Working Papers on The Nature of Evidence:
How Well Do 'Facts' Travel?
No. 27/08

Regulating Data Travel in the Life Sciences: The Impact of Commodification

Sabina Leonelli

Department of Economic History
London School of Economics

April 2008

“The Nature of Evidence: How Well Do ‘Facts’ Travel?” is funded by The Leverhulme Trust and the ESRC at the Department of Economic History, London School of Economics.

For further details about this project and additional copies of this, and other papers in the series, go to:

http://www.lse.ac.uk/collection/economichistory/

Series Editor:

Dr. Jon Adams
Department of Economic History
London School of Economics
Houghton Street
London, WC2A 2AE

Tel: +44 (0) 20 7955 6727
Fax: +44 (0) 20 7955 7730

Regulating Data Travel in the Life Sciences: The impact of commodification[1]

Sabina Leonelli

Abstract

The travel of small facts (such as data) across geographical locations and disciplines is increasingly regulated by the private and public sponsors of digital databases. My analysis focuses on the contrast between the strategies supported by the public and private sectors in governing bioinformatic strategies of data exchange. Up to now, private sponsors have encouraged *product-driven competition* among database curators and users, which results in the creation of databases whose use and survival is bound to the specific projects in which they are employed. Public sponsors have tended instead to favour *resource-driven competition*, where databases are seen as resources for all biologists in the long term, irrespectively of the specific context of use. By focusing on this difference and its consequences for the advancement of biomedical research, I show how the ongoing commodification of the life sciences affects the ways in which small facts travel across research contexts. I conclude that the values and methodological criteria currently endorsed by privately sponsored research have a disruptive impact on the ability of researchers to build on each other's work, an issue that is increasingly recognised both by governmental agencies and by the corporations involved in data production.

Introduction

Philosophers of science tend to focus their attention on the conditions under which scientific knowledge is produced and applied. This paper considers instead the conditions under which knowledge is *exchanged* in science, with particular attention to the boom in bioinformatic resources characterising contemporary biology and

[1] This paper will appear in the collection "The Commodification of Academic

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 build 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, [2] 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 uncertainty about the value of some types of data as evidence.[3] The regulation of data circr

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 paper 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 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.

1. 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).

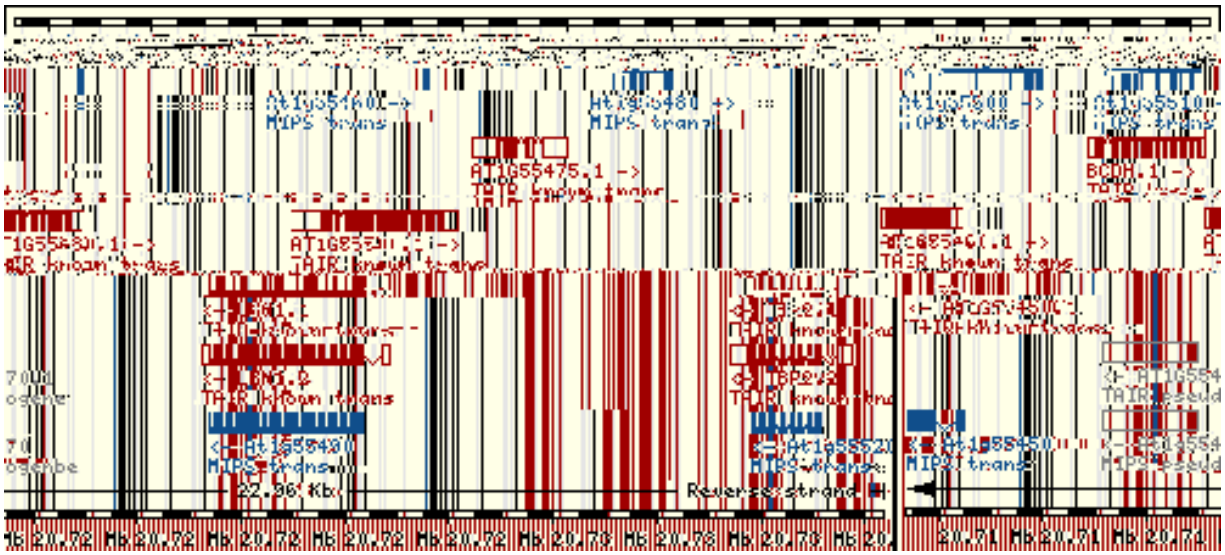

Figure 1. The red, blue and gray marks indicate the position of gene markers on a chromosome (represented by the dotted black lines at the top and bottom margins of the image) as detected by various investigators (the data of each contributing research group is marked by a different colour). *Courtesy of the Munich Information Centre for Protein Sequence.*

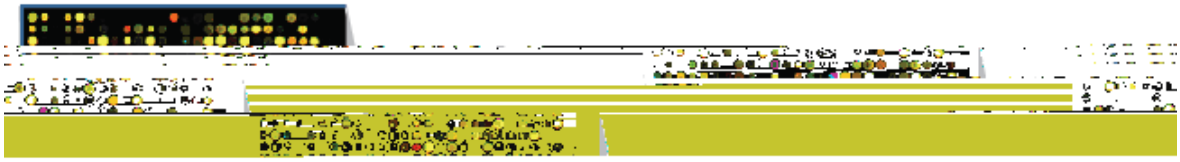

Figure 2. The coloured dots visible in the enlarged section of this microarray cluster represent the expression levels of specific genes in a particular region of a chromosome. (*Downloaded from the Internet, June 2007*)

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. Thus, 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.[4] 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 return 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 observation 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 eventually 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

[4] '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).

can take advantage of the ocean of data produced by each of the thousands of laboratories involved. And 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.

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.

2. Regulating data travels: the public and the private sector

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 of using those results to develop commercially interesting results faster than their competitors. 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. While around 70% of green biotechnology research is still officially in the hands of the private sector, 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 cannot 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.[9] Not only do sponsors allocate the material resources necessary to the development

[9] 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 public good and the related ‘philosophy of free access’ is evoked by all

u(ambguioustools toclaessfby the spa)5.2(e)-0.2nsois of biain
ousef biinfo(mati inisitbutsy fund)536ked byboithtypes(of)]TJ
sponsos’ isinecessaye inordher oe mkhe ensefia thetwory cr

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 dec fi

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 international data traffic cannot be underestimated. In 2007, following over a decade of investments in this direction, the NSF 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.'[11] 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 can be 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).[12] 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 the Perkin-Elmer Corporation, sponsoring the company Celera headed by Craig Venter, the creator of the shotgun sequencing

[11] From mission statement on the EMBRACE homepage, http://ec.europa.eu/research/health/genomics/newsletter/issue4/article04_en.htm
[12] For general information about the HPG, see the following official website: http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml

techniques that effectively allowed the HGP to keep

judgement that wa

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 vice versa: 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. In addition, 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

tools facilitating data circulation to anyone interested is the most

that it secures (as documented by surveys and website statistics).[15] 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 in two databases collecting 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).[16]

In all these different ways, resource-driven competition becomes a tool towards achieving an array of resources and methods facilitating

[15]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.

[16] 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 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 however impact my argument in this paper, as all agencies agree on treating resource-driven competition as an efficient strategy to circulate data.

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[17]). 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 find in a database.[18]

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.[19] This already constitutes a huge advance over the product-driven competition favored

[17] See Ruttenberg et al 2007.
[18] Another problematic issue, which is not however directly related to resource-driven competition, is the lack of commitments 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.
[19] This point was forcefully advocated by Olson and Green (1998) in the context of the HPG dispute.

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 value of their data.

4. 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.

4.1 Disclosure

There are no general rules in science about how researchers should treat the data that they produce. They can 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 formal 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

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 only one claim, little attention is paid to the format with which data are published. Journals seldom have rules on which format data ought to be reported in a publication, which means that researchers present data in whichever format 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.[20]

[20] 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 at 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, 2006).[21] 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 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.

4.2 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. 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 I hitherto referred to as 'databases,' such as TAIR), 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[22]). 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 the 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 specificre-

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.[23] The exercise of expert skills 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.[24] 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

[23] 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. Though my philosophical research on databases and biological knowledge, I have become reasonably skilled 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.

[24] 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 (2009).

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.

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 become 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 2008).

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

in Radder, ed.

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.[28] Providing means for adequate data circulation maximises the usefulness of research that has already been done and paid for. From a profit-driven perspective, it is just as important to maximise the flow of data across research contexts 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 when 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 strongly 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 exposed by 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 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 representative democracies, is a much more stable and durable entity than a company and can afford to invest capital in projects guaranteed

[28] In this sense, these values constitute good examples of the 'deflated' Mertonian norms proposed by Radder (forthcoming).

Bibliography

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

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

Bammler, T. et al (2005) Standardizing global gene expression analysis between laboratories and across platforms. *Nat Methods* 2, 5:

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

Shoof, H. (2003) Towards interoperability in genome databases: the MAtDB (MIPS Arabidopsis thaliana database) experience. *Comparative and Functional Genomics*, 4: 255-258.

Spannagl et al (2007) MIPSPlantsDB – plant database resource for integrative and comparative plant genome research. *Nucleic Acids Research*, 35: database issue.

Spradling, A., Ganetsky, B., Hieter, P., Johnston, M., Olson, M., Orr-Weaver, T., Rossant, J., Sanchez, A. and Waterson, R. (2006) New Roles for Model Genetic Organisms in Understanding and Treating Human Disease: Report from the 2006 Genetics Society of America Meeting. *Genetics* 172: 2025-2032. (Genet7-0.0005 T10.0TD0 Tc0 Tv

LONDON SCHOOL OF ECONOMICS
DEPARTMENT OF ECONOMIC HISTORY

WORKING PAPERS IN: THE NATURE OF EVIDENCE: HOW WELL DO "FACTS" TRAVEL?

For further copies of this, and to see other titles in the department's group of working paper series, visit our website at:
http://www.lse.ac.uk/collections/economichistory/

2005

01/05: Transferring Technical Knowledge and innovating in Europe, c.1200-c.1800
Stephan R. Epstein

02/05: A Dreadful Heritage: Interpreting Epidemic Disease at Eyam, 1666-2000
Patrick Wallis

03/05: Experimental Farming and Ricardo's Political Arithmetic of Distribution
Mary S. Morgan

04/05: Moral Facts and Scientific Fiction: 19th Century Theological Reactions to Darwinism in Germany
Bernhard Kleeberg

05/05: Interdisciplinarity "In the Making": Modelling Infectious Diseases
Erika Mattila

06/05: Market Disciplines in Victorian Britain
Paul Johnson

2006

07/06: Wormy Logic: Model Organisms as Case-based Reasoning
Rachel A. Ankeny

08/06: How The Mind Worked: Some Obstacles And Developments In The Popularisation of Psychology
Jon Adams

09/06: Mapping Poverty in Agar Town: Economic Conditions Prior to the Development of St. Pancras Station in 1866
Steven P. Swenson

10/06: “A Thing Ridiculous”? Chemical Medicines and the Prolongation of Human Life in Seventeenth-Century England
David Boyd Haycock

11/06: Institutional Facts and Standardisation: The Case of Measurements in the London Coal Trade.
Aashish Velkar

12/06: Confronting the Stigma of Perfection: Genetic Demography, Diversity and the Quest for a Democratic Eugenics in the Post-war United States
Edmund Ramsden

13/06: Measuring Instruments in Economics and the Velocity of Money
Mary S. Morgan

14/06: The Roofs of Wren and Jones: A Seventeenth-Century Migration of Technical Knowledge from Italy to England
Simona Valeriani

15/06: Rodney Hilton, Marxism, and the Transition from Feudalism to Capitalism
Stephan R. Epstein

2007

16/07: Battle in the Planning Office: Biased Experts versus Normative Statisticians
Marcel Boumans

17/07: Trading Facts: Arrow’s Fundamental Paradix and the Emergence of Global News Networks, 1750-1900
Gerben Bakker

18/07: Accurate Measurements and Design Standards: Consistency of Design and the Travel of ‘Facts’ Between Heterogenous Groups
Aashish Velkar

19/07: When Rabbits became Human (and Humans, Rabbits): Stability, Order, and History in the Study of Populations
Paul Erickson and Gregg Mitman

20/07: Contesting Democracy: Science Popularisation and Public Choice
Jon Adams

21/07: Carlyle and the French Enlightenment: Transitional Readings of Voltaire and Diderot
T. J. Hochstrasser

22/07: Apprenticeship and Training in Premodern England
Patrick Wallis

2008

23/08: Escaping the Laboratory: The Rodent Experiments of John B. Calhoun & Their Cultural Influence
Edmund pR8s0 TcPamP