Published in Social Studies of Science, 2012

When Humans Are the Exception: Cross-Species Databases at the Interface of Biological and Clinical Research

Sabina Leonelli
ESRC Centre for Genomics in Society (Egenis)
University of Exeter
Byrne House, St Germans Road
EX4 4PJ Exeter, UK

s.leonelli@exeter.ac.uk

Abstract

Cross-species comparison has long been regarded as a stepping stone for medical research, enabling the discovery and testing of prospective treatments before they undergo clinical trial on humans. Post-genomic medicine has made cross-species comparison crucial in another respect: the 'community databases' developed to collect and disseminate data on model organisms are now often used as a template for the dissemination of data on humans and as a tool for comparing results of medical significance across the human-animal boundary. This paper identifies and discusses four key problems encountered by database curators when integrating human and non-human data within the same database: (1) picking criteria for what counts as reliable evidence, (2) selecting metadata, (3) standardising and describing research materials and (4) choosing nomenclature to classify data. An analysis of these hurdles reveals epistemic disagreement and controversies underlying cross-species comparisons, which in turn highlight important differences in the experimental cultures of biologists and clinicians trying to make sense of these data. By considering database development through the eyes of curators, this study casts new light on the complex conjunctions of biological and clinical practice, model organisms and human subjects, and material and virtual sources of evidence – thus emphasising the fragmented, localized and inherently translational nature of biomedicine.

Keywords: databases, cross-species research, biomedicine, data, model organism, human subjects.

Acknowledgments

This research was funded by the ESRC as part of the ESRC Centre for Genomics in Society. I am grateful to Alberto Cambrosio for helpful comments on an early draft and to my colleagues in Egenis for their support and feedback. Also, I benefited from comments by audiences at a CRASSH conference in Cambridge in April 2009, an Egenis workshop in Exeter in March 2010 and a Gordon Cain Conference hosted by the Chemical Heritage Foundation in May 2010. Last but not least, I am very thankful for the detailed and insightful comments I received from three anonymous referees on an earlier version of this paper, which made an enormous difference to the resulting manuscript.

'Man is the most unsatisfactory of all organisms for genetic study'

A. Sturtevant 1954

Introduction

STS scholarship has long noted the important role played by online databases within scientific research. Large research efforts have been invested in the construction of databases (Wouters and Schroeder 2003, Hey and Trefhesten 2005, Stein 2008), resulting in major shifts in the ways in which scientific work is organized (Leigh Star and Rhleder 1996, Bowker 2001), ordered (Hine 2006) and communicated (Hilgartner 1995). The dissemination of data through electronic means is now an essential complement to traditional publication strategies and the consultation of databases has become part and parcel of everyday routines within experimental research (Lenoir 1999). The heightened need for specialist skills in computer programming has also affected the division of research labor and the ways in which scientists are trained. While university curricula in natural science are giving new prominence to information technologies, database 'curators' have emerged as a professional figure whose responsibilities lie in developing databases that satisfy the needs of prospective user communities (Leonelli 2009; Baker and Millerand 2010; Chow-White and Garcia-Sanchos, 2011). The impact of online resources is particularly evident within the biological and biomedical sciences, where research communities dedicated to the study of popular model organisms have developed sophisticated databases for the organization, dissemination and comparative analysis of genomic data coming from different species (Leonelli 2010a). These tools are often treated as a model for how cross-species data mining should be organized. This has special relevance for the development of information infrastructures for post-genomic, molecular-based medicine, which requires digital platforms through

which data on human and non-human organisms can be integrated and compared. Indeed, the databases developed within model organism biology have been hailed as critical to 'unlocking the very essence of biologic life and enabling a new generation of medicine' (Buetow 2005). Databases are expected to facilitate the achievement of these ambitious goals by fostering the integration of biological and biomedical knowledge, and thus supporting translational research towards new forms of medical diagnosis and treatment.

This paper investigates the role played by databases within biomedical research through an examination of the practical difficulties encountered by database curators in fulfilling such huge expectations. The hype attached to database development as an easy solution to the current 'data deluge' has taken attention away from the problems involved in actually using data found online towards further research: in particular, from the difficulties of matching *in silico* representations of the world with experimentation *in vivo* and clinical intervention, and aligning the experimental practices characterizing research on humans with the ones used to research model organisms. I here propose to view the process of database development not primarily as a means towards the solution of those problems (thought this might certainly be the case), but rather as an excellent site where diverging stakes, values and epistemologies characterizing experimental cultures in biomedicine can be identified and discussed. The idea for this paper emerged from in-depth conversations which I carried out with the curators of several major sites primarily devoted to collecting and disseminating data on model organisms, including The Arabidopsis Information Resource, WormBase, FlyBase, the Gene Ontology Consortium and the General Model Organism Database. These conversations, which took place between 2004 and 2010 in the context of research on the epistemology of model organisms, brought to my attention the acute discomfort that many curators felt at the perceived shortcomings of their databases. In particular, I found that curators worried about developing databases containing both human and non-human data. In what follows, I analyze some of the sources of that worry and show

how it relates not simply to differences in the biology of the organisms being studied, but rather to differences in how biologists and clinicians conduct research, collect data and interpret their findings.[i]

The use of cross-species databases is an excellent instance of 'biomedicine', defined as the set of practices which brings biological and clinical knowledge and techniques to bear on each other (Keating and Cambrosio 2003). Historians and sociologists of medicine have long pointed to the extensive fragmentation characterizing the epistemic communities involved in biomedical research and have analyzed the complex relations and intersections between them (e.g. Loewy 1986, Quirke and Gaudillière 2008). In particular, STS scholarship has documented how scientists attempt to overcome this pluralism in order to achieve common standards and procedures, for instance in the case of clinical trials (Kohli-Laven et al 2011), the trading of biological data and materials (Parry 2004), and the standardization of microarray experiments (Rogers and Cambrosio 2007). Database curation is another area where the introduction of standards, norms and specific technologies clashes with a highly fragmented and localized landscape of research habits and practices. This holds especially when considering the materials – the organisms - on which data are produced and disseminated. While it is tempting to neatly define clinical research as conducted exclusively on humans, much clinical work involves the collection and use of data acquired on rats, mice and other, more distant relatives of *Homo sapiens* (Davies 2010; Spradling et al 2006). Similarly, biological research largely revolves around few key model organisms, such as the nematode *Caernorhabditis elegans*, the fruit-fly *Drosophila melanogaster*, the plant *Arabidopsis thaliana* and the zebrafish *Danio rerio* (Davies 2004); and yet, biologists do not refrain from using human data whenever needed to further their understanding. Database curators are well aware of their important role in facilitating the comparison and integration of data on humans with data on model organisms. They are also aware that the success of their products depends on how useful they prove to be to experimenters, as this determines the levels of funding and

community support that they will receive. Thus, the career of curators depends at least in part on their ability to identify, embrace and constructively engage as many epistemic cultures in biomedicine as possible; which, in practice, means making their digital representation of data at least compatible with, and at best conducive to, widely diverse forms of intervention on actual organisms. This inescapable commitment makes curators' perspectives into very valuable sources of insight on the interface between biology and medicine and the experimental cultures that characterize those two highly intertwined, and yet distinguishable, realms.

I start my analysis by discussing the emergence of cross-species databases, the conception of model organisms that they embody and the status of human data within them. I emphasize how curators of cross-species databases, who are mostly biologists rather than clinicians, find themselves treating humans as a model organism with no special epistemic status – that is, as yet another species on which we happen to have huge amounts of data that need to be compared with data from other species in order to provide biological understanding. I then single out four technical problems in the development of cross-species databases, which curators view as evidencing potential discrepancies between clinical and biological research practices: (1) the criteria for what counts as reliable evidence, (2) the selection of meta-data, (3) the standardization and description of research materials and (4) the choice of nomenclature used to classify data. In closing, I show how the controversies surrounding these aspects of database development reveal their significance in demarcating, and possibly reinforcing, epistemic differences between the lab and the clinic and between human and non-human research.

Model Organism Databases and the Incorporation of Human Data

Within model organism biology, huge efforts have been invested in database development over the last

two decades (Bult 2006). On the one hand, these investments were certainly fuelled by the growing recognition, across biomedicine as a whole, that collection and dissemination practices affect whether and how data are re-used towards new discoveries (Buetow 2005). On the other hand, the extent to which model organism communities have engaged in database development is linked to their unique historical role in fostering a collaborative ethos within the notoriously competitive culture of biomedical research. Many of the most popular models, including the fruit-fly, the thale-cress, the nematode and the mouse, owe some of their success as laboratory organisms to the collaborative ethos and interdisciplinary ambitions fostered by the scientists who pioneered their use in biology (Kohler 1994, Ankeny 2001, Leonelli 2007, Rader 2004).[ii] This emphasis on sharing was not simply a matter of individual commitment, but was actually essential to the research programs set up by biologists such as T. H. Morgan in the 1920s, Sydney Brenner in the 1960s and Chris Sommerville in the 1970s. Their explicit long-term goal was understanding organisms in all of their complexity through an interdisciplinary approach that would include genetics as well as cell biology, physiology, immunology, morphology and ecology; and their strategy to achieve this was to accumulate and integrate knowledge on the biology of one species, which would then provide a blueprint and reference point for comparative, cross-species research.[iii] Over the last two decades, this attitude has been incorporated into the building of model organisms databases, which are often referred to as 'community databases' to stress their role in serving researchers by gathering and integrating all the information available on a specific organism of interest to them (Rhee and Crosby 2005). These databases, which are freely accessible online thanks to public funding from national and international agencies, have arguably become an important component of the very identity and status of model organisms in research, on a par with other characteristic features such as their capacity to represent other species, their tractability in the lab and the extent to which they embody biological processes of interest. Model organisms in the 21st century are organisms on which much is known, and knowledge

of which can be freely and easily accessed and used to study other organisms (Ankeny and Leonelli 2011). [iv]

The Arabidopsis Information Resource (TAIR), the main database collecting data on *Arabidopsis*, is a good example of the success of this research strategy. Its first Director, Sue Rhee, is an eloquent advocate of the value of sharing resources (an approach she calls 'share and survive', as opposed to the 'publish or perish' mentality characteristic of mainstream biomedical research; Rhee 2004) and set up TAIR as a platform for the collection and dissemination of all information of relevance to the study of *Arabidopsis*. Indeed, a decade after the end of the *Arabidopsis* sequencing project, this database provides access to several different types of data about the plant, as well as rich meta-data, search tools and modeling techniques to analyze the datasets and general information about the history of research on the plant. TAIR also co-operates with the *Arabidopsis* stock centers, which store hundreds of thousands of seed stocks of *Arabidopsis* mutants, so that users can order the specimens needed from their experiments directly from their website (Rosenthal and Ashburner 2002). As a result of these efforts, TAIR has become a key tool in plant science research and is routinely used to research not only Arabidopsis, but any other plant species, including crops. Other examples of well-curated and widely used community databases include FlyBase, WormBase, Mouse Genome Informatics and the Zebrafish Model Organism Database.[v] All of these databases were initially funded by public agencies to disseminate one specific type of data, that is the data coming out of sequencing projects; yet, they took advantage of the funding to build tools potentially incorporating other types of data on the same organism, and have aimed to increase the diversity of the data that they host ever since.[vi]

These efforts to develop databases have resulted in curators acquiring a sophisticated understanding of

the factors that influence the future adoption and use of data collected on model organisms across research contexts. These include the need to integrate different *data types* produced through various kinds of instruments and techniques, ranging from sequence data to photographs or tissue samples; to collect *meta-data* documenting the provenance of data; to develop *representations* of data that facilitate searches and the visualization of results retrieved from databases (e.g. maps, models, simulations); to be able to order the *materials* on which data were originally acquired, such as specimens of the same mutant; and to adopt intelligible *keywords* for the classification and retrieval of data. Curators have successfully proposed themselves as possessing the right skills to perform these complex tasks, and regular Biocurator meetings are now held across the globe to facilitate cooperation and interoperability across different databases (Howe et al 2008). On the basis of their growing expertise and increasing need for comparative analyses, the curators of community databases have recently engaged in the development of several cross-species databases, where existing data on different organisms can be searched, viewed and compared. A well-known initiative of this kind is the Gene Ontology (GO), a bio-ontology developed jointly by the curators of several community databases for the cross-species annotation of gene products (Gene Ontology Consortium 2000). All the curators involved in GO hold a PhD in a branch of experimental biology, and use that expertise to inform their curatorial activities (they know what at least some of their users want, because they have been potential users themselves). The GO currently includes data from dozens of species, including several grains, yeast, slime mold, rat, several microbes and *Homo sapiens*, and is coordinated by a central office at the European Bioinformatics Institute near Cambridge, UK. Its funding depends largely on the public supports provided to the model organism databases involved in its development, which is why curators refer to GO as a 'consortium' of research efforts; further, a small National Human Genome Research Institute (NHGRI) grant supports the activities of GO coordinators in Cambridge.[vii] Another important initiative is the General Model Organism Database project (GMOD), also referred to as the 'myriads' database

because of the sheer number of species that it incorporates. The GMOD project is again the result of an extensive cooperation involving over 100 participating databases, including repositories of human data such as Human 2q33, Chromosome 7 Annotation project, Xmap, Ymap and HapMap. The main goal of the GMOD is to help species-specific databases to coordinate their efforts, so as to guarantee interoperability across databases and thus facilitate cross-species analyses. To this aim, the GMOD encourages database curators to use a common set of software packages, such as tools for browsing and annotating genomes, and to take account of the standards already employed by the main model organism databases when setting up new tools and resources (http://gmod.org/wiki/Main_Page). One of the most used GMOD tools is the Generic Genome Browser, which has been adopted by several databases across the globe, including most community databases as well as HapMap and the Human Genome Segmental Duplication Database (Stein et al 2002). Before proceeding further, it is important to note how the funding allocated to both GO and GMOD compares to the large investments made in disease-specific databases. Cross-species databases are much less well-funded and rely on international cooperation much more than their purely medical counterparts. This feature might seem puzzling given the revolutionary results that these databases are expected to yield, but it becomes easy to understand given the current financial crisis and the pressure on scientific institutions to invest only in projects with short-term, measurable returns. Disease databases are deemed to be directly relevant to finding efficient treatments, and thus receive considerable support from both public and private institutions; while the contributions that cross-species databases could make to medical knowledge are much less clearly defined, and more difficult to understand and advertise to the general public, which results in scarce funding (mostly from public sources).

Thanks to initiatives such as GO and GMOD, several features of the community databases developed within model organism research have been proposed as standards for the online gathering and

distribution of all sorts of biological data, including data on humans. Indeed, the assimilation of human data into model organism databases has been fostered precisely for its potential value to both biological and medical research. The GMOD project itself is coordinated by the Ontario Institute for Cancer Research, where the International Cancer Genome Consortium is also based. Further, many community databases for model organisms, including the GMOD and GO projects, are sponsored by the National Human Genome Research Institute, with the following motivation:

> 'These bioinformatic resources will allow the scientific community efficient access to genomic data, which will enable new types of analyses. The analyses, in turn, will allow for the computer modeling and subsequent experimental validation of the complex pathways and networks that ultimately determine the phenotype of a cell or the causes of many human diseases' (NIH website, accessed December 2009).

So, at least in the context of biocuration, model organisms have become models for data mining in humans. This move was prompted by the technical expertise accumulated by biologists in disseminating data obtained from model organisms, which is viewed as immediately relevant also to human data; and yet, it is not obvious that such expertise is sufficient to coordinate the dissemination of data obtained through clinical research. The paradox of proposing to treat humans as a model organism, while at the same time acknowledging that this effort is not currently carried out in co-operation with human geneticists and clinicians, is captured by the following quote from a paper summarizing the discussions over the role of model organisms in understanding and treating human disease held at the 2006 meeting of the Genetics Society of America:

> 'A critical need is better cross-organism databases that enable one to compare the genes, expression patterns, gene functions, cell types, tissue organization, and biological subprocesses

> across organisms, including humans. Maintaining and expanding our community resources, such as mutant collection and siRNA libraries for many organisms, including those not amenable to standard genetic techniques, is crucial. They provide access to the genetic power of the different model organisms and enable investigators to take full advantage of whole-genome sequence information. Finally, we must look for ways to interact with clinician scientists and human geneticists and bring their knowledge and perspectives to the modeling efforts' (Spradling et al 2006).

This quote clearly indicates that clinicians have not been involved with developing cross-species databases, and that this lack of involvement is potentially problematic and needs to be remedied. This situation is puzzling especially since one of the key purposes of cross-species comparisons is to achieve a better understanding of humans, leading to improvements in medical knowledge, diagnosis and treatments. In the words of a curator I interviewed in 2008,

> 'model organism databases also included human because obviously people are interested in what goes on in human, so that gets included even though there isn't an organism database'.

The curator is referring to the fact that there is no unique 'model organism database' for *Homo sapiens* (even if of course there are hundreds of disease / system / organ-specific human databases). There are several practical reasons for this: the sheer diversity and scale of data collecting practices on human beings; the multiplicity of sites where such collections are taking place, and the impossibility to coordinate and standardize the formats of collection; and the restriction to interoperability and access to human data, motivated by ethical concerns with privacy as well as by intellectual property issues in clinical research, particularly as carried out by pharmaceutical companies. If we stick to the above characterization of model organisms as ones on which data of all types, ranging from sequence data to morphological observations, can be collected and exchanged without restrictions, it is clear that *Homo sapiens* is not a model organism, nor could it become one in the future. Yet, as evident in the above

quote, in the context of cross-species databases, human data are treated in the same way as data coming from model organisms.

This observation opens a host of ethical questions about privacy concerns and the status of individuals and populations in biomedical research. These questions are being examined by scores of excellent social scientists, so I will not focus on them here. What I wish to explore is, rather, the differences in experimental practices characterizing human and non-human research that are brought to the fore by current attempts to develop cross-species databases. In what follows, I focus on four sets of issues that curators perceive to be emerging when human data are added to model organism databases: those concerning data, meta-data, materials and terminology. My analysis is based on a cross-examination of the content and guidelines of the GO and GMOD websites; and multi-sited ethnographic research on curation practices and database building carried out between 2004 and 2010, which included attendance at scientific meetings concerning biocuration in both model organism biology and medicine; visits to laboratories engaged in extensive bioinformatic work, including the development of cross-species databases; and extensive interviews with curators of cross-species databases based in the UK, Germany and the US. As a result of this long-term engagement, I have developed an ongoing dialogue with several curators, sometimes leading to collaborative activities (such as writing scientific papers together or inviting scientists to science studies meetings) and advisory roles within those communities (I am presently a member of the advisory board of the Genomic Network of Arabidopsis Research). My close personal and professional ties with database curator communities provide me with in-depth insight in their working habits and daily struggles, which as I will show is relevant to understanding the issues emerging when disseminating cross-species data. At the same time, this research methodology situates my analysis within the boundaries of my own interpretive understanding of a specific empirical context – which, as any ethnographic account, leaves open the question of whether these results capture

other (understandings of) efforts to order and disseminate data. In particular, the concerns I shall voice are the ones expressed by curators trained in the biological sciences, who have had little or no exposure to clinical research. This is in itself a remarkable fact, underscoring the difficulties in involving clinical researchers in the curation of cross-species databases. The implication is that this study does not represent views held by clinicians and by curators involved solely in human databases, a shortcoming that I hope will be addressed through future empirical research.

Issue 1: Data. What Counts as Reliable Evidence?

The first issue I wish to highlight concerns a divergence in the criteria used to determine what counts as reliable evidence. The problem is exemplified by the unclear status of microarray data as a source of evidence about gene expression. A great deal of standardization of terminology, experimental protocols and instruments is required to describe a microarray experiment – and, at the same time, to make sure that the procedures and techniques used within such an experiment are intelligible and replicable across different laboratories. The MIAME project, which stands for Minimal Information About a Microarray Experiment, has been set up precisely to address this need and streamline the process of agreeing upon, and implementing, such standards. Still, the development of standards such as MIAME has been fraught with difficulties and controversies (Roger and Cambrosio 2007), and MIAME standards are still far from being widely applied. This means that the quality and reliability of microarray data is hotly contested. A recent review in Nature Genetics set out to evaluate the replicability of 18 datasets obtained through microarray experiment and found that ten could not be reproduced on the basis of the information provided. The conclusion was that ‘repeatability of published microarray studies is apparently limited. More strict publication rules enforcing public data availability and explicit description of data processing and analysis should be considered’ (Ioannidis et al 2009, 149).

Despite remaining highly disputed,[viii] this kind of assessment provides the background to understand why several model organism databases do not accept microarray data as a valuable source of information about an organism. Many of the curators of these databases still view results of microarray experiments as highly dependent on the specific circumstances and expertise of the scientists who carry them out. The following quote exemplifies the feelings of several curators whom I interviewed on this subject:

> 'I'm doubtful that we would include any micro-array results at the moment. [..] You get very variable results from micro-array and you get lots of indications that genes are involved in certain processes when they may not be. They're up-regulated because of various different reasons which may not be related to the experimental conditions that are used. So, yeah, they're a bit doubtful'.

In clinical settings, the variability and lack of experimental 'validation' of microarray data, which make them so problematic to accept within model organism science, do not seem to raise the same amount of skepticism. While it is certainly true that the evidential value of genome-wide association studies is being widely debated, there is a widespread agreement that microarray experiments play an important role as sources of genomic evidence, especially since microarray results are being used in conjunction with other sources of evidence on the same genes/processes. The idea of experimental replication as a way to validate results is not as strong in clinical research as it is in biological research, for the simple reason that replicating experiments on the same tissues / humans is expensive, if it is at all possible given that samples are unique. Further, microarray experiments are seen to have an exploratory value: they can point to interesting correlations and patterns that might, upon further research, turn out to have biological meaning, yet they provide no clear evidence that those patterns exist, and would certainly not be trusted in isolation from other types of data.[ix] There is no intrinsic reason why this approach

should not be equally powerful in the biological realm. However, curators perceive biologists as showing a low level of trust in results acquired by other researchers, and ideally wanting to be able to assess the reliability of microarray data on a case-to-case basis. This perception is illustrated by the following quote from a curator of a model organism database:

> 'It's a tricky situation at the moment. Some people do annotate to high throughput experiments. We're thinking about doing it, we haven't done it as yet because we're still working out how we would do it and to what extent we would go, because there are lots of experiments out there that maybe wouldn't give very good results…very reliable results. But there are some certain high throughput experiments that are quite robust and so you would trust results you get. So we're kind of in the process at the moment of deciding which experiments we would be happy to include'.

Another interesting case of potential discrepancy between data mining in clinical and biological contexts concerns the ways in which data are extracted from publications. Several model organism databases rely on text-mining, or in some cases even manual annotation, to extract published data from available literature on a specific organism. This arduous task is made marginally easier by the existence of a rather coherent corpus of literature on each popular organism (Davies 2004)**.** The situation is perceived to be different in clinical research on humans, as another curator relates:

> 'You kind of feel like 'yeah, it would be nice if we had an organized set of human literature', the kind of thing that FlyBase provides for *Drosophila* and that MGI provides for the enormous body of mouse literature. There are times when I feel like it would be nice if the sequence data and the transcript, you know, expression data and proteomic data and function data, the localization data were more unified. But the big one is the literature that I think is probably the biggest single thing that model organisms have because of model organism databases and that is

missing from the systems that deal with the human genome sequence or human gene expression data'.

It is worth noting that, while it is true that the publications on humans are so numerous and dispersed across disciplines to make it very difficult to assemble them together, this is due to the fact that medical research is organized in ways that differ widely from the typical set-up of a model organism community. In this latter case, funding usually comes largely from a restricted number of governmental funding bodies, and massive resources are invested in community-formation and identity politics. Researchers identify themselves as 'rat-people', 'worm-people' and 'fly-people', and much effort goes into making sure that all researchers working on the same organism know each other, attend common meetings and exchange data. As I argued above, this ethos is at the core of contemporary model organism biology: the very notion of model organism is linked to the construction of infrastructure and communities that can integrate data on a single species, with the future goal of using that as a reference point for a comparative understanding of organisms as wholes. Research on humans has a much longer and complex history, is located in several different types of settings ranging from research laboratories to clinics, involves countless more professionals than the few thousands involved in research on model organisms, and partly as a consequence of this is not focused on the integration of data as much as on understanding and treating specific diseases.[x]

Issue 2: Meta-Data. Detailing Experimental Protocols

A second headache for database curators is the lack of agreement on what information needs to be included about the experimental circumstances in which data are originally obtained – in other words, information about the *provenance* of data, the processes through which data was produced and formatted for dissemination (Bowker 2001, 664; Evans and Foster 2011**).** This information, technically

referred to as meta-data, is crucial to assessing the evidential value and reliability of data found in a database. By accessing meta-data, users get to know who gathered the data of interest, the methods employed to do so and the research interests that motivated data production in the first place. These are all elements that help researchers to evaluate whether data are trustworthy, how they compare to other datasets available on the same phenomena and, as a result, what biological interpretation they could credibly support. Meta-data thus constitute ways to represent the tacit expertise underlying the production of data, in a way similar to the one described by Michael Lynch in the case of protocols (Lynch 2002). Contrary to Lynch's case, however, the representation does not primarily serve administrative purposes: its most important function is to encourage the critical scrutiny of these practices by as wide a community of peers as possible, so as to facilitate the proliferation of different interpretations of the same data.

The process of gathering meta-data is complicated by the fact that, even within model organism biology, different labs disagree on what elements are crucial in describing the provenance of data. Further, experimental protocols and procedures are constantly shifting, making it difficult to settle on fixed types of information as meta-data. Still, the curators of model organism databases argue for the importance of settling at least minimal standards for what counts as important information about an experiment.[xi] The most fundamental piece of information that needs to accompany each dataset is, unsurprisingly, the specific organism on which the data was obtained. The very idea of comparing data obtained on different organism depends on clearly identifying the species, and sometimes even the individual specimen, on which the data were originally collected. And yet, precisely on this crucial point curators find that clinical and biological researchers differ in how they conduct and describe experiments. Clinical researchers are perceived as frequently mixing organic materials coming from different organisms. According to the curators I interviewed, they often contaminate human samples

with materials coming from other organisms – RNA probes, for instance – and do not care to specify this when writing up their results. They sometimes even fail to specify whether they are working on human cells or mouse cells, on the grounds that they are convinced, almost certain, that this will not matter for their conclusions. This attitude clashes with the strict standards for annotating experimental materials and procedures adopted within model organism biology. This sometimes results in curators refusing to include data in a cross-species database, because they cannot classify them according to the organism on which they were acquired. In the words of one curator:

> 'when people publish, they…a lot of times don't tell you what protein they're working on, whether it's mouse or human. They'll tell you the protein name, but that could be 99% identical between human and mouse and they won't tell you which species it's from. And so in that case we can't annotate, we don't know exactly the species'.

Further, curators are committed to distinguishing results acquired through experimental procedures (referred in the quote below as 'primary annotations') from the interpretation of those results given by experimenters ('author statement'). One of the main worries underlying the contamination of samples is that experimenters tend to decide, on the basis of their own experience and of the specific circumstances in which data are produced, whether contamination is relevant or not to interpreting the results. The reasons for this important decision are thus kept tacit and inaccessible to the users of databases that report those data, who are left with the only option of trusting the scientific judgment (and thus the beliefs and expertise) of the original data producers. This situation generates uneasiness among curators, since efficient data re-use is understood to involve the possibility to scrutinize (and if necessary, challenge) the beliefs and context in which data were originally produced, as noted by this curator:

> 'People will be interested in humans, but do the experiments in model organism or take a cloned

> human gene and work with it in a system that's otherwise model organism in cultured cells or something. And *it can be very, very difficult to find out which species the sequences came from*, which species the cells came from, in a paper. It's not that it's necessarily unreasonable to infer that the human does the same thing as the mouse thing or the rat thing in many of these instances, but it would still be nice to know whether you stuffed the human gene into some cultured mouse cells or vice versa just in case there's a difference and *so that you know which one's the primary experimental annotation and which one's inferred from your belief that…that these genes correspond or are somehow equivalent*.'

Remarkably, a consequence of such uneasiness is that human data on gene products are often annotated as author statements, because experiments are not carried out entirely on human tissue. Curators sometimes try to resolve this issue by emailing the authors of papers directly:

> 'Quite often I send an e-mail to the author. I'll write and say, 'can you tell me which species it's from?' And a lot of the time I get response, maybe 50% of the time you get response and you can annotate that then, whereas before you'd have to just chuck the paper away, can't do anything with it.'

One way to explain the perceived difference in the ways in which clinicians and biologists annotate their experimental results is to think of the different priorities and commitments involved in their daily activities. It is often said that the medical project differs from the biological one in its emphasis on intervention and treatment: while clinicians aim to cure, biologists aim to explain. This distinction cannot be applied too neatly to experimental cultures in the two realms, since they both attempt to understand biological processes (whether general processes like metabolism and development or specific syndromes such as breast cancer) and to successfully manipulate organisms. However, the above remarks by curators on how experimenters annotate and assess their data point to some interesting differences in the ways in which biological and clinical experimental results are valued and

used. These differences might be at least partly explained by the ways in which experimentation in the two realms is evaluated by funding bodies. Biologists are increasingly under pressure to produce results of social and economic relevance, and yet the quality of biological research is still primarily assessed through peer review of the procedures through which a new explanation was crafted and tested. As a result, enhancing the quality and credibility of experimental research in biology involves major efforts in documenting and validating the sources of the evidence used to back specific claims. By contrast, clinicians' experimental results are explicitly valued not only for the biological insight that they generate, but also for their fruitfulness in supporting effective treatment of patients. This difference in priorities and evaluative cultures might contribute to explaining the relative disinterest of clinicians in documenting procedures and testing the reproducibility of results.

Issue 3: Materials. Which Materials Get Standardized and How?

Another reason that might account for the difference in experimental annotations concerns the very relationship built by researchers with the organisms that they study. This brings me to the third issue I wish to discuss, which is the experimental procedures used to select, manipulate and standardize organisms (both individual specimens and parts such as tissues, cell cultures, blood, organs). Within model organism biology, the standardization of organisms is of paramount importance: being able to access specimens that are genetically and/or phenotypically identical to the ones on which experiments are carried out is seen as crucial to validating experimental results and pursuing research that builds on previous efforts (Rosenthal and Ashburner 2002). Model organisms are standardized through two types of processes. The first consists of the processes of *transformation* from organisms found in the field to tractable laboratory specimens. The very act of transporting an organism into a laboratory environment occasions several changes to its biology (ranging from its physiology to its genome), due to the need to live in an environment where the basic rules of survival in the wild are subverted. For instance, most

basic needs, such as food and light, are provided for and with much less variety than in the field; the organisms are protected from the vagaries of the elements and live in relative isolation from each other as well as from other forms of life. On top of the biological transformations due to this change of context and habits, which enhance the tractability of the organism by researchers who need to handle them, there are often genetic modifications specifically intended to make organisms more suited to the research goals at hand. The oncomouse is probably the best known case for research aimed to understand human biology (Clarke and Fujimura 1992; Murray 2010); Arabidopsis has been extensively modified to increase its susceptibility to changes in light and temperature, in the attempt to uncover the genetic basis of processes such as vernalization and photosynthesis. The second type of standardization processes used to produce model organisms is the one involved in the *dissemination* of specimens and related findings across research communities. For organisms to become favoured scientific materials, it is not sufficient that they are tractable in a laboratory environment and useful for the research that is carried out. Once that research starts to be disseminated across a wider community of experimenters, the organisms themselves need to be able to travel across different labs and research contexts, so that researchers can verify those results and/or further them through more experiments on precisely the same type of organism. The requirement to be physically sent to laboratories across the globe contributes to defining the characteristics of the organism selected for research: for instance, bigger organisms fare worse than smaller organisms and organisms that easily survive displacement are favoured.[xii]

As illustrated by these procedures, the need to standardize guides and conditions all stages of researchers' interactions with model organisms. This situation is quite different from the ways in which researchers interact with human subjects, and indeed neither of the two processes of standardization described above maps neatly onto the treatment of humans in clinical research. Let us consider the

process of transformation first. It is true that human subjects are selected as subjects for research according to their biological characteristics, including sometimes their genetic make-up or their ethnic background. Some clinical studies look for 'adequate' populations across the globe – where 'adequate' means representative of the traits that researchers wish to study, and/or amenable to the kind of treatment and sampling required for clinical research purposes. Not surprisingly, however, the latter interpretation of what constitutes adequate populations is under heavy ethical scrutiny (Petryna 2009), not least because treatments are supposed to be tested on any group that might benefit from them, including ethnic minorities, affluent populations, children and pregnant women**.** Further, the very notion that human beings might be used as instruments for research, to the point of infringing on their basic rights (among which the right to privacy), is extremely controversial, and is often argued to require tighter regulation and more media attention than it is now (e.g. Waldby and Mitchell 2006). Another possible parallel to the process of transforming a model organism is the way in which patients are 'prepared' for participation in a clinical study, for instance through a specific diet and/or by imposing a set of appropriate behaviors and habits as a condition for participation (e.g. stopping to smoke or drinking alcohol). Even when taking this into account, however, human subjects cannot be viewed as undergoing physical modifications comparable to the transformation of model organism specimens so as to fit research needs – and again, there are excellent ethical and practical reasons why humans are not, nor should be, engineered in this way. Turning now to the process of dissemination, the parallels with the treatment of model organisms are more striking, even if still limited and controversial. While individual subjects are not routinely shipped around the world as a research commodity, samples of their tissues, cells or blood are disseminated through biobanks and thus selected on the basis of clear standards for what constitutes an acceptable donation (for instance in terms of its integrity, characteristic features, provenance and means through which it was collected). Still, such dissemination of samples is subject to stringent regulation that vary across national borders (Gere and

Parry 2006; Kaye and Stranger 2009). Also, variability across human individuals plays an important role in clinical research – each sample is unique and not easily cloned or reproduced, which makes samples into a precious commodity whose dissemination is only agreed upon under specific circumstances (Parry 2004). Overall, ethical, practical and regulatory constraints make it impossible to think of human subjects in the same way as we think of specimens in non-human research. Indeed, this is the very reason why, despite the well-known ambiguities in inferring medical insights from research on model organisms and the controversy surrounding the use of animals in research, experimentation on non-human remains a stepping stone for clinical research. At least within Western scientific culture, the idea that humans deserve more respect and protection than non-humans has a strong hold.

This set of considerations adds another important layer to curators' worries about extensive differences in how researchers treat organisms. Remarkably, thinking about the ways in which organisms are manipulated shifts the focus away from a rigid divide between biological and clinical research, both of which are likely to use cross-species data. The relevant difference here is instead the one between researchers who study model organisms and researchers who study humans. Clinicians working on mice are much more likely to adhere to the practices recommended by database curators to describe their specimens, while researchers carrying out experiments on human subjects and their parts operate in quite a different experimental culture. It is then not surprising to note that, while clinicians working on humans were never officially involved in the effort to develop the GO, prominent representatives of the mouse community were among its founders (together with researchers working on fruit-flies and worms; Leonelli 2010b).

Issue 4: Terminology. Extending the Gene Ontology

The last issue I wish to discuss is the choice of terminology used to classify and retrieve data across

organisms. For cross-species databases to work, such terminology should function as a *lingua franca* intelligible to all potential database users (Gene Ontology Consortium 2000). Already within model organism communities, the problem of choosing terms that different groups will recognize and understand is one of the most urgent issues confronted by curators.[xiii] Achieving terminological compatibility – if not integration – across the human/non-human boundary and across biological and clinical practice is even more daunting, especially given the amount of efforts already invested by the medical community towards terminological homologation (e.g. the Medical Subject Heading created by the National Library of Medicine to index medical literature, see website http://www.nlm.nih.gov/mesh/introduction.html). Attempts to integrate the terminologies used in medicine with the ones used in biology are under way since decades. Some attempts, such as the ones focused on specific areas / diseases, are making headway, as illustrated by NCI Enterprise Vocabulary Services (EVS), a set of tools developed by National Cancer Institute to integrate molecular and clinical research on cancer (http://evs.nci.nih.gov/). To exemplify the issues that might emerge when merging vocabularies coming from model organism research and research on humans, I will focus briefly on the recent merger of GO terms with the Unified Medical Language System, a metathesaurus of 900.000 medical terms developed by the National Library of Medicine which is recognized as one of the most authoritative reference for standard medical terminology (Nelson et al 2002; Bodenreider 2004). Despite curators' published claims to the effect that the merger had been relatively smooth (Lomax and McCray 2004), my interviews with curators involved in this effort reveal that this attempt towards integration generated some interesting paradoxes and led to revisions of GO, many of which are still under way. For example, a key organizing principle within GO is to distinguish terms that describe a molecular function from terms that describe a biological process. Within medical discourse, such partition does not make sense, since the molecular function of gene products is automatically equated with a characterization of the biological process in which that gene is involved; thus, the

UMLS nomenclature does not classify its terms according to it. Further, even in cases where terms overlap between the two nomenclatures, the meanings assigned to those terms might differ in a number of respects: the definition assigned to the terms might of course be different, a case where epistemic dissonance might be easily resolved through modifying the definitions; more problematically, however, the same term defined in the same way might acquire different meanings depending on its position in the semantic network of the nomenclature. To understand this point, one needs to know that terms in nomenclatures such as ULMS and GO are organized hierarchically through a series of relationships. Both systems rely heavily on the 'is_a' identity relation (as in 'nuclear membrane is a membrane'), yet differ in the other relationship types that they use. For instance, basic relationships in GO are mereological or 'part_of' relationships and functional relationships such as 'regulates'. In contrast to GO, ULMS uses a wider and broader range of relationships including `physically related to,' `spatially related to,' `temporally related to,' `functionally related to,' and `conceptually related to'.[xiv] Given these differences in semantic structure, terms shift their meaning depending on where they are situated in the network – in much the same way as the interpretation of single words in everyday communication depends on the linguistic and social context in which they are used.

Yet another issue emerges in relation to the process through which curators select which terms should be used to classify given sets of data. In biology, annotations tend to be based on peer-reviewed publications relating datasets to specific processes, functions or entities. In clinical research, however, it might be hard to find a direct, well-established link between a dataset and a term of interest – for instance, a disease. Still, there might be good reasons to suspect that such a link exists, and thus to annotate those data under the term referring to the disease in question. Trying to accommodate these different criteria is puzzling to curators trying to work on both realms, as evident in the following quote:

> ‘The clinical research tends to give you more or less detailed GO terms, so, for example, epilepsy. You might, if you have nothing else for that protein but you know that that a mutation in that gene causes epilepsy then you could annotate to neuron development or something. So it’s a bit of a strange annotation to make, because you don’t know if it’s a direct link for that protein to cause…that protein is involved in neuron development, because it may be way downstream. But if you don’t have any other functional information for protein, then it’s good to make that annotation anyway.’

This brings us back to the divergence in priorities that I discussed with reference to data assessment and meta-data annotation, and enables me to add a further layer to it. Why would clinical researchers be satisfied with incomplete information in this case? One possible answer is that this is because, in their worldview, having some information is better than having no information at all. Biologists are typically more cautious in claiming causal links between biological processes, aware as they are of the risk of spurious correlations. Clinicians are more used to make causal claims and take account of information deemed to be relevant to a given disease without fully understanding the mechanisms causing it. This is because they do not have the luxury of waiting until detailed mechanistic understanding is achieved; they are under pressure to intervene on patients on the basis of the best knowledge available at any given time. Thus, in biomedical research, any hint that points to the etiology or treatment of abnormal human phenotypes merits mention; experimental research in this realm can always be characterized as largely exploratory.

Finally, I wish to mention a more fundamental, conceptual problem with integrating nomenclatures across the human/non-human boundary. Databases such as GO have been built to focus on non-pathogenic entities and processes – which are referred to as ‘normal’ (Gene Ontology Consortium 2000). The reason for this is clear: model organisms are supposed to be representative for the biology

of a wide set of organisms, and are thus conceptualized as 'typical' of 'normal' gene functions found in a given species. Clinical research on humans has almost the opposite connotation: because the main interest is in understanding and treating specific pathogenic conditions, cross-species research is centered on diseases and the vast majority of available human data document so-called 'pathological states'. This situation causes serious problems when it comes to incorporate data on diseased organisms into GO, with the consequence of making cross-species databases potentially less interesting to clinical researchers, as one of the curators I interviewed makes clear:

> 'Some people at the National Cancer Institute have opted not to do very much with GO because it's not ontology things, cancer or [GO is] explicitly excluding abnormal things such as oncogenesis or tumourogenesis that makes it unsuitable for their purposes. There are plenty of biologists who use it, but wish GO had more depth in this area or better annotations in that area'.

This feature may account, at least in part, for the limited participation of clinicians working on human pathologies to the development of cross-species databases. At a more fundamental level, it raises the deep philosophical question of what constitutes 'pathogenesis' and 'normality' in the biological and clinical realms.[xv] An adequate discussion of this issue deserves more space that I can devote to it here. It is however an important dimension to mention in closing this section, as the way in which researchers answer this question deeply affects their conceptualization of organisms and of how experimental results should be collected, disseminated and interpreted**.**

Conclusion: Identifying and Aligning Experimental Cultures in Biomedicine

In her study of how databases are used to organize scientific work, Christine Hine has observed that the significance of these tools needs to be evaluated with reference to existing epistemic cultures, since

'while practices and outcomes of knowledge production may change with increasing use of information and communication technologies, such changes do not do away with existing framework' (Hine 2006, 290). In this paper, I have turned Hine's argument on its head by using the study of database construction as a starting point to identify extant differences and tensions among the epistemic cultures that use these technologies. Examining curators' perceptions of the difficulties involved in developing cross-species databases, and particularly in combining data coming from model organisms and humans, has enabled me to identify and discuss several characteristics of the complex interface between biology and medicine.[xvi] One set of the resulting observations reinforces the idea of a strong divergence in experimental practices, goals, and values between biologists and clinicians. For a start, there is a difference in research priorities and goals. Both sides aim to understand and change the world. Yet, biologists use their experimental skills as a way to understand organismal biology, while clinicians view the accumulation of biological understanding as essentially aimed to treat patients. This difference in emphasis is amplified by the evaluative cultures surrounding these two realms (issue 2). Clinicians are working in an environment where research is evaluated both for its contributions to medical knowledge and for its impact on treating patients. Despite the increasing push towards applied research, this is not the same for biologists, whose outputs are evaluated mainly through the quality of their publications. This in turn reinforces differences in how biologists and clinicians conduct research. For instance, clinicians tend to use data that biologists consider to be potentially unreliable, as illustrated by the controversies around the inclusion of high-throughput data in cross-species databases (issue 1) and the inclusion of causal, yet unproven, information in GO (issue 4). More generally, clinicians tend to value causal information that biologists do not see as conclusive, because of its potential value towards finding treatments. In their eyes, inserting such information in databases means increasing the chance of gathering useful clues towards understanding phenomena of interest; biologists seem to be more risk-averse, fearing that lowering standards for what counts as evidence will weaken the overall

reliability of information found in databases. One final source of contrast is the way in which biologists and clinicians seem to direct their attention to different aspects of organismal biology altogether: 'normal states' versus 'pathological states' (issue 4). Again, this is a difference in emphasis rather than a strict divide – after all, many of the mutants used to study healthy organisms might be defined as diseased, while clinicians obviously need an idea of what constitutes a healthy organism in order to identify and treat unhealthy ones. And yet, when added to the other evidence I mentioned, it seems to point to a rather stark separation between experimental cultures in medicine and biology. This simple picture is however disrupted by other aspects emerging from this study. Many sources of controversy highlighted by database curators stem not from cultural divergences between clinicians and biologists, but rather from differences in the research practices of experimenters who work with non-human organisms and experimenters who work with humans. Clearly, experimenting on humans brings ethical, financial and material constraints that are not present in model organism research. My analysis has highlighted the material aspects in particular, by showing the differences in how human and non-human organisms are prepared and standardized for research (issue 3); and the difficulties in re-using human samples and reproducing results from clinical trials (issue 1). These material conditions give rise to differences in the ways in which researchers communicate and in the labels that they choose (issue 4).

These results emerge from an analysis of curators' struggle to understand the needs and working conditions of their users. Curators are dealing with two cross-cutting ways to identify user categories: they need to pay attention to the difference in research cultures underlying clinical and biological research; at the same time, they have to cater for model organism researchers and for researchers working on human subjects. This shows how databases are fast becoming crucial sites for the encounter of those diverging cultures, the identification of differences and the expression of conflict (which may

or may not pave the way to its resolution). The recent deluge of genomic data is making it ever more difficult for biologists and clinicians to interpret the wealth of information found online in ways that help understanding the material bodies they work with – whether they are bodies of insects, plants, animals or humans. This process of aligning the informational with the material is specific to data-intensive modes of experimental research and constitutes one of the foremost scientific challenges of the 21st century. The divides between biologists and clinicians on the one hand, and human and non-human research on the other, make this process of alignment even more complex to achieve. The work done by curators is key to confronting this challenge. Curators are engaged in a process of alignment at all three levels: while attempting to foster the use of *in silico* information to understand organisms *in vivo*, they have to navigate the experimental cultures involved in biomedicine, and attempt to identify and align divergence between biological and clinical, and human and non-human, research. This process is crucial to what Keating and Cambrosio (2003) have described as the ongoing re-alignment of biology and medicine characteristic of post-genomic science. From the analysis above, it is evident that the production of adequate databases, and consensus on how these can and should be used, is a crucial platform for the development of cross-species research; and that cross-species research is in turn crucial to biomedical inquiry in all its forms, whether it is taking place in the lab or in the clinic, and whether it is focusing on the discovery of biological mechanisms or on therapeutic application. And indeed, how curators end up dealing with experimental pluralism in all its different forms is likely to have a huge effect on how different constituents of biomedical research relate to each other. The ways in which databases are structured, and the choice of which data gets included and how, can dilute or reinforce the differences in experimental cultures noted above. Sociological research on database development and use thus constitutes an important platform for the analysis and understanding of conflict, collaboration and integration in biomedicine. Further, it can improve our understanding of the relations between basic and applied research, and particularly of what constitutes translation in biomedicine and

beyond. Keating and Cambrosio have already stressed the need to 'understand the interactions between fundamental and clinical research in terms other than subordination or application' (Keating and Cambrosio 2004, 368). The study of databases provides an excellent alternative to these binaries. We are not witnessing a simple movement from 'bench to bedside' here, nor a straightforward transfer of biological knowledge to medical practice. Rather, the focus on how data and materials are traded online, and on the related shifts in the skills and tools used to interpret such data and materials, facilitates the understanding of biomedicine as a complex web of practices which defies a neat classification into the very categories of 'basic' and 'applied'. Research currently invested on the study of simple model organisms, such as *Arabidopsis* or *C. elegans*, plays as important a role within this web of practices as the recent attempts to develop genome-wide association studies for the study of cancer. Making sense of the intersections between human and non-human research is key for both scientists trying to further biomedicine and for social scientists interested in understanding their work – and as I hope to have shown, the development of cross-species databases is a good place to start.

References

Ankeny RA (2001) The natural history of *C. elegans* research. *Nature Reviews Genetics* 2: 474–8.

Ankeny RA and Leonelli S (2011) What is so special about model organisms? *Studies in the History and the Philosophy of Science: Part A*. 42 (2): 313-32

Baker KS and Millerand F (2010) Infrastructuring ecology: challenges in achieving data sharing. In Parker JN, Vermeulen N and Penders B (eds) Collaboration in the New Life Sciences. Ashgate.

Bodenreider O (2004) The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research* 32: 267-270.

Bolker JA (1995) Model systems in developmental biology. *BioEssays* 17: 451–5.

Bowker GC (2001) Biodiversity datadiversity. *Social Studies of Science* 30 (5): 643-683.

Buetow KH (2005) Cyberinfrastructure: empowering a "third way" in biomedical research. *Science*, 308 (5723): 821 – 824.

Bult CJ (2006) From information to understanding: the role of model organism databases in comparative and functional genomics. *Animal Genetics* 27 (1): 28-40.

Canguilhem G (1991 [1966]) *The Normal and the Pathological.* New York: Zone Books.

Chow-White PA and Garcia-Sanchos M (2011) Global genome databases bidirectional shaping and spaces of convergence: Interactions between biology and computing from the first DNA sequencers to global genome databases. *Science, Technology and Human Values.* Published online 27 February 2011. DOI: 10.1177/016224391039796

Clarke AE and Fujimura JH (1992) *The Right Tools for the Job. At Work in Twentieth-Century Life Sciences.* Princeton, NJ: Princeton University Press.

Davies G (2011) Playing dice with mice: building experimental futures in Singapore. *New Genetics and Society,* in press.

Davies G (2010) Captivating behaviour: mouse models, experimental genetics and reductionist returns in the neurosciences. In Parry S and Dupré J (eds) *Nature After the Genome*. London: Sage.

Davies RH (2004) The age of model organisms. *Nature Reviews Genetics*, 5: 69-76.

Editorial Nature (2009) The sharing principle. *Nature* 459:752.

Evans JA and Foster JG (2011) Metaknowledge. *Science* 331 (6018): 721-725.

Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics* 25: 25-29.

Gene Ontology Website (accessed April 2010), http://www.geneontology.org

Gere C and Parry B (2006) The flesh made word: banking the body in the age of information. *Biosocieties* 1 (1): 83-98.

Hey T and Trefheten AE (2005) Cyberinfrastructure for e-science. *Science* 308: 817-821.

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

Hine C (2006) Databases as scientific instruments and their role in the ordering of scientific work. *Social Studies of Science* 36(2): 269-298.

Howe D et al (2008) Big data: the future of biocuration. *Nature* 455: 47-50.

International Arabidopsis Informatics Consortium (2010) An international bioinformatics infrastructure to underpin the Arabidopsis community. *Plant Cell* 22(8): 2530-2536.

Ioannidis et al (2009) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics* 9: 356-369.

Kaye J and Stranger M (ed) (2009) *Principles and Practice in Biobank Governance*. London: Ashgate.

Keating P and Cambrosio A (2003) *Biomedical Platforms: Realigning the Normal and the Pathological in Late-Twentieth-Century Medicine.* Cambridge, MA: MIT Press.

Keating P and Cambrosio A (2004) Does Biomedicine Entail the Successful Reduction of Pathology to Biology? *Perspectives in Biology and Medicine* 47(3): 357-371.

Kohler RE (1994) *Lords of the fly: Drosophila genetics and the experimental life*. Chicago, IL: University of Chicago Press.

Kohli-Laven N, Bourret P, Cambrosio A and Keating P (2011) Cancer clinical trials in the era of genomic signatures: biomedical innovation, clinical utility, and regulatory-scientific hybrids. *Social Studies of Science* 41(4) 487–513

Leigh Star S and Rhleder K (1996) Steps towards an ecology of infrastructure: design and access for large information spaces. *Information Systems Research* 7 (1): 63-92.

Lenoir T (1999) Shaping biomedicine as an information science. In Bowden ME, Hahn TB and Williams RV (eds) *Proceedings of the 1998 Conference on the History and Heritage of Science Information Systems*. Medford, NJ: Information Today, Inc., ASIS Monograph Series, 27-45.

Leonelli S (2009) Centralising labels to distribute data: the regulatory role of genomic consortia. In Atkinson P, Glasner P and Lock M (eds) *The Handbook for Genetics and Society: Mapping the New Genomic Era.* Routledge, London, pp. 469-485.

Leonelli S (2010a) Packaging data for re-use: databases in model organism biology. In Howlett P and Morgan MS (eds) *How Well Do Facts Travel? The Dissemination of Reliable Knowledge.* Cambridge, UK: Cambridge University Press.

Leonelli S (2010b) Documenting the emergence of bio-ontologies: or, why researching bioinformatics requires HPSSB. *History and Philosophy of the Life Sciences*, 32 (1): 105-126.

Loewy I (1986) *Between Bench and Bedside: Science, Healing, and Interleukin-2 in a Cancer Ward*. Cambridge, MA: Harvard University Press.

Lomax J and McCray AT (2004) Mapping the Gene Ontology into the Unified Medical Language System. *Comparative and Functional Genomics* 5: 354-361.

Lynch, M (2002) Protocols, practices and the reproduction of techniques in molecular biology. *The British Journal of Sociology* 53 (2): 203-220.

McMullen PD, Morimoto RI and Amaral LAN (2010) Physically grounded approach for estimating gene expression from microarray data. *PNAS* 10(31): 13690-13695.

Murray, F (2010) The Oncomous that roared: hybrid exchange strategies as a source of distinction at the boundary of overlapping institutions. *American Journal of Sociology* 116(2): 341-388.

Nelson SJ, Powell T, Srinivasan S and Humphreys BL (2002) The Unified Medical Language System (UMLS) project. In Kent A and Hall CM (eds) *Encyclopedia of Library and Information Science.* New York: Marcel Dekker, Inc., 369-378.

National Human Genome Research Institute, National Institute of Health Website (accessed December 2009) http://www.genome.gov/10001735

O'Malley MA (2008) Exploratory experimentation and scientific practice: metagenomics and the proteorhodopsin case. *History and Philosophy of the Life Sciences* 29(3): 337–358.

Parry B (2004) *Trading the Genome*. New York: Columbia University Press.

Petryna A (2009) *When Experiments Travel: Clinical Trials and the Global Search for Human Subjects*. Princeton, New Jersey: Princeton University Press.

Quirke V and Gaudillière JP (2008) The era of biomedicine: science, medicine, and public health in Britain and France after the Second World War. *Medical History* 52(4): 441-452.

Rader K (2004) *Making Mice*. Princeton, NJ: Princeton University Press.

Rhee SY and Crosby B (2005) Biological databases for plant research. *Plant Physiology* 138(1): 1–3.

Rhee SY (2004) Carpe diem: retooling the ‘‘publish or perish’’ model into the ‘‘share and survive’’ model. *Plant Physiology* 134: 543–547.

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

Rosenthal N and Ashburner M (2002) Taking stock of our models: the function and future of stock centers. *Nature Reviews Genetics* 3: 711–7.

Spradling A et al (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.

Star SL and Griesemer JR (1989) Institutional ecology, 'translations' and boundary objects: amateurs

and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science* 19 (3): 387–420.

Stein LD (2008) Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges. *Nature Reviews Genetics*, 9 (9):678-688.

Stein LD et al (2002) The Generic Genome Browser: a building block for a model organism system database. *Genome Research* 12: 1599-1610.

Sturtevant A (1954) The social implications of the genetics of man. *Science* 120: 60.

Taylor C et al (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nature Biotechnology* 26, 8: 889-896.

Unified Medical Language System Website (accessed 2010) http://www.nlm.nih.gov/research/umls/knowledge_sources/index.html#semantic

Waldby C and Mitchell R (2006) *Tissue Economies: Blood, Organs and Cell Lines in Late Capitalism.* Durham and London: Duke University Press.

Wouters P and Schröder P (eds) (2003) *The Public Domain of Digital Research Data* Amsterdam: NIWI-KNAW.

[i] A referee has pointed out that instead of focusing on differences between experimental cultures, thus study could have

interrogated what constitutes 'the same practice' across those communities. While agreeing that this would have constituted another interesting take on my data, I am not able to incorporate this into the already extensive scope of this paper.

[ii] The importance of model organisms as tools for interdisciplinary collaboration has also been stressed by several pioneering STS studies, including most famously the notions of model organisms as 'right tools for the job' (Clarke & Fujimura 1992) and 'boundary objects' (Star and Griesemer 1989).

[iii] It is true that the founders of these model organism communities grounded their interdisciplinary approach firmly on genetic studies, thus becoming vulnerable to anti-reductionist critiques (Bolker 1995). Nevertheless, historical research shows how strongly these scientists were committed to the long-term ideal of interdisciplinary integration (Ankeny and Leonelli 2011).

[iv] It is tempting to define model organisms as non-human organisms that are used to understand human biology, and of course model organisms are often used to that aim. However, it is important to remember that this is not always the case. Research on plants, for instance, is mainly focused on improving food production and acquiring new sources of energy; while some research on animals aims to make sense of specific features of their biology, which could be put to use in human societies (e.g. dog and sheep breeding, egg production, migration patterns in fish and bird populations). My definition of model organisms, based on the extent to which they are researched and understood rather than on what they are meant to represent, is thus broader and more satisfactory than a definition based on their role as human models.

[v] See the webpage of the National Human Genome Research Institute, listing the main community databases funded by the National Institute of Health (http: www.genome.gov/10001837).

[vi] The important role played by these databases in model organism research is underscored by the strongly critical reactions by biologists to recent announcements of funding cuts. TAIR in particular is under threat as NSF funding for the project has been phased out, and the vehement protests of the Arabidopsis research community resulted in the formation of an International Arabidopsis Informatics Consortium , which recommended the continuation of TAIR under the 'new' name Arabidopsis Information Resource (International Arabidopsis Informatics Consortium 2010).

[vii] Details on the history, personnel and characteristics of GO can be found in Leonelli (2009, 2010b).

[viii] Studies using richer models of the process by which microarray data are constructed claim to produce much better and more reproducible results of even existing microarray data (McMullen et al, 2010).

[ix] For an in-depth discussion of the characteristics of exploratory research, see O'Malley (2008).

[x] Research on mice represents an extremely interesting middle ground between the ethos characterizing model organism

biology and the ethos of clinical research, which deserves a whole study in its own right. As in other model organism communities, mouse-people form a tight network and strive to build infrastructures that will ensure the standardization and centralized access to specimens, protocols and data. At the same time, these efforts are routinely undermined by the scientific, social and economic context in which research on mice takes place, which tends to be extremely competitive and fragmented into small communities with little financial and social incentive to communicate with each other (see for instance the debates around the creation of a centralized stock centre, Editorial Nature 2009**).** For an in-depth study of current research on mice, see Gail Davies (e.g. 2010).

[xi] An example of this is the MIBBI project (Taylor et al 2008). It must be noted that, as remarked by a referee, few of these discussions on standards are applied within databases as yet. The level of standardization adopted by Gene Ontology curators is one of the highest among model organism databases.

[xii] It is no wonder that plants, whose specimens can be sent around in the form of seeds, are among the best stocked and standardised organisms (Leonelli 2007), while mice and rat researchers rely both on whole specimen collections and tissue cultures (Davies 2011)**.**

[xiii] The very success of the GO, as of all bio-ontologies within the Open Biomedical Ontologies Consortium, stems from attempts to solve this problem (Leonelli 2010b).

[xiv] From the GO and ULMS websites, accessed in March 2010.

[xv] Clearly, GO is not sharing in George Canguilhem's insight that that 'the menace of disease is one of the components of health' (1991).

[xvi] Admittedly, these issues are likely to emerge in any research environment making use of biological databases, and are thus not per se characteristic of the case of cross-species databases. For instance, biologists working on different organisms can strongly disagree on the reliability of microarray experiments; picking meta-data is an arduous task no matter which types of data one is trying to disseminate and re-use; and the process of labeling data for travel is always charged with specific interests, values and beliefs. However, looking at the specific manifestations of these problems in the development of cross-species databases can teach us much about the intersections between medicine and biology, as I argue here.