in biology wants to be able to search other researchers' datasets in order to quickly discover whether data produced by

[5] Dupré (1993) and Mitchell (2003) are among the many philosophical contributions to the discussion of disunity in biology. For a discussion of the notion of epistemic culture, see Knorr Cetina (1999).

[6] For philosophical analyses of model organism research in biology, see Ankeny (2007) and Leonelli (2007).

others can be relevant to their own project.

This situation makes the search for tools to circulate data into the holy grail of contemporary biology. Researchers need efficient ways to exchange datasets with biologists working in other research contexts, without however losing time and focus on their own specific project and goals. Bioinformatics is the biological field devoted to tackling this need. The idea is to exploit developments in information and communication technologies so as to build *databases* 'smart' enough as to store data and transmit them through the internet to whoever might need them. This strategy has hitherto been extremely successful, with databases steadily increasing their size, numbers and popularity, and funding for bioinformatics acquiring priority over the development of other biological resources. Some of the most successful databases host data about specific model organisms. The Arabidopsis Information Resource (TAIR), for instance, brings thousands of different types of data gathered on the flowering plant *Arabidopsis thaliana* under the same virtual roof and facilitates access to that information through user-friendly search engines and apposite visualisation tools. Other databases focus on data concerning the same level of organisation of organisms (for example, Reactome gathers available data on biological pathways) and allow researchers to compare datasets derived from different organisms (the Munich Information Protein Service, or MIPS, enables comparisons between sequences of rice, Arabidopsis, maize, tomato and various other plants; The Institute for Genomic Research, or TIGR, allows for cross-examinations of functional genomics data in humans, mice and several

species of plants, microbes and fungi).[7]

I already remarked on the trust that researchers tend to grant to data as 'indisputable facts'. In fact, displaying trust in data coming from other research contexts is a matter of necessity. Within the competitive context of cutting-edge biology, short-term projects earn the highest rewards; researchers have quite literally no time to check on data produced by someone else, unless this is made unavoidable by questions, problems or discrepancies emerging in the course of applying those data to resolving new issues. Databases respond to this situation by incorporating some standards for format and quality control over data. In practice, this responsibility falls on the curators who develop and maintain databases. They are the ones deciding on issues such as which datasets are circulated and which background information is included on their provenance (protocols, instruments and materials used in producing them); the standards used to share data, such as the format used to publish and compare data of the same type; and the technical means (software, visualisation tools) by which data are circulated.

3. Regulating Data Travels: the Public and the Private Sector

Before addressing these technical hurdles in more detail, it is important to note that resolving technical difficulties is not the only challenge faced by curators. Biologists devoting their efforts

[7] For more details about how these databases operate, see the descriptions published by the team of curators responsible for each of them: for TAIR, Rhee *et al*. (2003); for Reactome, Vastrik *et al*. (2007); for MIPS, Spannagl *et al.* (2007); and finally for TIGR, Bammler *et al*. (2005).

to facilitating data exchange need to confront the contemporary regulatory context for scientific collaboration, which is strongly affected by the need to translate the results of basic research into commodities of use to society at large. In its broadest sense, commodification is of course constitutive of scientific research. For science to remain a viable and socially relevant enterprise, the value of scientific discoveries needs to be evaluated not only through epistemic criteria, but also through social and economic ones**.** Using science towards the development of new commodities (or the bettering of old ones) is one important way in which scientific understanding informs our capability to interact with the world. The push to commodify research becomes problematic only when epistemic and social criteria are neglected in favour of purely economic considerations.

I hardly need to point out the commercial significance of constructing efficient means for distributing information across biology. Future developments in biomedical research depend heavily on how data are managed and on who controls the flow of information across research contexts. At stake is the future of 'red' biotechnology (medical applications of biological research) as well as 'green' biotechnology (production of genetically modified organisms for agricultural purposes). Both pharmaceutical companies and agricultural corporations have become heavily involved in basic research on model organisms, precisely because such research yields knowledge about how to intervene on plants and animals in ways seen as desirable to potential customers. These same industries have long sought to acquire exclusive control over the flow of data produced through their research and development efforts, in the hope to use those results to develop commercially interesting results faster than their competitors. Around 70% of green biotechnology research is officially in the hands of the private sector. Academic research is following in the same path, as it becomes increasingly tied to the private sector and driven by the

necessity to produce marketable goods. The public sector is pushing biologists to pursue research with obvious biotechnological applications. Research projects aimed at acquiring knowledge of basic biological mechanisms are weeded out, as long as they do not guarantee to yield profitable applications within a short period of time.

One crucial factor in understanding the impact of profit-driven ambitions on biological research is the role played by the sponsors of such research in the governance of science. Both public and private agencies play a pivotal role in the regulation of the means through which data is distributed across research communities.[8] Not only do sponsors allocate the material resources necessary to the development of bioinformatics, but they also act as governing bodies over processes of data circulation. Their economic (and in the case of public institutions, political) power is taken to legitimise their role as legislators over goals, strategies and rules adopted by databases. Database curators are not at liberty to decide who has the right to consult the database and use data therein stored. Nor can they determine the goals and procedures to be followed in

[8] In their excellent analysis of bioinformatic networks, Brown and Rappert (2000) have argued that the labels 'public' and 'private' only serve as 'idealised codes to which various actors, whether they are universities or commercially funded initiatives, can appeal' (ibid., 444). While I agree that the notion of a public good and the related 'philosophy of free access' is evoked by all participants in bioinformatics to fit their own agenda, I view the distinction between private and public as a valid and unambiguous tool to classify the sponsors of bioinformatic efforts. As noted by Brown and Rappert, there are of course bioinformatic institutes funded by both types of sponsors; yet, recognising the difference in the values and commitments of those sponsors is necessary to make sense of the work carried out within these institutes.

storing and circulating data. Sponsors take upon themselves the responsibility of making those decisions.

Who are these sponsors? On the corporate side, we have giant industries such as Monsanto, Syngenta, GlaxoSmithKline and giant biotechnology and pharmaceutical corporations with extensive R&D facilities. These companies maintain databases for all their research output. Further, there is a boom in smaller companies providing specialised services to data producers and curators. Affimatrix©, for instance, is the most popular company assisting the production of microarray data, which are now the main source of information about gene expression outside of the nucleus. As I already remarked, universities are now closely aligned with the interests of these various companies, since most of their staff is involved in contract research in some way or another.[9] Remarkably, this is one of the reasons why universities do not play a decisive role in the regulation and development of bioinformatics efforts. This regulatory power is assigned either to the companies owning rights on the data being produced, or to the governmental funding agencies that sponsor the development of databases.

The most active public institutions allocating funding to bioinformatics are the National Science Foundation and the National Institute of Health in the United States (NSF and NIH, respectively); the European Union (EU); and national funding agencies around the world, such as the Biotechnology and Biological Sciences Research Council (BBSRC) in Britain, the German Federal Ministry of Education, Research and Technology and the Ministry of Education, Science, Sport and Culture in Japan. The extent to which these agencies are committed to regulating

[9] Krimsky (2003) documents how contract research has been steadily displacing governmental funding towards most biomedical and genomic research in the last three decades.

international data traffic cannot be underestimated. Following over a decade of investments in this direction, the NSF just launched a funding programme called 'Cyberinfrastructure', devolving 52 million dollars to the development of integrated bioinformatics tools. The EU has been almost equally generous with its Embrace programme, set up to 'improve access to biological information for scientists both inside and beyond European border'.[10] The funding program has run since February 2005 and involves 17 institutes located in 11 European countries.

The reasons for the heavy involvement of governmental agencies in regulating and funding bioinformatics are illustrated by a brief reference to one of the best-known instances of the clash between private and public interests over this issue. This is the dispute surrounding the disclosure and circulation of data from the Human Genome Project (HGP). Officially running from 1990 to 2003, the HGP was a multinational project set up to sequence the whole human genome. Its resonant success in this task made it an exemplar for many other 'big science' collaborations (such as the projects devoted to sequence the worm *C. elegans*, the mouse *Mus Musculus* and Arabidopsis).[11] The sequencing effort was funded by both the private and the public sectors. Research on the public side involved a multinational effort coordinated by Francis Collins. The main corporate investor was Perkin-Elmer Corporation sponsoring the company Celera headed by Craig Venter, the creator of the shot-gun sequencing techniques that effectively allowed the HGP to keep up with its completion schedule.

At the turn of the millennium, conflict erupted over the means through which data would

[10] From the mission statement on the EMBRACE homepage, http://ec.europa.eu/research/health/genomics/newsletter/issue4/article04_en.htm

[11] For a general account ofthe HPG, see Sulston and Ferry (2002) and Bostanci (2004).

be disclosed to the wider community. On the corporate side, Venter proposed to take over the remaining sequencing efforts from public funding and to create a database enabling access to both public and private data. In exchange for relieving governmental budgets of such expenses, Venter asked for the right to patent several hundreds of the genes mapped through the HGP, as well as the right to control access to the database for a period of at least five years, during which only researchers busy with non-profit projects would be given permission to view and use data. Speaking for his publicly funded, multinational research group, Collins put forward a number of critiques of the terms set by Celera. First, he remarked, there is no unambiguous way to demarcate profitable from non-profit research, as by now any project in basic biology might yield insights that can be commercially exploited at a later time. This meant that Venter's conditions effectively blocked the great majority of researchers from gaining access to the database. Collins also claimed that public agencies could grant Celera no longer than one year of unilateral control of the data. Five years of exclusive access would prevent the development of research that builds on the sequencing data, thus halting genomics in the most exciting moment of its history and barring biologists from exploring the significance of those data to other research fields, ranging from cell biology to ecology. Finally, Collins condemned Venter's requirements as an attempt to take over the results of investments by the public sector and exploit them for the commercial purposes of his company. Collins argued that accepting Venter's proposal meant fostering a monopoly over the access to and use of HGP data. Given the importance of such data to future biomedical research, sanctioning corporate claims of exclusivity would have been not only misguided, but also immoral – a judgement that was shared by other researchers working for public agencies.

Eventually, Celera gave in to most of Collins' demands and disclosed its data through

publication in *Science* at the same time as the publicly funded researchers published in *Nature*. As discussed in detail by Bostanci (2004), however, the disagreements between private and public parties of the HGP remained, and the dispute over the means of data disclosure symbolised a deeper disagreement about the means and goals of research.[12] This is the point that I wish to emphasise in the next section.

4. Product-driven versus Resource-driven Competition

As evident in the HPG dispute, both public and private sponsors are susceptible to the demands of commodification and posit financial profit as an important goal of scientific research. However, they have different ways of specifying this minimal sketch of what commodification involves. Private sponsors see data as means to achieve marketable commodities. Data are in this view indispensible to developing the knowledge needed to develop new products. By contrast, public sponsors value data themselves as commodities with great potential for multiple uses: each dataset can potentially serve the development of a variety of ideas and products, which makes it a vital resource to whoever is involved in research. These two approaches encourage contrasting sets of criteria for what constitutes 'good' science. As a consequence, public and private sponsors adopt diverging strategies towards regulating data distribution in biology.

Let us tackle private sponsors first. Corporations involved in scientific research have a strong preference for short-term efforts to produce immediately applicable results. Their

[12] For documentation on this case, see also Marshall (2000; 2001).

assessment of the value of biological data is based on an estimate of the commercial value of products that are likely to be obtained from analyzing those data. Most importantly, products chosen as targets of a company's R&D efforts need to be developed and marketed before competitors in other industries or in the public sector reach the same result. The priority is to be the first to create a product of a specific type. As a consequence of such *product-driven competition* between companies, R&D departments are reluctant to share the data that they produce in-house, since the possession of unique datasets might constitute an advantage over competitors (and viceversa, data that are disclosed might end up helping competitors in their own quest). Data are not interesting in themselves, but rather as a means to achieve the scientific and technical knowledge that might allow for a commercially marketable discovery.

Thus, researchers working under private contracts take a short-term view on the quality and maintenance of data that are produced. Data quality is assessed in relation to the way in which data serve the creation of a viable product. Data are considered to be good when they guide biologists towards the realization of efficient means of intervention on an organism. Hence, privately sponsored research seldom adopts standards for data quality that do not depend on the specific research context; also, private sponsors are not interested in investing money towards the long-term maintenance of data produced in the course of a project, unless those data are thought to be potentially useful for in-house projects to come. As long as data are no longer of use to the company itself, no more time and money should be spent on them.

In practice, this set of values leads private sponsors to favor *project-directed databases*, i.e. databases that gather all available data that is relevant to exploring the specific problem tackled by researchers in a given period. These databases are quick to set up and yield results, since the range of data involved is very limited and there is little curation work involved.

However, they are maintained only as long as they are useful to the production of the range of products of interest to the company. Data stored within those resources thus risk to be lost, as the databases are discarded on completion of the project at hand. Also, since sharing data could enhance the chance of a competitor developing the same product in a shorter amount of time, project-based databases sponsored by private companies bar access and/or permission to use data to researchers who have no direct ties to their sponsors (note that they often give the option of building such ties as a way to gain admission to the database).

Public sponsors have a different view of both the role of data in science and the role of communication among researchers trying to transform data into products. The key value here is a long-term view on the possible developments in biology and the ways in which a strategic management of present knowledge might foster high returns in the future. Public sponsors invest large quantities of money in producing data and are interested in maximizing that investment by making sure that those results are used in as many ways and with as much impact as possible. This leads to a view of data as more than a means to the fast production of commodities: data are themselves seen as commodities whose potential utility is not yet clear and should be explored through appropriate resources. This standpoint is reinforced by the realization that, in practice, exploring the relevance of data is not compatible with retaining control on who can use data and when. In order to determine whether a given dataset might be relevant to their research, biologists need to be able to access it directly, compare it against all other available datasets and interpret it in the framework of their own research. Given the large amount of data whose relevance needs to be assessed, it is vital that biologists have unrestricted and quick access to all available datasets, thus increasing the possibility of finding datasets suiting their research interests. Ultimately, constructing tools facilitating data circulation to anyone interested is the

most efficient way to yield profitable results out of the efforts involved in producing data in the first place. Data need to be made accessible and usable to any researcher interested in assessing their significance, no matter who funds them or what they are aiming to produce.

Public sponsors have therefore moved from an emphasis on product-driven competition to encouraging *resource-driven competition.* This kind of competition acts at two levels: between research groups and between databases themselves. Between research groups, sponsors exploit competitive forces to push researchers to donate their data to public databases. There is actually no consensus yet on what constitutes an appropriate reward for 'data donors', since despite the efforts and time spent in disclosing data of good quality, data donation is not yet officially recognized as part of a researchers' curriculum vitae. Public agencies are acutely aware that this situation needs to be changed: research groups should be encouraged to compete not only for the number of publications or patents produced, but also for the number and quality of donations achieved. Strategies hitherto used to this end include context-specific rewards, such as the offer of specific services or materials in exchange for a donation to a database,[13] and disclosure obligations tied to publicly awarded grants, which imply that researchers sponsored by those grants disclose the resulting data to public repositories (this is a policy currently endorsed by the BBSRC, NIH and NSF).

At the same time, governmental agencies encourage competition between databases for who provides the best service to their users. The success of a database, and thus decisions on its

[13] The BBSRC-funded Nottingham Arabidopsis Information Centre, for instance, offers to perform micro array experiments at a low price in exchange for the permission to disclose all data obtained through this procedure to public repositories.

long-term survival through follow-up grants, is judged on the basis of the amount of users that it secures (as documented by surveys and website statistics).[14] This encourages database curators to put the interests and expectations of their users before their own. There is a constant trade-off between what the curators view as efficient ways to package data and what users from various contexts see as useful search parameters and forms of display. As a result of current public policy, curators need to be aware of what biologists expect to find on the database and how they will be handling the data, since user satisfaction will be the determinant factor for the survival of their database. A further effect of governmental insistence on competition for user shares is the progressive diversification of databases seeking to please different needs. Curators have realized that there is no point for two databases to collect precisely the same type and amount of data in the same ways, as they would be competing for the attention of same users and one of them could eventually lose out. As a result of this insight, the landscape of existing databases is exhibiting more and more self-regulating division of labor – and at the same time, extensive networks of collaboration among databases are emerging (since, even if sponsored by different agencies, database curators can usefully exchange notes on how best to serve their user communities and how to boost each other's work by building links between databases).[15]

[14]Again as an example, the Nottingham Arabidopsis Stock Centre was recently granted funds by the BBSRC on the grounds of user satisfaction surveys and statistics documenting how many researchers accessed and used their existing database.

[15] Yet another interesting instance of competition in this context is the one existing between different funding agencies, such as the competition between NSF and NIH in the United States, or between American and European agencies. These agencies might be characterised as pushing

In all these different ways, resource-driven competition becomes a tool towards achieving an array of resources and methods facilitating all foreseeable types of research. This approach can certainly have unintended consequences which are potentially damaging to science. For instance, the division of labor occasioned by resource-based competition risks to diminish opportunities for dissent among database curators and pluralism among packaging strategies, as it reduces the chances to develop and test different packaging processes for the same data. Also, with databases building more and more of their work on each other's efforts, chances of perpetuating errors and ultimately wrong approaches increase (although it should be noted that comparisons across databases can also highlight inconsistencies, thus signaling places where the quality and reliability of available data could be improved[16]). Last but not least, user interest alone is not enough to guarantee user satisfaction, as researchers might be consulting databases because they are the only source of information available, without however approving of the choices made by curators in packaging the data. To maximize the chance of data re-use across research contexts, public sponsors need to find better ways to assess what researchers wish to

---

different versions of resource-driven competition, insofar as some of them (e.g. the NSF) favour a centralised approach to database construction, with one group of 'superexperts' responsible for a whole sector, while others (e.g. the BBSRC) prefer to decentralise funding into different curator pools. While interesting in themselves, these differences in regulatory policy do not impact my argument in this chapter, as all agencies agree on treating resource-driven competition as an efficient strategy to circulate data.

[16] See Ruttenberg *et al*. (2007).

find in a database.[17]

These are surely only some of the possible complications involved in adopting resource-driven competition as a mechanism pushing data circulation. Their damaging effects may or may not be averted by improved policies and scientific practice. What I wish to emphasize here is that resource-driven competition does enforce the development of standards for producing and handling data that *do not* depend on the demands of one research context only.[18] This already constitutes a huge advance over the product-driven competition favored by private sponsors, as public institutions encourage the construction of databases aiming to serve biological research as a whole. This places careful maintenance and free circulation of data as important criteria for what constitutes 'good science'. Indeed, resource-driven competition has hitherto proved very productive from the scientific point of view. Within barely a decade, publicly sponsored databases have made enormous leaps in the quality of their services and of the data that they contain. Scientists note the increasing usefulness of databases in their research and are therefore becoming more aware of the advantages of contributing their data to these resources, which are seen as crucial services yielding high returns to whoever can afford a long-term view on the

[17] Another problematic issue, which is not directly related to resource-driven competition however, is the lack of commitment of funding agencies to maintaining databases in the long term. Up to now, most governmental funding of bioinformatics is on a limited time-scale, which encourages curators to constantly improve their services, but offers no secure support for the long-term storage of data.

[18] This point was forcefully advocated by Olson and Green (1998) in the context of the HPG dispute.

value of their data.

5. Data Travels in Commodified Science

I now turn to examine the three stages through which scientists actually use databases to distribute data. These three stages of data travel involve three sets of actors: database curators, scientists who produce data in the first place ('producers') and users of data retrieved through databases ('users'). In each of these stages, a number of difficulties need to be overcome for data to be shared across research communities in a manner that facilitates as much as possible the overall advancement of research. The contrasting values adopted by database sponsors have a strong impact on how producers, curators and users deal with those technical difficulties. This analysis highlights how the product-driven competition encouraged by the private sector fails to reconcile the roles of bioinformatics as a research field and service to scientists with its role as an industry seeking to profit from available data.

5.1 Disclosure

There are no *general* rules in science about how researchers should treat the data that they produce. While in some cases the disclosure of data is policed by journal editors or funding agencies (see above and section 7 of Brown's chapter in this volume), the majority of researchers can still choose to discard specific datasets when they do not fit their interests or goals, so that no

one will be able to see them again. Indeed, there are as yet no standard mechanisms within science regulating the selection of data to be disclosed from the wider pool of data produced by any one research project. This is partly because there is no consensus on what data are produced for. Clearly, data are produced as evidence for the hypotheses and beliefs characterising a specific research context. It makes perfect sense, in this interpretation, to disclose only data of direct relevance to the questions investigated in that context. At the same time, however, data can be seen as a heritage to be shared among various researchers interested in different aspects of the same phenomenon. Making every bit of data produced in one's research accessible to others could prevent useless duplication of efforts, thus giving biologists more time to probe the significance of existing data and/or produce new ones.

This ambiguity in the goals of data production leaves scientific sponsors at liberty to impose their own values and regulations on the disclosure of data. As I pointed out in the previous section, private sponsors encourage scientists towards selecting data on the basis of their usefulness within the specific project in which they are produced. This is due to the instrumental constraints imposed by product-driven competition, in which there is simply no time to store and manage data that are not immediately relevant to the project at hand. In the private context, disclosure also depends on the level of control that sponsors wish to retain on the data. Producers are often asked to refrain from disclosing them for a specified time period, thus giving time to the sponsors to fully reap the commercial fruits of related discoveries. Alternatively, privately funded researchers may disclose data through various types of IPRs granting exclusive legal ownership of the material being disclosed, including the power to control who gets to use data and under which conditions.

Researchers whose contract allows for public disclosure of (at least some of) their data

have a choice between two means of disclosure. One is publication in a scientific journal. The incentives to disclose data through publications are very high for producers working in academia, where the number of one's publications constitutes the main indicator for the quality of one's research. Through publishing, producers earn academic recognition for their efforts and thus the right to apply for (or maintain) jobs in scientific institutions. The disadvantage with this method of disclosure is that it mirrors many of the values and methodological criteria underlying the product-driven competition fostered by private sponsors. Researchers disclosing data through publications tend to select those that directly support the specific claim made in their paper(s). This means again that the majority of data actually generated is never seen by other biologists. Also, because data are treated as the evidential means towards demonstrating one claim, little attention is paid to the format with which data are published. Journals seldom have rules on which format data should be reported in a publication, which means that researchers present data in the format that best fits their present purposes. This has two crucial implications. First, only biologists with a direct interest in the topic of the paper will access those data, regardless of the fact that the same data could be useful to investigating other biological questions. Second, without some expertise in the topic addressed by the paper, it can be very difficult to extract data from it.[19]

[19] The NSF-sponsored TAIR database has been searching for efficient ways to extract data from publications since almost a decade. This process, aptly dubbed 'text-mining' by bioinformaticians, is known to be both time-consuming and exceedingly subjective, as curators need to interpret the biological significance of the claims made in the paper in order to adequately export data from that context (Pan *et al*. 2006).

There is an alternative to this method for disclosure and to the assumption that data are only produced to provide evidence for one specific claim, no matter their potential relevance to other research projects. This is donation to public repositories, also referred to as 'large-scale public databases' (Rhee, Dickerson and Xu, 2006).[20] Researchers can choose to donate all of their data to a repository (such as GenBank). This method of disclosure adheres quite closely to the resource-driven competition characterising public governance of data sharing. Public repositories provide a platform for producers to contribute the results of their work so that database curators can use them to construct databases that the whole community (including the original producers) can enjoy. As I detail below in the circulation and retrieval stages, contribution to a public repository is the first, indispensable step towards enabling efficient data sharing across biologists.

If the goal of producing data was solely to provide a legacy to biology as a whole, this form of disclosure would indisputably constitute the best option for everyone's benefit in this case. However, disclosure through public repository requires extra work on the side of producers, who have to format their data according to the minimal standards demanded by the repositories and have to take account of all the data that they produce, rather than simply the ones relevant to answering their own research question in a satisfactory way. Further, donation to public repositories is not yet fully recognised as a valuable contribution to science. It is certainly valued by individual scientists as a gesture of good will and openness, but it will not get people jobs or boost their CV. These are big issues for researchers under strong pressure to move

[20] Hilgartner (1995) has put forward the idea of referring to journals and databases as two different *communication regimes*.

quickly from one project to the next and to maximise the recognition that they receive for each piece of research. Another, stringent reason for researchers to prefer disclosure through publications over donations to repositories is the issue of ownership of data. Donation to public repositories requires producers to relinquish control of the data that they submit, so that they can be freely accessed and used by other members of the community. This clause is in direct conflict with their sponsors' demand to retain control over the spread and use of the data. Thus, privatisation drives researchers away from freely donating their data to public repositories.

5.2 Circulation

The mere disclosure of data through public repositories is not sufficient for biologists to be able to access and use those data in their own work. Due to both the amount and the diversity of data hosted by them, accessing data through repositories is not an easy task. There are no categories through which to search for specific sets of data; the formats in which data are presented are still rather heterogeneous, since each contributor of data tends to interpret and apply the standards imposed by the repository in her own way; and, most importantly, there are no tools through which users can visualise correlations among existing sets of data (such as for instance tools to assemble all data relevant to the sequence of genes on a chromosome, or models allowing one to view and compare all available data on a specific metabolic pathway).

These are the problems that the so-called 'community-databases', i.e. the entities that I hitherto referred to as databases, are funded to tackle. Their role is to extract data from either public repositories or other forms of disclosure (such as publications or even through direct

interaction with data producers) and standardise those data in order to make them easily accessible to all biologists, no matter their specific expertise or location. Database curators are responsible for decisions concerning data selection (which data will be inserted in the database and which information on data source will be made available) and the 'packaging' of data (the standard format in which data of the same type should be presented and the taxonomy through which data should be ordered in order to be easily retrieved by users[21]). Publications have tacit rather than formal rules as to what information – and to which level of detail – to insert about protocols, instruments and assumptions used in a study. Databases are much more exigent in their requirements, because, as I noted above, curators are responsible for verifying the quality and reliability of data hosted in their databases.

Notably, the role played by curators here is peculiar to resource-driven competition and indeed these databases are sponsored almost exclusively by public agencies. These databases typically seek to serve the whole community of potential users by *making data usable for multiple purposes*. Efficiency, in the view of their curators, consists in enlarging the number of research contexts in which the same sets of data can be relevant. Product-directed databases are not interested in the outreach of data (which in fact they seek to control) as much as they are interested in their applicability to a specific context. In that context, there is neither time nor resources to curate data so that they are reusable in other contexts. This factor alone greatly limits the extent to which these data can be distributed, as users have to do a lot of work to retrieve

[21] These taxonomies, which bioinformaticians refer to as 'bio-ontologies', include precisely defined categories that allow users to search and compare data. On bio-ontologies, see Baclawski and Niu (2005).

them.

5.3 Retrieval

Users exercise two kinds of expertise to adequately retrieve data from databases. The first kind of expertise concerns the actual act of searching for data. Users need to be able to log into a database; move efficiently through the database interface; phrase their query in a way that is compatible with the parameters and visualization tools built into the database; and, finally, maneuver through the results displayed by the database until they obtain a visualization of data that is satisfactory to them. These are what I call 'access skills'. Without them, a user cannot hope to retrieve the data that she wishes to consult – which is why a lot of the curators' work consists in making these skills as easy to acquire as possible, thus minimizing the time that users have to spend in familiarizing themselves with the database and improving the chances that they get what they want from it.

The second kind of expertise needed by users is the ability to actually use the data acquired through the database within their own research. This implies an altogether different set of skills, which I call 'expert skills' and which are acquired as part of biologists' own training and practice, rather than in direct connection to database use.[22] The exercise of expert skills

[22] A good example of the difference between access and expert skills is the difference between the skills exercised by myself and by a practicing biologist in accessing a database. Through my philosophical research on databases and biological knowledge, I have become reasonably skilled

requires a thorough knowledge of both the practices and the theoretical apparatus used within the disciplines dealing with the broad research question that is being asked.[23] It is on the basis of this background knowledge that biologists determine which sets of data could potentially inform their investigation of the research question. Through scrutiny of data accessed through a database, a biologist with adequate expert skills can substantially increase the precision of her research question as well as use the new information to design her future research.

Consider the example of a biologist specialized in plant growth, who wishes to study how a specific hormone influences the expression of a particular phenotypic trait. For a start, she might check whether there are any data already available on which gene clusters are affected by the hormone. If she discovers that there are indeed specific genes whose expression is strongly enhanced or inhibited by the hormone, she will have grounds to think that whichever phenotypic trait is controlled by those genes will be affected, too. Again, she can check whether there are any data already available documenting the correlation between the gene cluster that she has identified and specific phenotypic traits in her model plant. If that is the case, she will be able to form a hypothesis about which traits are influenced by the hormone, and she will thus modify her research design in order to test her hypothesis.

---

in accessing biological databases and getting some data out of them. However, I do not know how to use those data to pursue a specific research question in biology. This requires a commitment to goals that I do not share as well as a familiarity with cutting-edge techniques, methodologies and concepts in specialized areas of research that I do not have.

[23] A detailed analysis of how biologists coordinate embodied and theoretical knowledge of a phenomenon to acquire understanding of that phenomenon can be found in Leonelli (2009b).

Up to this point, the researcher has used her access to the database to identify possible causal links between the phenomena that she is interested in. This has helped her to construct a more detailed research question and experimental setting. To proceed with the investigation, the biologist might need to gather more information about the provenance of data, so as to assess with more detail their quality and reliability with regards to her specific research context. This is where the information on data sources provided by curators becomes extremely useful. As I noted in my first section, 'travelling' data are everything but local: their anonymity is a crucial factor in allowing them to circulate widely across research contexts. However, data become 'local' again once they are adopted into a new context and used to pursue new research questions. In this phase, information about their provenance is often important to evaluating their role in the new domain (Leonelli 2009a).

A resource-directed database is constructed to minimize the skills needed to access the database and the information on data sources. The database is specifically built for consultation by any disciplinary background: as we have seen in the circulation stage, data are standardized and ordered so as to travel across disciplinary boundaries. Further, curators invest much effort in adding information about the provenance of data, which is not crucial to circulating the data, but is often very helpful to researchers wishing to use retrieved data in their projects. Researchers wishing to exercise their expert skills in using retrieved data have the needed information immediately at their disposal.

By contrast, project-driven databases serve the specific disciplinary interests informing the work of whoever produces the data. This implies that curators do not take time to standardize the data and the tools through which data are displayed to the user. The access skills needed to retrieve data from such a database are specific to the specific field in question, which makes them

difficult to acquire for researchers working in other fields. This means that even if these databases were always freely accessible, the probability that a researcher will actually make the effort to retrieve data from them is very low. Further, project-driven databases do not invest effort into adding information about the local conditions where data were produced, as this would imply investing time and money in employing curators to do this work. The result is a list of anonymous data. These data can certainly be circulated if the access skills needed to retrieve them were easy enough to acquire. However, their usefulness within a new research context is severely compromised by the lack of information about their provenance.

6. Conclusion: Values in Data Circulation

My discussion of how the priorities of database sponsors affect the three stages of data travel brings me to the following conclusion. The privatisation of research does not affect the dissemination of data solely by attempting to control it through the exercise of Intellectual Property Rights, by distorting or spinning the data, or by affecting the research directions to which data are brought to bear (as illustrated respectively by Brown, Resnik and Musschenga, van der Steen and Ho in this volume). Private sponsors affect data circulation, and therefore the development of future research, by imposing criteria for what counts as data in science and how these data should be treated. These criteria are dictated by values such as speed and instrumentalism, which are in turn related to specific methodological procedures: product-driven competition and a preference for project-directed databases. Biologists are long discussing whether the insistence on seeking IPRs in contract research obstructs the community's *freedom*

to operate on the basis of the data that are produced in that context (e.g. Delmer *et al*., 2003). I wish to add that the very values and temporal constraints that privatisation currently imposes on scientific practices obstruct the development of future research.[24]

Science and technology are characterized by the ability of their practitioners to build new research projects on the insights acquired through old ones. The practices encouraged by product-driven competition force researchers to shy away from contributing to the bioinformatics effort towards improving existing resources for the circulation of data. As a result, they jeopardise current opportunities for an efficient transmission of knowledge. More specifically, *product-directed competition compromises the opportunity to use the same set of data for multiple scientific purposes.* This could be very damaging to science in the long term. Science and society at large seem to have everything to lose from the obstacles posed to data circulation by industries and, increasingly, universities.[25]

---

[24] Privatisation is of course not the only mechanism imposing the values characterising product-driven competition. The habit to assess scientists' output through number of publication generates similar problems: a tendency to value the usability of data towards 'minimally publishable units' rather than their usability in the long term.

[25]Arguably, technologies such as databases provide opportunities for collaboration never before seen in biology or other sciences, because they free existing datasets from their disciplinary and geographical provenance. It is also true that the contemporary setting of 'big science' differs so vastly from how science was conducted in earlier periods as to make comparisons almost impossible: the globalisation of scientific education and research, as well as the invention of

This situation is recognised by governmental agencies, which therefore support a resource-driven policy over a product-driven one. When it comes to determining procedures for data sharing, public agencies often act as gate-keepers for what Dick Pels calls 'self-interested science'[26] by endorsing the following key values:

1. *equal access to resources*: especially in the context of biological research, where expertise is fragmented into specialized niches and division of labor is efficiently used to achieve common research goals, it is of paramount importance that researchers of any specialty have equal access to basic resources such as data;
2. *competition between different methods to achieve a common goal*: research groups are encouraged to compete on creating and improving resources and procedures useful to carrying out research (rather than competing purely on the quantity and quality of research results, i.e. number of publications);

---

technologies gathering data of all types at increasing speed, make the question of data circulation more pressing than it has ever been in the history of science.

[26] Pels introduces the idea of 'self-interested' science as a useful way to overcome the idea that the current commodification of science is destined to completely erode the boundary between scientific and political or commercial activities (Pels 2003, 30). As Pels notes, science should work with a distinctive methodology and values compared to other human activities: the reasons for this are less to do with Enlightenment ideals, however, than with the scientists' interest in safeguarding their own profession from excessive manipulations from 'external' forces (such as the market or the state), which may compromise its functioning by distorting its methodology and procedures.

3. *long-term vision*: investing time, as well as money and human resources, is of the essence in scientific research: 'science is typically of the 'long breath', depending on long-term cycles of investment in human and material resources, whereas politics expects quicker returns within a much shorter time-span' (Pels 2003, 32).

Adherence to these values allows public agencies to keep their commitment to the goals and means of commodified science, without however losing sight of key methodological requirements for 'good science', such as the need to share data freely and efficiently.[27] Providing means for adequate data circulation maximises the usefulness of research that has already been done and paid for. In fact, it could be argued that it is just as important to maximise the flow of data across research contexts from a profit-driven perspective as it is from a Mertonian perspective. The construction of platforms through which data can be circulated and thus re-used towards further research represents a great improvement in the efficient use of public research funds to serve the public interest, even if the latter is defined through appeal to the potential commodification of research.

In closing, I want to draw attention to the peculiar situation that allows publicly sponsored research to support the free exchange of scientific knowledge. If the advantages of this strategy are so great, why is it that private sponsors do not embrace them? For the same reasons as the ones motivating public sponsors, it would seem rational for them to pursue resource-driven competition rather than insisting on the short-sighted strategy of product-driven competition – a point that some of the main biotechnology and pharmaceuticals corporations are starting to take

[27] In this sense, these values constitute good examples of the 'deflationary' Mertonian norms proposed by Radder in this volume.

on board. At least a partial explanation for this difference is provided by the social roles and economic power characterising private and public institutions. By its very nature, publicly sponsored research is at an advantage with respect to privately sponsored research. A government, at least among the majority of Western representative democracies, is a much more stable and durable entity than a company and can afford to invest capital in projects guaranteed to yield returns in the long term. Thus, public agencies can better afford to adopt resource-driven competition. Further, investing in facilitating data circulation has political as well as economic benefits. By encouraging cooperation among databases, resource-driven competition opens opportunities for international cooperation among countries involved in the same type of research, thus fostering diplomatic ties and political trust.

Individual companies, and particularly small businesses, do not enjoy these advantages. They need short-term profit to survive: a long-term vision on scientific research is difficult to maintain by an entity whose very existence depends on monthly revenues and the support of shareholders. As a consequence, they are more strictly bound to the market rules dominating international trade, which do not offer opportunities for long-term analysis. A fact that seemingly proves this point is that the only corporations willing to donate some of their data to publicly funded databases are giants like Monsanto. The company justifies this policy of disclosure by pointing out that public databases such as TAIR take better care of data on Arabidopsis than Monsanto itself would (as Monsanto does not intend to invest more money in maintaining the data). The underlying reality is that Monsanto can afford to make such a donation and reap its benefits in the long term. The same cannot be said of the hundreds of satellite companies specialising on one project at a time and producing much smaller and less organically compiled databases.

Acknowledgments

I am grateful to Bram Bos, Hans Radder and Mary Morgan for their insightful comments on an earlier draft. I also benefited from discussions with the participants to the Amsterdam workshop held on 21-23 June 2007; with various biologists and database curators, particularly Sean May of NASC; and with my colleagues at the London School of Economics. This research was funded by the Leverhulme Trust (grant number F/07004/Z) and the ESRC as part of the project 'How Well Do 'Facts' Travel?' at the Department of Economic History, LSE.

Bibliography

Ankeny, R. (2007). Wormy Logic: Model Organisms as Case-Based Reasoning. In: Creager, A.H., Lunbeck, E. and Wise, N. (eds.) *Science without Laws: Model Systems, Cases, Exemplary Narratives*. Chapel Hill, NC: Duke University Press, pp. 46-58.

Baclawski, K. and Niu, T. (2005). *Ontologies for Bioinformatics*. Cambridge, MA: MIT Press.

Bammler, T., Beyer, R.P., Bhattacharya, S., Boorman, G.A., Boyles, A., Bradford, B.U., Bumgarner, R.E., Bushel, P.R., Chaturvedi, K., Choi, D., Cunningham, M.L., Deng, S., Dressman, H.K., Fannin, R.D., Farin, F.M., Freedman, J.H., Fry, R.C., Harper, A., Humble, M.C., Hurban, P., Kavanagh, T.J., Kaufmann, W.K., Kerr, K.F., Jing, L., Lapidus, J.A., Lasarev, M.R., Li, J., Li, Y., Lobenhofer, E.K., Lu, X., Malek, R.L., Milton, S., Nagalla, S.R., O'Malley,

J.P., Palmer, V.S., Pattee, P., Paules, R.S., Perou, C.M., Phillips, K., Qin, L., Qiu, Y., Quigley, S.D., Rodland, M., Rusyn, I., Samson, L.D., Schwartz, D.A., Shi, Y., Shin, J., Sieber, S.O., Slifer, S., Speer, M.C., Spencer, P.S., Sproles, D.I., Swenberg, J.A., Suk, W.A., Sullivan, R.C., Tian, R, Tennant, R.W., Todd, S.A., Tucker, C.J., Van Houten, B.,Weis, B.K., Xuan, S. and Zarbl, H. (2005). Standardizing Global Gene Expression Analysis Between Laboratories and across Platforms. *Nat Methods* 2, 5: 351-356.

Bogen and Woodward (1988). Saving the Phenomena. *The Philosophical Review*, 97, 3: 303-352.

Bostanci, A. (2004). Sequencing Human Genomes. In: Gaudilliere, J.P. and Rheinberger, H.J. (eds.) *From Molecular Genetics to Genomics.* New York: Routledge. Pp.158-179.

Brown, N. and Rappert, B. (2000). Emerging Bioinformatic Networks: Contesting the Public Meaning of Private and the Private Meaning of Public. *Prometheus* 18, 4: 437-452.

Delmer, D.P., Nottenburg, C., Graff, G.D. and Bennett, A.B.(2003). Intellectual Property Resources for International Development in Agriculture. *Plant Physiology*, 133, 1666-1670.

Dupré, J. (1993). *The Disorder of Things*. Cambridge, UK: Cambridge University Press.

Fuller, S. (2000). *The Governance of Science*. Buckingham, Philadelphia: Open University Press.

Hilgartner, S. (1995). Biomolecular Databases: New Communication Regimes for Biology? *Science Communication* 17: 240-263.

Knorr Cetina, K. (1999). *Epistemic Cultures*. Cambridge, MA: Harvard University Press.

Krimsky, S. (2003). *Science in the Private Interest*. Oxford: Rowman & Littlefield Publications.

Leonelli, S. (2007) *Weed for Thought: Using Arabidopsis thaliana to Understand Plant Biology*. Doctoral Thesis in Philosophy, Vrije Universiteit Amsterdam. http://hdl.handle.net/1871/10703

Leonelli, S. (2009a) On the Locality of Data and Claims About Phenomena. *Philosophy of Science* 76, 5.

Leonelli, S. (2009b). The Impure Nature of Biological Knowledge and the Practice of Understanding. In: De Regt, HW, Leonelli, S and Eigner, K (eds.) *Philosophical Perspectives on Scientific Understanding*. Pittsburgh: Pittsburgh University Press.

Marshall, E. (2000). Talks of Public-Private Deal End in Acrimony. *Science* 10 March 2000, 287, No. 5459: 1723-1725.

Marshall, E. (2001) Sharing the Glory, Not the Credit. *Science* 16 February 2001, 291, No. 5507: 1189-1193.

Mitchell, S. (2003). *Biological Complexity and Integrative Pluralism.* Cambridge, UK: Cambridge University Press.

Olson, M. and Green, P. (1998). A 'Quality-First' Credo for the Human Genome Project. *Genome Res*. 8: 414-415.

Pan, H., Zuo, L., Kanagasabai, R., Zhang, Z., Choudhary, V., Mohanty, B., Lam Tan, S., Krishnan, S. P. T., Veladandi, P.S., Meka, A., Keong Choy, W., Swarp, S. and Bajic, V.B. (2006). Extracting Information for Meaningful Function Inference through Text-Mining. In: Eisenhauer, F. (ed.) *Discovering Biomolecular Mechanisms with Computational Biology*. Austin, TX: Landes Bioscience and Springer, pp.57-73.

Pels, D. (2003). *Unhastening Science*. Liverpool: Liverpool University Press.

Rhee, S.Y., Dickerson, J. and Xu, D. (2006). Bioinformatics and Its Applications in Plant Biology. *Annu. Rev. Plant Biol.*, 57: 335-360.

Rhee, S.Y., Beavis, W., Berardini, T.Z., Chen, G., Dixon, D., Doyle, A., Garcia-Hernandez, M., Huala, E., Lander, G., Montoya, M., Miller, N., Mueller, L.A., Mundodi, S., Reiser, L., Tacklind, J., Weems, D. C., Wu, Y., Xu, I., Yoo, D., Yoon, J. and Zhang, P. (2003). The Arabidopsis Information Resource (TAIR): a Model Organism Database Providing a Centralised, Curated Gateway to Arabidopsis Biology, Research Materials and Community. *Nucleic Acid Research*, 31, 1: 224-228.

Ruttenberg, A. et al (2007). Advancing Translational Research with the Semantic Web. *BMC Bioinformatics* 8 (Suppl. 3), S2: 1-16.

Spannagl, M., Noubibou, O., Haase, D., Yang, L., Gundlach, H., Hindemitt, T., Klee, K., Haberer, G., Schoof, H. and Mayer, K. F. X. (2007). MIPSPlantsDB – Plant Database Resource for Integrative and Comparative Plant Genome Research. *Nucleic Acids Research*, 35: database issue.

Sulston, J. and Ferry, G. (2002) *The Common Thread: A Story of Science, Politics, Ethics and the Human Genome*. London: Bantam Press.

Vastrik, I., D' Eustachio, P., Schmidt, E., Joshi-Tope, G., Gopinath, G., Croft, D., de Bono, B, Gillespie, M., Jassal, B., Lewis, S., Matthews, L., Wu, G., Birney, E. and Stein, L.(2007). Reactome: A Knowledge Base of Biological Pathways and Processes. *Genome Biology* 8, 3: R39.