Sabina Leonelli, s.leonelli@exeter.ac.uk
ESRC Centre for Genomics in Society (Egenis)
Byrne House, St German's Road
Exeter EX4 4PJ, United Kingdom

Contribution to the *Handbook of Genetics and Society: Mapping the New Genomic Era*, section 'New Forms of Knowledge'. Routledge, 2009.

Centralising Labels to Distribute Data: The Regulatory Role of Genomic Consortia

Introduction

Given the new opportunities for data-driven research afforded by genomics in conjunction with bioinformatics, biologists and their sponsors have been struggling for agreement on data dissemination strategies. The factors involved in regulating the disclosure and circulation of genomic data range from the conflicting interests and ethos of the researchers involved to the clash in goals and procedures characterising biotechnology and pharmaceutical industries, national governments and international agencies. This situation gives rise to new types of organisations functioning as platforms for networking, debate and joint action among relevant actors. Examining the circumstances in which these organisations emerge, as well as their effects on research practices and regulatory structures, illuminates important aspects of governance in contemporary biomedical research and its effects on knowledge production. This is a case where, in the words of Andrew Barry (2001), the space of governance is being reconfigured, largely as a consequence of adopting new technologies for the production and exchange of data.

In this chapter, I focus on the case of genomic consortia. These are self-organised committees gathering specialists from various scientific fields to negotiate common standards for data dissemination, often with substantial consequences for the regulation of data sharing on a global scale. In particular, I discuss the case of bio-ontology consortia, organisations created to develop and maintain a labelling system for the distribution of data across research contexts. Bio-ontology consortia function as a much-needed interface between bottom-up regulations arising from scientific practice, and top-down regulations produced by governmental and international agencies. They achieve this by focusing on practical problems encountered by researchers who use bioinformatic tools such as databases. A good example is the problem of data classification: that is, the tension that is bound to exist between the stability imposed by classificatory categories used in databases and the dynamism and diversity characterising the scientific practices through which data are produced. Bio-ontology

consortia provide an institutional solution to the problem, by setting up mechanisms to select and update the labels given to data so as to mirror the expectations and needs of data users.

As I intend to show, the study of consortia is crucial to the regulatory impact of data sharing practices on biomedical science. This is because they have become key institutional loci for the actual governance of data sharing: consortia are deliberately created by scientists to develop relevant bioinformatic tools, supervise the implementation of those tools within existing scientific practices, and encourage feedback on those procedures from other researchers as well as top-down regulators. Looking at the issues confronted by consortia, like the problem of data classification, helps to understand what governance in genomic research actually consists of, and how it can be successfully managed.

1. Data-Centric Biology

One of the most important characteristics of contemporary research in the life sciences, and particularly genomics, concerns the status accorded to data as a source of biological knowledge. Thanks to high throughput technologies such as shot-gun sequencing and micro array experiments, data production has become increasingly automated and technology-driven. As a consequence, since the 1990s the activity of data gathering has acquired relative independence from other scientific activities such as hypothesis-testing and explanation. Genome sequencing projects have given new legitimacy to the idea of data as prime motors of research, which are gathered in and by themselves rather than in order to corroborate existing hypotheses. The underlying belief is that it is now possible to obtain scientific knowledge from the (statistical) analysis of data, rather than from the testing of hypotheses: research can be data-driven as well as hypothesis-driven. In the words of a scientific commentator,

> 'if it becomes cheaper to just collect all data required than to run after a hundred consecutive, plausible, but wrong hypotheses, starting with a hypothesis becomes an economic futility' (van Ommen 2008: 1).

Most biologists would disagree with the idea that data can fully replace hypotheses: both elements are recognised as playing an indispensible role in research (e.g. Kell and Oliver 2003). Still, many researchers are discussing the possible re-birth of the inductive method in biology.[1] Not only are biologists flooded with more data than ever before, as high-throughput technologies can produce several billions of data per day: they are also prepared to use those data as reliable evidence, at least as long as no data of better quality are available and standards are developed to enable the comparison of data obtained in diverse experimental circumstances.[2] The accumulation of masses of data on biological entities and processes is widely seen as necessary to improving current understandings of those entities and processes.

This renewed trust in inductive inferences constitutes a powerful intellectual movement, which might be dubbed data-centric biology. Underlying the decision to invest on high-throughput

[1] See the controversies on the role of induction surrounding Holliday (1999) and Allen (2001) in *BioEssays*.
[2] For a study of the development of standards to enable data analysis on a large scale, see Rogers and Cambrosio (2007) on microarrays.

technologies in the first place, for instance, is a fascination with the power of evidence as a potentially 'objective' ground for decision-making. The idea of relying on hard data to verify or refute scientific hypothesis is very much alive within scientific and policy circles, where Popper continues to be hailed as the philosopher of science who most closely captured the features of scientific research (see again the discussion surrounding Holliday 1999). In this context, high-throughput data constitute another opportunity to provide firm footing for the life sciences, notoriously seen as 'softer' than the physical sciences and fraught with uncertainty. The use of data-centrism as guarantor of a more 'objective' science is also evident in medicine, where evidence-based methods promise to substitute the tacit, potentially untrustworthy expertise of doctors with 'hard facts' obtained through randomised clinical trials (Lambert 2006). The prioritisation of evidence derived from clinical trials as the most credible (objective) source for medical knowledge bears obvious similarities with the prioritisation of genomic data obtained through high throughput technologies as the most credible (objective) source for biological knowledge.

Data-centrism is driven by a multiplicity of factors, ranging from the changing structure and resources available to the biomedical sciences (Keating and Cambrosio 2003), to the availability of new technologies (Gaudillere and Rheinberger 2004), the commodification of academia (Leonelli forthcoming) and fundamental advances in the scientific understanding of mechanisms of heredity. Further, data-centric biology has proved its scientific worth in several ways. One of its most significant applications consists in the development of so-called 'model organism biology'. Organisms have always been crucial to experimental research in biology, and typically various branches of the life sciences used different organisms depending on the issues that they investigate and the facilities at hand (developmental biology, for instance, has long focused on chicks because of the ease of keeping eggs in a laboratory and the large size of the embryos, which make them easy to study). More recently however, the notion of model organism has changed connotations. Popular organisms such as fruit-flies (*Drosophila melanogaster*), thale cress (*Arabidopsis thaliana*), yeast (*Saccharomyces cerevisiae*) and mice (*Mus musculus*) are now seen as boundary objects through which various biological disciplines can meet and cooperate; and cooperation comes first and foremost through the accumulation of data on various aspects of their biology, starting with their DNA sequence (Ankeny 2007, Leonelli 2007). Focusing on the study of one species, rather than of several species at once, enables large groups of researchers to channel their efforts into gathering data on virtually every aspect of the same organism. It is expected that these data can then be used as evidential platform to understand the biology of the organism as a complex whole, as well as drawing accurate comparison across species representing different families. As put by the principal investigator of The Arabidopsis Information Resource, a database serving a user community of over 16.000 biologists:

> 'my long-term goal is to discover the rules and mechanisms underlying the workings of *Arabidopsis thaliana* by building an infrastructure to bring all the available data together, developing computer programs that infer knowledge based on the available data, and engaging the research community to test the inferences'.[3]

In what follows, I reflect on what it means to build infrastructures 'to bring all available data together', and on the effect that the development of these infrastructures is having on the regulation of biomedical research. In other words, I wish to focus on the institutional impact of

[3] Sue Rhee website, accessed August 2007: http://carnegiedpb.stanford.edu/research/research_rhee.php.

data-centric biology, by examining the collaborative work needed to develop and implement tools for data sharing.

2. Data Sharing through Bioinformatics

In the words of Paul Wouters and Colin Reddy, 'the increasing role of huge data sets in scientific research has important implications for the way the research is conducted, for the way it should be organised and funded, and for the training of new researchers' (2003: 13). Several international organisations, most prominently the Organisation for Economic Cooperation and Development, have argued that advances in biomedical research depend on scientists' ability to consult and use all available data, independently from where they were originally produced: data sharing on a global scale is the best way to 'advance science for the public good' (Azberger et at 2004: 1777).[4] This view has been adopted by all the main funding bodies for scientific research, including the National Science Foundation and the National Institute of Health in the United States, the European Union and research councils in the UK, Germany, France and Japan. The assumption underlying this policy is that the more scientists are allowed to access the same sets of data, the more those data will be used to produce new knowledge about biological phenomena (Leonelli 2008). The special status accorded to data as sources of knowledge has become a strong argument for making them available to any interested researcher without restrictions.

Indeed, governmental bodies, scientists and, increasingly, industry agree that the efficient re-use of data presupposes data sharing on a global scale (thus settling on a 'politics of coordination' of the type outlined by Stemerding and Hilgartner 1998: 60). What remains contentious is the goal itself: whether all data should be made available for re-use in new research contexts. The main argument underlying the public disclosure of results from the Human Genome Project was that data produced through governmental funding are a public good, helping scientists to produce innovations for the benefit of all, and that therefore no restriction should be placed on accessing them (Sulston and Ferry 2002). This view is still controversial among scientists, policy-makers and both private and public research sponsors. Even aside from issues of privacy arising from the dissemination of data obtained from human subjects (e.g. Martin 2001, Gibbons 2008), there are concerns about intellectual property. Data are hard-won resources obtained through substantial investment. To data producers, giving data away means losing competitive advantage over research groups working on similar topics. Data sharing practices thus challenge the competitive nature of research in both academia (with its 'publish or perish' policies of grant allocation) and industry (where the goal is to be the first to license a marketable product).

The development of tools facilitating data sharing is a great challenge in itself. Given their prominence as criteria to measure the quality of scientific research and allocate funding, publications in journals remain the preferred way to disclose results. However, this system is not ideal as a platform for data sharing. It requires scientists to select the data that best fit the claims published in each paper. This means that the majority of the data produced in each experiment is either kept within the walls of one laboratory or discarded outright. The problem is exasperated by the growing size of biomedical research and related publications, currently split in countless subfields ranging from cell-to-cell transmission to theoretical ecology. Given

[4] See also OECD Guidelines for Data Sharing (2007).

the vast amounts of journals supporting each of these subfields, the chance of published data being noticed and used across different research contexts is extremely small, as researchers are barely able to keep up-to-date with publications within their own area.

A recent solution to these issues has been to introduce online databases as tools to search, retrieve and analyse datasets. The field of bioinformatics has become the main provider of digital tools to store and distribute data online. The resulting ensemble of databases, software and web services used for data sharing, often referred to as *cyberinfrastructure*,[5] is generously supported by governmental bodies throughout the developed world, including the National Science Foundation and the European Union.[6] There are different types of databases in use, ranging from data repositories, holding vast amounts of data of the same type (e.g. GenBank, holding sequencing data from over 260.000 organisms) to community databases, storing data of various different types gathered on the same organism or process (like the above-mentioned Arabidopsis Information Resource, which provides access to 'the complete genome sequence along with gene structure, gene product information, metabolism, gene expression, DNA and seed stocks, genome maps, genetic and physical markers, publications, and information about the Arabidopsis research community'[7]). Thanks to their capability to store huge amounts of information, their ease of consultation over the internet and their flexibility in letting users choose among search parameters, databases constitute ideal means to communicate data to a wide audience. Stephen Hilgartner has highlighted the function of these tools as a 'new communication regime', which complements traditional ways of communicating results by focusing specifically on the dissemination of data (Hilgartner 1995). Community databases are especially interesting, as many of them enable not only the retrieval, but also the analysis of data through apposite models and the perusal of information concerning the original sources of data.[8] Further, databases can be easily linked with one another, enabling cross-searches across a widening network of datasets.

3. Confronting The Problem of Classification: Bio-Ontologies

The increasing use of bioinformatic tools for data sharing has had large effects on the governance of science, many of which I cannot discuss here. I wish to focus on one area whose regulation is of primary interest to scientists, while at the same time requiring the attention – and, arguably, the intervention - of other relevant stakeholders such as funding bodies, governmental organisations and industry. This is the process underlying the choice of categories for data classification to be implemented across databases.

The fast development of cyberinfrastructure has brought a new urgency to the need for common criteria under which data gathered across a variety of contexts can be classified and retrieved. Categories appropriate for this task need to be intelligible and useable by any researcher interested in re-using genomic data stored in databases. This is an extremely challenging requirement, however, given that biology is a highly fragmented field,

[5] See for instance Stein 2008 and Buetow 2005. Lee et al (2006) provide an excellent analysis of the 'human infrastructure' of cyberinfrastructure.
[6] A typical example is the website set up by the National Science Foundation as a virtual 'Office for Cyberinfrastructure', with news on projects and apposite funding (http://www.nsf.gov/dir/index.jsp?org=OCI).
[7] TAIR Homepage, accessed 24/09/2008 (http://www.arabidopsis.org).
[8] Such information, usually referred to as meta-data, is crucial for users to assess for themselves the trustworthiness and evidential value of data found in a database (Leonelli 2008).

encompassing numerous epistemic cultures with diverse commitments, interests, research methods, and tacit knowledge (Knorr Cetina 1999). Differences among epistemic cultures may depend on the disciplinary or geographical location of the researchers involved, on whether or not they use specific technologies or model organisms (and which ones), or on the context in which they work. These differences may shift very rapidly depending on the alliances developed to study a specific topic: epistemic cultures can form or dissolve on the basis of which projects are funded and which collaborations prove useful in the long term. Despite their fragility, substantial differences among epistemic cultures often manifest themselves through the elaboration and use of local terminologies, shaped by shared tacit knowledge and interests, to refer to biological objects and practices. For instance, what ecologists see as a symbiont might be classified as a parasite by immunologists; and molecular and evolutionary biologists often attribute different meanings to the term 'gene' (Stotz, Griffith et al 2004).

Making information travel across communities using very different languages is no small feat. The use of language is unavoidable when trying to classify enormous masses of data for dissemination – and here we find ourselves on familiar territory to historians and social scientists, that is the issue of classification and its use in standardisation practices. Classifying data for dissemination means finding categories that can accommodate the diversity characterising users' research practices. At the same time, a classification system needs to retain some internal consistency in order to provide access to all available data at once. This is true also in the case of multiple classification systems giving access to data obtained in different areas: these systems need to be linked with each other, so as to enable comparative and cross-field searches, which again requires some level of unification among the categories and standards used to classify data. Further, as highlighted by several STS scholars[9], any classification system has a stabilizing force. Stable categories are needed to search and retrieve data from databases. Yet, this requirement also clashes with the practices of the user community: research at the bench is typically quick to produce new results, some of which help to overturn previously held knowledge. As observed in a recent review on standards in bioinformatics, 'this is one of the most important general problems in building standards for biology – our understanding of living systems is constantly developing' (Brazma et al 2006: 595).

An appropriate classification for data sharing needs to be dynamic enough to support the ever-changing understanding of nature acquired by the biologists who use it; at the same time, it needs to retain enough stability to enable data of various sorts and significance to be quickly surveyed and retrieved. Can data classification through standard categories enable collaborative research without at the same time stifling its development and pluralism? In the late 1990s, biologist Michael Ashburner proposed an answer to this question in the form of a functional approach to data classification. As I will argue, the success of this approach is due as much to its technical characteristics as to its institutional implementation; indeed, the development of this system required the creation of a new set of organisations bringing together researchers from different communities and serving as an interface between bioinformaticians, researchers using data and regulatory bodies such as funding agencies and international organisations.

[9] Most notably Bowker and Star (1999).

Ashburner's idea, first presented at the Montreal International Conference on Intelligent Systems for Molecular Biology in July 1998, was to classify data on the basis of the biological entities and processes that genomic data were used to research. This view was born out of Ashburner's experience in developing one of the first community databases (FlyBase, for the fruit-fly *Drosophila melanogaster*) and was shared with many other database developers interested in serving 'not just organism-specific communities, but also pharmaceutical industries, human geneticists, and biologists interested in many organisms, not just one' (Lewis 2004: 103.2). In practice, it involved classifying data gathered on each gene according to the known molecular function and biological role of that gene. The implication was that the terms used for data classification should be the ones used by biologists to describe their research interests, i.e. terms referring to biological phenomena. Thus, for instance, a database user wishing to investigate cell metabolism should be able to type 'cell metabolism' into a search engine and retrieve all available genomic data of relevance to her research.

This approach was implemented as a 'ontology': a strategy for ordering and storing information already popular in computer science and information technology, which enables programmers to produce a formal representation of a set of concepts and of the relationships among those concepts within a given domain. The choice to use the word 'ontology' has little to do with the long tradition in the philosophical study of being. Rather, it has to do with signalling to other scientists and to the world that what is at stake in the development of labels is the very core of scientific and technological innovation: that is, the map of reality used by scientists to coordinate efforts and share resources. This map clearly needs to be drawn on the basis of pragmatic considerations, rather than theory or ideology.[10] When applied in the biological domain, each concept is used to refer to an actual biological entity and, at the same time, to classify available data. Thus were born the so-called 'bio-ontologies', defined as 'formal representations of areas of knowledge [..] that can be linked to molecular databases' (Bard and Rhee 2004: 213) and thus can be used as classification systems for data sharing and retrieval.

The first bio-ontology to achieve prominence among databases was the Gene Ontology, or GO. The GO was developed as a standard for the classification of gene products. It encompasses three different ontologies, each of which mapping a different set of phenomena: a *process* ontology describing 'biological objectives to which the gene or gene product contributes' (Ashburner et al 2000: 27), such as metabolism or signal transduction; a *molecular function* ontology representing the biochemical activities of gene products, such as the biological functions of specific proteins; and a *cellular component* ontology, referring to the places in the cell where a gene product is active (nuclear membrane or ribosome).[11] The GO is now incorporated into most community databases for model organisms, including WormBase, the Zebrafish Information Network, DictiBase, the Rat Genome Database, FlyBase, The Arabidopsis Information Resource, Gramene and the Mouse Genome Database. Many other ontologies have appeared since its creation. Some are devoted to classifying data gathered on a given type of object, such as the Cell Ontology or the Plant

[10] As remarked in a recent review of standardisation efforts in model organism biology, 'there is a considerable difference between building a 'perfect ontology' for knowledge representation, and building a practical standard that can be taken up by the entire community as a means for information exchange. If the ontology is complex, it is unlikely that the wider community will use it consistently, if they use it at all' (Brazma et al 2006: 601).

[11] All three GO ontologies are designed to represent the processes, functions and components of a generic eukaryotic cell; at the same time, they can incorporate organism-specific features (GO includes data from over 30 species). See Ashburner et al (2000) and the Gene Ontology Consortium (2004, 2006 and 2007).

Ontology; other ontologies focus on data gathered through specific practices, such as the Ontology for Clinical Investigation and the Ontology for Biomedical Investigation (Smith et al 2007).

To understand the success of bio-ontologies, it is important to keep in mind that they were explicitly created to confront the challenge of classification: that it, to serve the diverse and shifting interests of database users, while at the same time efficiently enabling data retrieval. This is evident when considering some of the processes through which bio-ontologies are created:

- the process of *selection* of terms to be used as keywords for data searches. Terms are chosen for their popularity within current scientific literature and thus their intelligibility to potential users of databases. As emphasised by the founders of the Gene Ontology, 'One of the factors that account for GO's success is that it originated from within the biological community rather than being created and subsequently imposed by external knowledge engineers' (Bada et al 2004). The developers of bio-ontologies, appropriately called 'curators', are encouraged to pick terms in use within current scientific literature, so as to make sure that they do not insert their own pet terms in the classification system. Further, they are responsible for compiling lists of synonyms for each term, so that research communities using different terms for the same entity would still be able to access the wished-for data.

- the process of *definition*, that is the specification of what the terms actually mean. Bio-ontologies are not simply lists of keywords. Rather, they are 'controlled vocabularies'**:** collections of terms whose definitions and relations to each other are clearly outlined according to specific rules. The meaning of each term is unambiguously fixed via a definition in which curators specify the characteristics of the phenomenon which the term is intended to designate (Baclawski and Niu 2006, 35). For instance, the GO defines the term 'ADP metabolism' as 'the chemical reactions and pathways involving ADP, adenosine 5'-diphosphate'.[12] The definition of terms matters greatly to the success of bio-ontologies as classification systems: researchers can only make sense of data retrieved through a bio-ontology term if they know precisely which entity that term refers to.

- the process of *mapping* terms to specific datasets. This is not an automatic process, as datasets do not come with a ready-made tag indicating the research areas in which they might prove relevant - and thus the bio-ontology terms that will be used for their classification. The mapping of data to bio-ontology terms needs to be done manually, through a process called 'annotation', either by curators or by data producers themselves. Again, how curators or experimenters decide to label data has a strong impact to the functioning of bio-ontologies for data retrieval: database users need to trust that the data classified under a specific term are actually relevant as evidence for the investigation of the related phenomenon.

The ways in which the processes of selection, definition and annotation of terms are carried out are crucial to the success of bio-ontologies as tools for data sharing. This is also because all three processes are subject to constant revision. Curators are well aware that the

[12] GO Website, accessed 24/09/2008 (http://www.geneontology.org/).

knowledge captured by bio-ontologies is bound to change with time and further research, as well as manifesting themselves differently in each research contexts. The advantage of bio-ontologies as digital tools is that they can be updated to reflect developments in the relevant scientific fields. They are built to be a dynamic, rather than a static classification system:

> 'By coordinating the development of the ontology with the creation of annotations rooted in the experimental literature, the validity of the types and relationships in the ontology is continually checked against the real-world instances observed in experiments' (Bada et al 2004: 237**).**

By stressing flexibility to users with diverse epistemic cultures as well as to shifts in scientific knowledge, it may seem that bio-ontologies succeed in solving the problem of classification presented above. As I emphasised, however, whether bio-ontologies can actually match the challenge depends on the way in which they are developed and maintained - that is, on how curators select and update the labels used in bio-ontologies to mirror the expectations of their users.

4. Bio-Ontology Consortia: Institutionalising Collaboration

Curators carry out their work through a combination of skills. They need to have a good understanding of cutting-edge information technology, which enables them to collaborate with programmers and computer engineers in developing appropriate software.[13] They also need a basic training in various biological disciplines: without a generalist understanding of at least two or three different disciplines (for instance developmental and molecular biology), curators would not be aware of the differences among the epistemic cultures characterising different fields and would not be able to build bridges between them. Most importantly, a curators' expertise needs to include some familiarity with experimentation 'at the bench'. This provides curators with an awareness of what users need, expect and look for in a database; it also enables them to understand the experimental settings in which the data have been obtained, which helps the process of annotation.[14] Finally, curators need to endorse a 'service' ethos: they must embrace the idea that their work is meant to facilitate experimental research, which often implies that their contribution is perceived as 'second-class', rather than as complementary to it. In the current scientific retribution system, services such as bioinformatics and database building are not yet fully recognised as forms of research in their own right (Howe, Rhee et al 2008). As many other innovations, bio-ontologies were started through the efforts of personalities already predisposed to collaborative work and willing to put their own career at risk to develop new modes of research. Of course key motivations for entering this profession include self-interest and the desire to expand one's career and impact. Still, it is important to note that the 'service ethos' is frequently emphasised within

[13] As nicely documented by Goble and Wroe (2004), there is actually much tension between computer scientists and biologists on the criteria and priorities to be adopted in bioinformatics. At least within the bio-ontologies sanctioned by the Open Biological Ontology Consortium, the priority is clearly given to biologists: as the ultimate users of the tool being produced, they should be the ones determining how it is developed.

[14] 'One of the strengths of the GO development paradigm is that development of the GO has been a task performed by biologist-curators who are experts in understanding specific experimental systems: as a result, the GO is continually being updated in response to new information' (Hill et al 2008).

curator circles, with normative effects on what counts as good behaviour in those communities.[15]

Paradoxically, the very expertise that enables curators to develop and maintain a database constitutes an obstacle to the communication between curators and database users. Many researchers do not have the skills to provide feedback to curators on how well their systems serve their research. Providing feedback unavoidably means engaging with the practices through which bio-ontologies are developed, and thus acquiring some of the skills involved in curation. Understandably, given the time, interest and effort involved, this is something that database users are often reluctant to do. A molecular biologist I interviewed in March 2007 summarised the problem as follows: 'biologists want to get information and then go back to their question'. To researchers subscribing to this view, the elaboration of bio-ontologies is not a matter of democratic 'voting' on which terms and definitions to adopt, but a matter of division of labour between people busy with experiments and people busy with developing databases storing the results of experiments. In their eyes, the production of a reliable bio-ontology is the job of curators; all they need to do is trust the curators' judgement.

The difficulties in obtaining feedback from users constitute a serious problem. Left to their devices, curators carry out the processes of selection, definition and annotation on the basis of their own perception of what is happening in experimental research. Despite their skills, their judgement is unavoidably one-sided and risks representing research in ways that do not match actual practice – thus betraying the trust of database users. This is especially true when it comes to updating existing bio-ontologies, an activity that benefits tremendously from direct consultation with users. A solution to these issues comes in the form of the institutional set-up in which bio-ontologies are developed and maintained. In order to develop labels that would effectively serve the diverse needs of data users, scientists from various disciplines have joined forced with bioinformaticians and research sponsors to create *genomic consortia*. Consortia are typically born out of the initiative of database developers, who are well aware of the issues surrounding data classification and who decide to join forces in order to bring visibility to those issues within the wider scientific community. Over and above more traditional scientific institutions and funding bodies, it is these organisations that have taken responsibility for enforcing collaboration and dialogue among curators as well as between curators and users – and that therefore play a crucial role in the regulation of data sharing.

An exemplary case is the Gene Ontology Consortium, which was instrumental to the development and current success of the GO as a classification tool. The GO was not started through appositely allocated funds, but rather through an informal network of collaboration among the curators of prominent community databases such as FlyBase, Mouse Genome Informatics and *Saccharomyces* Genome database (Lewis 2004). These researchers were aware that the problems emerging in relation to data classification could not be solved by single-handed initiatives. They therefore decided to use some of the funding allocated to each of their databases to support an international collaboration among database developers, which they named GO Consortium. Institutionalising their informal network into an independent organisation served several purposes: it allowed them to attract funding

[15] As GO co-funder Suzanna Lewis observes: 'careers are measured by the success of the project and the strengths of an individual's contribution to the project's goals. This attitude allowed us to remove both out egos and our concerns for individual recognition from the search for a solution to the data-interconnection problem' (2004: 103.3).

specifically supporting this initiative; it gave visibility to their efforts; and gave other curators the opportunity to join in. Within little less than a decade, the GO Consortium was able to attract funding from both private and public agencies (e.g. a pump-priming grant by AstraZeneca, a major pharmaceutical company, in 1999 and a grant by the National Institute of Health in 2000, which has been hitherto renewed), enabling to fund four full-time curators to work in the Consortium's main office in Cambridge. At the same time, the consortium expanded to incorporate several new databases as members of the consortium (Bada et al 2004). Today, the consortium includes over fifteen members, each of which is required to 'show a significant and ongoing commitment to the utilization and further development of the Gene Ontology' (GO website, accessed 25 November 2008). This implies funding some of their staff to work on GO and contribute to its content, sending at least one representative to GO Consortium meetings (held on average twice a year), and being prepared to host those meetings at their own institutions.

A different example is provided by the Open Biomedical Ontology Consortium, an umbrella body for curators involved in the development of bio-ontologies that was started by Ashburner and Lewis in 2001. The initial motivation was to develop criteria through which the quality and efficiency of bio-ontologies as classificatory tools could be assessed and improved. These include open access to data (with exceptions in the case of sensitive data such as derived from clinical trials); active management, meaning that curators would be constantly engaged in improving and updating their resource; a well-defined focus, which would prevent redundancies between ontologies; and maximal exposure to critiques, for instance through frequent publication in major biology journals (thus advertising the ontology and attracting feedback from potential users) and the establishment of mechanisms to elicit comments from users. Not incidentally, these are also 'the key principles underlying the success of the GO' (Smith et al, 2007: 1252). In other words, the OBO Consortium set out to make the GO into an exemplar in the Kuhnian sense: a textbook example of what a bio-ontology should be and how it should function, a 'model of good practice' (Smith et al 2007: 1253). At the same time, the OBO consortium used the feedback gathered through interaction among curators to develop rules and principles that could be effectively applied to ontologies aimed at different types of datasets.[16] The GO itself ended up being substantially reformed as a result of this process.

Within six years, over sixty ontologies becoming associated with the OBO Consortium (where association involves similar requirements for collaboration as membership in the GO Consortium), and many more curators learning from the experiences gained through these co-operations. The participating ontologies range from the Foundational Model of Anatomy to the Cell Ontology, the Plant Ontology and the Ontology for Clinical Investigations. In several cases, each of the participating ontologies also maintains its own consortium (for instance, in the case of the Plant Ontology Consortium), which again helps curators to interact with experts in the specific fields addressed by the bio-ontology. Governmental agencies have started to pay close attention to the efficiency with which consortia operate, and to reward it by allocating apposite funding.

[16] This is particularly relevant to the OBO Foundry, a subset of OBO ontologies whose curators are actively engaged in testing and developing further rules for ontology development (Smith et al 2007).

The main function of bio-ontology consortia such as the OBO and the GO is to effectively *enforce collaboration* across three main groups involved in the regulation of data sharing.[17] The first group comprises *database curators*. Consortia provide an institutional incentive for the exchange of ideas, experiences and feedback among curators busy with different projects, thus speeding up developments in bio-ontology curation, enhancing curators' accountability to their peers, increasing effective division of labour among curators and at the same time helping to maintain and legitimise a collaborative ethos. Exchanges are achieved through regular face-to-face meetings and weekly communications through various channels, ranging from old-fashioned emails to wikis, blogs and apposite websites (such as the BioCurator Forum[18]). Indeed, consortia play a key role in training curators: it is through consortia that curators discuss what counts as expertise in bioinformatics, and many consortia organise training workshops for aspiring curators. Last but not least, consortia promote the interests of curators and their rights as researchers, much as a workers' union would do. The status of bioinformatics within biology is on the rise (Howe, Rhee et al 2008), a fact that, among many other factors, has at least something to do with the visibility obtained by curators through savvy steering of consortia.

The second group consists of the many *stakeholders* with an interest in data sharing practices and their regulation, encompassing for instance research sponsors, publishers and journal editors. Consortia function as a platform for communication between curators and the outer world. For instance, several consortia are engaged in dialogue with industry, in an attempt to align data classification practices in that context with the practices characterising research that is publicly sponsored. Also, the GO Consortium – and particularly curators from The Arabidopsis Information Resource – has started to collaborate with the editors of *Plant Physiology,* the foremost journal in plant science (Ort and Grennan 2008). This involves asking researchers who submit a paper to the journal to disclose related data through The Arabidopsis Information Resource, which in turn involves making use of GO labels to classify the data. Both curators and editors hope that this mechanism will force experimenters (who certainly need to publish) to learn more about how bio-ontologies work, thus enhancing their ability to provide feedback to database curators, as well as their interest in doing this.

The third group whose ability to intervene on data sharing regulation is massively increased by consortia is, of course, the vast and diverse community of *database users*. Notably, communication through consortia transcends local commitments such as national culture, geographic position or disciplinary training. Consortia such as the OBO are intended to provide a space for confrontation among all users of bio-ontologies, regardless of their affiliation or location. Consortia attempt to increase feedback through apposite mechanisms for communication between curators and users. An effective mechanism is the so-called 'content meeting', a workshop set up by curators to discuss specific bio-ontology terms in the presence of experts from several related fields. For instance, the GO Consortium organised a GO Content Meeting at the Carnegie Institution's Plant Biology Department in 2004, in which the GO terms 'metabolism' and 'pathogenesis' were critically discussed and redefined with the help of experts in immunology, molecular biology, cell biology and ecology. Similar to this are so-called 'curator interest groups', in which users are invited to provide feedback on specific ontology contents; and online discussion groups coordinated through wikis or blogs.

[17] Lewis describes effective collaboration as an 'unforeseen outcome', yet points to it as 'the single largest impact and achievement of the Gene Ontology consortium to date' (2004: 103.3).
[18] See http://www.biocurator.org.

Some consortia have also discussed implementing peer review procedures on each annotation process, thus asking two referees from the bench to assess the validity and usefulness of specific bits of curators' work. This procedure, though time-consuming, might become popular especially for complex annotations relating to pathways or metabolic processes. Last but not least, consortia forcefully promote user training, both via workshops at conferences and in home institutions and by pushing the insertion of bioinformatic courses within biology degrees, often already at the undergraduate level. More than any other factors, this influence on science education is likely to reduce the gap in skills and interests currently separating curators from users.

The more success bio-ontologies enjoy as efficient tools for data sharing, the more power consortia acquire to enforce collaboration among all stakeholders in data sharing and to influence top-down regulations from governmental and international agencies to fit their agenda. The timing of these organisations is remarkable, as their activities fit squarely into the broader shift in science policy towards the use of cyberinfrastructures as main tools for collaborative and interdisciplinary work (Lee et al 2006, Stein 2008).[19]

Conclusion: Centralising Regulation to Distribute Data

By focusing on the bottom-up institutionalisation of bio-ontologies, I have shown that the process of labelling data for dissemination is both an outcome of and a platform for the regulation of data sharing practices. The use of bio-ontologies and databases is meant to increase the fluidity of communication and the ease in exchanging resources among scientists. In practice, this can only happen within the right institutional setting. A key concern in governing data sharing is to enforce effective dialogue among the developers and the users of bio-ontologies, so as to secure the trustworthiness and usability of these classificatory systems. In particular, the developers of cyberinfrastructure have made themselves accountable for developing systems to elicit users' feedback and act upon it. Bio-ontology consortia have emerged from the deliberate, reflexive efforts by curators to collaborate towards the improvement of data classification processes.

I have argued that bio-ontology consortia play an important role in developing, maintaining and legitimising practices of data sharing. Consortia operate as collectives gathering relevant stakeholders and forcing them to interact with each other. Regulatory measures emerge from consensus achieved through frequent confrontation among different parties. As pointed out by Cambrosio et al (2006: 193), this kind of consensus does not have to concern all aspects of scientific work, but rather the modalities of use of technologies that need to be shared across large and diverse communities. Further, consensus is conceived pragmatically as a temporary achievement, which needs to be frequently challenged and revised through the expression of diverse viewpoints. Explicitly formulated dissent among different epistemic cultures is necessary to maintaining such a dynamic set of conventions. Like many other organisations devoted to the regulation of cyberinfrastructure, bio-ontology consortia construe themselves as platforms to voice the epistemic diversity characterising local research cultures. In their attempts to classify data for dissemination, consortia exemplify the need for

[19] This shift is also discussed by a growing literature on 'collaboratories' (intended as 'laboratories without walls'), as exemplified by Bafoutsou and Mentzas (2002) and Finholt (2002).

forms of scientific governance where, as in other realms of social life, centralisation processes at once emerge from and fuel diversity.

By fostering consensus through the acknowledgment of epistemic pluralism, consortia are making a political move: they are proposing themselves as *regulatory centres* for data sharing processes. As I have shown, they indeed play a central role in shaping the *expertise* required to build and maintain tools for data sharing. They are also centralising *procedures*, as demonstrated by their attempts to establish common rules for bio-ontology development. And they promote common *objectives* for the whole scientific community, such as the willingness to integrate the tools used to share materials and resources from which knowledge can be extracted (resources such as data, but also tissue samples, in the case of bio-banks, or specimens, in the case of natural history collections or stock centres for model organism research). In her reflections on pre-GO attempts to integrate community databases, Lewis notes how collaborations set up in the absence of a common focus ended up in failure (2004: 103.2). Further, the unity of purpose characterising consortia involves a self-appointed sense of responsibility for the regulation of data sharing practices, which remains largely uncontested by users and enables consortia to attract funding and support.

This type of centralisation has several epistemic and institutional advantages. It enhances the power of labels and standards to cross epistemic contexts; it enables constructive dialogue between curators and users of bioinformatic tools; and it favours the cooperation between academia, governmental agencies and industry towards disclosing and disseminating data. The coordinators of the OBO consortium acknowledge the diversity of expertises and stakes in genomic research, as well as the need for data users to work within their own network and their local epistemic culture:

> ‘Our long-term goal is that the data generated through biomedical research should form a single, consistent, cumulatively expanding and algorithmically tractable whole. Our efforts to realize this goal, which are still very much in the proving stage, reflect an attempt to walk the line between the flexibility that is indispensable to scientific advance and the institution of principles that is indispensable to successful coordination’ (Smith et al 2007: 1254)

At least in principle, the centralisation of regulatory power in the hands of consortia fosters the distributed and plural nature of biological research. Walking the line between flexibility and stability in regulating data sharing might prove a viable way to confront and finally solve the classification problem. As long as consortia keep up their efforts to walk that line, bio-ontologies have a chance to develop as a dynamic classification system. The solution to the classification problem is therefore institutional as much as it is technological: bio-ontologies provide both the means and the platform to constantly update classificatory categories, while at the same time cultivating epistemic diversity through recourse to the ‘right’ expertises and institutional settings.

It might be objected that the total decentralisation achieved through the ‘wikification’ – also referred to as ‘crowdsourcing’ and ‘community annotation’ (Ledford 2008) - of data sharing constitutes an increasingly popular and effective alternative to the centralisation exemplified by consortia. Within that model, users are free to add their own annotations and corrections to a given gene or pathway through tools such as the Gene Wiki. This is certainly a promising avenue for the involvement of users in the development of databases, and one that consortia

are seeking to exploit. Indeed, wikis should be considered as complementing, rather than substituting, the work of organisations such as consortia. This is because the existence of common terminological standards is a necessary requirement for the very functioning of wikis, as widely acknowledged by defenders of the role of local agency in designing tools for data sharing. Consider for instance the following statement, which concludes a recent survey of the usefulness of distributed agency in the development of ontologies:

> 'Local agency, *when incorporated into the wider design concept*, is increasingly seen as a resource for maintaining the quality, currency and usability of locally generated data, and as a source of creative innovation in distributed networks' (Ure et al 2008: 9; my emphasis).

Interventions by database users are essential to maintaining the efficiency of ontologies as tools for data sharing. Yet, this can only happen in the presence of a 'wider design concept', which the regulatory framework supplied by consortia helps to achieve in a collaborative fashion.

As widely noted within the social sciences, the recognition of cultural diversity and of the need to facilitate inter-cultural communication are key characteristics of governance today. The biomedical sciences are no exception. The community involved in this type of research has never been so large, so geographically dispersed and so diverse in motivations, methods and goals. In such a context, efficient channels of communication are important means to 'make order' (Jasanoff 2008), that is to establish a structure through which individuals and groups can interact beyond the boundaries imposed by their location, disciplinary interest and source of funding. Consortia play an important role in the management and distribution of labour (and accountability) relating to data sharing, as well as in the regulation of data ownership. By making access to bioinformatic tools conditional on the adoption of specific data sharing practices, curators use consortia towards 'the co-production of technical and social orders capable of simultaneously making knowledge and governing appropriation' (Hilgartner 2004, 131). Thus, consortia serve a regulatory function that is complementary to legal frameworks, which are typically constructed by non-scientists and imposed by state agencies rather than emerging from the experience of practitioners.

Last but not least, the regulation of tools and practices facilitating communication among data users might have one additional consequence: to encourage both scientists and policy-makers to question and enrich their understanding of the role of data within research. Thanks to their exposure to diverse epistemic cultures within biology, the curators of community databases and bio-ontologies are acquiring an increasingly sophisticated understanding of the complex array of methods, tools, practical skills and conceptual baggage needed to evaluate the quality of data and to use data towards creating new knowledge. Through institutions such as consortia, data sharing is regulated so as to involve dialogue not solely over data, but also over the theoretical assumptions and tacit knowledge underlying their production and re-use. It remains to be seen whether consortia will continue to voice epistemic diversity and highlight local agency in ways that might help to push biomedical research beyond its current data-centric mode.

Acknowledgments
The inspiration for this piece came from various interviews to database curators and users carried out between 2004 and 2008. I am particularly grateful to the staff at The Gene Ontology Consortium and The Arabidopsis Information Resource. I also thank Alberto Cambrosio for his help in drafting this piece; Mary Morgan, the 'facts' group and my colleagues at Egenis for illuminating discussions; and the audience at the 4S meeting in Rotterdam, August 2008, for excellent feedback. This research was funded by the Leverhulme/ESRC project 'How Well Do 'Facts' Travel?', based at the Department of Economic History of the London School of Economics, and by the ESRC Centre for Genomics in Society of the University of Exeter.

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.

Allen, J.F. (2001) 'Bioinformatics and discovery: Induction beckons again', *BioEssays,* 23, 1: 104-107.

Ashburner, M. et al (2000) 'Gene Ontology: Tool for the unification of biology', *Nature Reviews: Genetics*, 25: 25-29.

Arzberger et al (2004) 'An international framework to provide access to data', Science, 303, 5665:1777-1778.

Baclawski, K. and Niu, T. (2006) *Ontologies for Bioinformatics*, Cambridge, MA: The MIT Press.

Bada, M. et al (2004) 'A short study on the success of the Gene Ontology', *J Web Semant*,1: 235-240

Bafoutsou G. and Mentzas G. (2002) 'Review and functional classification of collaborative systems', *International Journal of Information Management*, 22, 4: 281-305.

Bard, J.B.L. and Rhee, S. (2004) 'Ontologies in biology: Design, applications and future challenges', *Nature Reviews: Genetics*, 5: 213-222.

Barry, A. (2001) *Political Machines. Governing a Technological Society*, London and New York: The Athlone Press.

Brazma, A., Krestyaninova, M. and Sarkans, U. (2006) 'Standards for systems biology', *Nature Reviews: Genetics*, 7: 593-605

Boukwer, G. C. and Star, S. L. (1999) *Sorting Things Out. Classification and Its Consequences*, Cambridge, MA: The MIT Press.

Buetow, K.H. (2005) ‘Cyberinfrastructure: Empowering a "third way" in biomedical research’, *Science*, 308, 5723: 821 – 824.

Cambrosio, A., Keating, P., Schlich, T. and Weisz, G. (2006) ‘Regulatory objectivity and the generation and management of evidence in medicine’, *Social Science and Medicine* 63: 189-199.

Chicurel, M. (2002) ‘Bioinformatics: bringing it all together’, *Nature*, 419: 751-757.

Finholt, T.A. (2002) ‘Collaboratories’, in Cronin, B. (ed) *Annual Review of Information Science and Technology*, Washington, D.C.: American Society for Information Science and Technology, vol. 36: 73-107.

Gaudillière, J.P. and Rheinberger, H.J. (eds) (2004) *From Molecular Genetics to Genomics,* New York: Routledge.

The Gene Ontology Consortium (2004) ‘The Gene Ontology (GO) database and informatics resource’, *Nucleic Acids Research*, 32: D258-D261.

---------- (2006) ‘The Gene Ontology (GO) project in 2006’, *Nucleic Acids Research*, 34: D322-D326.

---------- (2007) ‘The Gene Ontology (GO) project in 2008’, *Nucleic Acids Research*, 36: D440-444.

Gibbons, S. (2008) ‘From Principles to Practice: Implementing Genetic Database Governance’, *Medical Law International*, 9, pp.101-109.

Goble, C. and Wroe, C. (2004) ‘A tale of two households’, *Comp. Funct. Genom.* 5: 623–632.

Holliday R. (1999) ‘The incompatibility of Popper's philosophy of science with genetics and molecular biology’, *BioEssays,* 21: 890-891.

Hilgartner, S. (1995) ‘Biomolecular databases: New communication regimes for biology?’, *Science Communication,* 17: 240-263.

---------- (2004) ‘Mapping systems and moral order: Constituting property in genome laboratories’, in Jasanoff, S. (ed) *States of Knowledge: The Co-Production of Science and Society*, London and New York: Routledge.

Howe, D., Rhee, S. et al (2008) ‘The future of biocuration’, *Nature,* 455, 4: 47-50.

Hill, D.P., Smith, B., McAndrews-Hill, M.S. and Blake, J.A. (2008) ‘Gene Ontology annotations: what they mean and where they come from’, *BMC Bioinformatics,* 9(5): S2.

Jasanoff, S. (2008) ‘Making order: Law and science in action’, in Hackett, E.J. et al (eds) *The Handbook of Science and Technology Studies. Third Edition*, Cambridge, MA: MIT Press.

Keating, P. and Cambrosio, A. (2003) *Biomedical Platforms*, Cambridge, MA: The MIT Press.

Kell, D.B. and Oliver, S.G. (2003) 'Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era', *BioEssays,* 26: 99-105.

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

Lambert, H. (2006) 'Accounting for EBM: Contested notions of evidence in medicine', *Social Science and Medicine,* 62(11): 2633-2645.

Ledford, H. (2008) 'Molecular biology gets wikified', *Nature Online* 23 July 2008 Doi:10.1038/News.2008.971.

Lee, C.P., Dourish, P. and Mark, G. (2006) 'The Human Infrastructure of Cyberinfrastructure', *Computer Supported Cooperative Work Conference (CSCW)*, Banff: Canada, http://www.dourish.com/publications/2006/cscw2006-cyberinfrastructure.pdf.

Leonelli, S. (2007) 'Growing weed, producing knowledge. An epistemic history of *Arabidopsis thaliana*', *History and Philosophy of the Life Sciences,* 29, 2: 55-87.

-------- (2008) 'Circulating evidence across research contexts: The locality of data and claims in model organism biology', *LSE Working Papers on the Nature of Evidence: How Well Do 'Facts' Travel?*, 25/08.

-------- (forthcoming) 'The commodification of knowledge exchange: Governing the circulation of biological data', in: Radder, H (ed) *The Commodification of Academic Research*, Pittsburgh University Press.

Lewis, S.E. (2004) 'Gene Ontology: Looking backwards and forwards', *Genome Biology*, 6, 1: 103.

Martin, P. (2001) 'Genetic governance: the risks, oversight and regulation of genetic databases in the UK', *New Genetics and Society*, 20, 2: 157-183.

OECD (2007) *OECD Principles and Guidelines for Access to Research Data from Public Funding*, http://www.oecd.org/dataoecd/9/61/38500813.pdf

van Ommen, G.B. (2008) 'Popper revisited: GWAS here, last year', *European Journal of Human Genetics*, 16: 1–2

Ort, D.R. and Grennan, A.K. (2008) '*Plant Physiology* and TAIR Partnership', *Plant Physiology*, 146: 1022-1023.

Rogers, S. and Cambrosio, A. (2007) 'Making a new technology work: The standardisation and regulation of microarrays', *Yale Journal of Biology and Medicine,* 80: 165-178.

Smith, B. et al (2007) 'The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration', *Nature Biotechnology,* 25, 11: 1251-1255.

Stein, L.D. (2008) 'Towards a cyberinfrastructure for the biological sciences: progress, Visions and Challenges', *Nature Genetics*, 9: 678-688.

Stemerding, D. and Hilgartner, S. (1998) 'Means of coordination in making biological science: On the mapping of plants, animals, and genes', in Disco, C. and van der Meulen, B. (eds) *Getting New Technologies Together*, Berlin and New York: de Gruyter Studies in Organization.

Stotz, K., Griffiths, P.E et al. (2004) 'How scientists conceptualise genes: An empirical study', *Studies in History & Philosophy of Biological and Biomedical Sciences,* 35, 4: 647-673.

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

Ure, J. et al (2008) 'Aligning technical and human infrastructures in the semantic web: A socio-technical perspective', paper presented at the *Third International Conference on e-Social Science*, http://ess.si.umich.edu/papers/paper188.pdf (revised version forthcoming with the *Journal of the Association for Information Systems*).

Wouters, P. and Reddy, C. (2003) 'Promise and practice in data sharing', in: Wouters, P. and Schröder, P. (eds) *The Public Domain of Digital Research Data*, Amsterdam: NIWI-KNAW.