Re-Thinking Organisms: The Impact of Databases on Model Organism Biology

Sabina Leonelli (corresponding author)
ESRC Centre for Genomics in Society, University of Exeter
Byrne House, St Germans Road, EX4 4PJ Exeter, UK.
Tel: 0044 1392 269137   Fax: 0044 1392 269135
Email: s.leonelli@exeter.ac.uk

Rachel A. Ankeny
School of History and Politics, University of Adelaide
423 Napier, Adelaide 5005 SA, AUSTRALIA.
Tel: 0061 8 8303 5570   Fax: 0061 8 8303 3443
Email: rachel.ankeny@adelaide.edu.au

> *'Databases for model organisms promote data integration through the development and implementation of nomenclature standards, controlled vocabularies and ontologies, that allow data from different organisms to be compared and contrasted'*
>
> *(Carole Bult 2002, 163)*

Abstract

Community databases have become crucial to the collection, ordering and retrieval of data gathered on model organisms, as well as to the ways in which these data are interpreted and used across a range of research contexts. This paper analyses the impact of community databases on research practices in model organism biology by focusing on the history and *current use of four community databases: FlyBase, Mouse Genome Informatics, WormBase and The Arabidopsis Information Resource. We discuss the standards used by the curators of these databases for what counts as reliable evidence, acceptable terminology, appropriate experimental set-ups and adequate materials (e.g., specimens). On the one hand, these choices are informed by the collaborative research ethos characterising most model organism communities. On the other hand, the deployment of these standards in databases reinforces this ethos and gives it concrete and precise instantiations by shaping the skills, practices, values and background knowledge required of the database users. We conclude that the increasing reliance on community databases as vehicles to circulate data is having a major impact on how researchers conduct and communicate their research, which affects how they understand the biology of model organisms and its relation to the biology of other species.*

Keywords: database, data, model organism, data-intensive science, curator.

1. Introduction: Community Databases and Data-Intensive Science

The development of cyberinfrastructure, in the form of databases, modelling tools and communication platforms such as wikis, is having an enormous impact on how biological research is done, and what it achieves. A critical question for natural and social scientists, as well as for philosophers and historians, is how to assess this impact. Is this impact simply a matter of speed and scale, as new technologies allow fast and efficient access to unprecedented amounts of data? Or do the impacts of cyberinfrastructure make more fundamental changes in the nature of the knowledge produced, how it is organized and how it is utilized, and hence have far-reaching implications for scientific practices in the biological sciences? In this paper we explore these questions in the case of a specific type of cyberinfrastructure, the 'community database' used the context of model organism research, and argue that fundamental epistemic changes are indeed occurring which in turn have important implications for understanding scientific practice.

So-called 'model organisms' have become the main focus of much recent research within molecular biology as well as the biomedical sciences more generally.[1] In their modern form, community databases were first constructed in the 1990s as portals for accessing a wide range of information on specific model organisms primarily used for genetic research (Rhee and Crosby 2005). They have become popular and useful within large model organism communities, where the number of disciplines, variety of types of resource materials and variety of types of data involved is very large and researchers find it next to impossible to keep abreast of all developments and resources of potential relevance to their work. Research on the mouse cress *Arabidopsis thaliana*, for instance, has unmistakeably 'made it big', encompassing over 16.000 laboratories distributed across the five continents (Sommerville and Koornneef 2002). Accordingly, in 2006 the keyword *Arabidopsis thaliana* was used in 43% of all plant life science publications (Jonkers 2009). This proliferation of journal articles, numbering in the thousands every year, makes it difficult to rely solely on publications as vehicles of information from one's research field. Community databases were established to enable researchers to locate information on a given organism without having to read all existing literature (published or informally distributed, often without any indices or similar mechanisms for ease of access) or being personally acquainted with the work of all the researchers involved.

[1] Due to space limitations, and because it is not our main focus in this paper, we examine the history of research with what have become known as 'model organisms' extremely selectively and do not define what counts (or should count!) as a model organism. For detailed arguments concerning the complexities of defining a model organism, as well as an overview of existing historical and philosophical research on model organisms with a much more extensive bibliography, see Ankeny and Leonelli 2011

The rise of community databases has been strongly correlated with the current emphasis on data-intensive science and automated data analysis. The need for such tools long preceded the advent of high-throughput technologies for data production. For example, as early as the 1920s, *Drosophila* researchers in T. H. Morgan's laboratory struggled with the difficulties of distributing data and specimens of fruitfly across an expanding global community (Kohler 1994, ch. 5). The mouse community created a Mouse Club, with related newsletter to share mouse-related information, in 1922 (Rader 2004). Similarly, researchers working on the nematode *Caenorhabditis elegans* have been concerned about community building and information sharing since the early to mid 1970s, when they actively promoted an ethos of co-operation and data sharing through mechanisms such as annual meetings and the Worm Breeder's Gazette, which was published beginning in 1975 (Ankeny 2001).

The sequencing projects that took place in the late 1990s and the 2000s brought new urgency to the problem of how to store and organise huge masses of data. Genome sequencing and the subsequent high-throughput 'revolution' heightened the importance of online information sharing as a newly available solution to an old problem. The pursuit of sequencing data provided a powerful common goal for model organism communities, as well as a common denominator which could serve as the basis for future collaborative work. Scientists agreed that gaining access to sequence data was of utmost importance for future research in all areas of biology, thus constituting a collaborative platform for the integration of knowledge about single organisms as well as for comparative research across species. At the same time, the structural nature of sequencing data, which could not by itself provide meaningful functional information about the biology of organisms, provided an excellent illustration of the limitations of a formal publishing system based entirely on the communication of claims, hypotheses and experiments, and which did not easily accommodate raw data (Hilgartner

1995; Leonelli 2010b). Such a system could not support the dissemination of data that are not (yet) attached to specific claims about biological phenomena. Nonetheless it was widely argued that sequencing data could not be interpreted as evidence for claims unless they were shared in their raw, uninterpreted form.

Community databases thus began in the late 1990s and early 2000s[2] with the immediate goal of storing and disseminating genomic data in a formalized manner, and the longer-term vision of (1) incorporating and integrating any data available on the biology of the organism in question within a unique dataset, including data on physiology, metabolism and even morphology; (2) allowing and promoting cooperation with other community databases so that the available datasets eventually would be comparable across species; and (3) gathering information about laboratories working on each organism and the associated experimental protocols, materials and instruments, thus providing a platform for community building.

While based on ongoing comparative research on several model organism communities, including zebrafish, rat and yeast, this paper focuses on four community databases: The Arabidopsis Information Resource (TAIR), gathering data on *Arabidopsis thaliana*; FlyBase, on *Drosophila melanogaster*; WormBase, on *Caenorhabiditis elegans*; and Mouse Genome Informatics, on *Mus musculus*. This focus is partly dictated by convenience, as a survey of all existing community databases would be beyond the scope of this paper, and partly by the observation that these databases are representative of four of the biggest model organism

[2] Community databases as we characterise them were often preceded by digitalised repositories aimed solely at storing genetic data, for example the *Arabidopsis thaliana* Database (AtDB) and A *C elegans* Database (ACeDB). Set up in the late 1980s, these repositories were soon seen as insufficient to store and efficiently disseminate the amount and variety of data available on these organisms. In the late 1990s, with support from the National Science Foundation, these repositories were reorganised, renamed and relocated across major scientific institutions (such as the Carnagie Institute of Plant Biology in Stanford in the case of TAIR and the California Institute of Technology, the Ontario Institute for Cancer Research, Washington University and the European Bioinformatics Institute in the case of WormBase).

communities to date, as well as being four of the best developed community databases. At the same time, the history of the use of these organisms in biological research is very diverse, involving different fields, networks and research interests. The *Drosophila* community originated in the early 1900s in Morgan's Columbia laboratory with the purpose of producing a genetic map of the fruitfly's chromosomes (Kohler 1994). The mouse community was launched by Clarence Cook Little not long after, but for the purpose of studying human diseases such as cancer (Rader 2004). *C. elegans* and *Arabidopsis* took their places in mainstream biology at later stages, the worm being used originally to investigate the nervous system by Sydney Brenner in 1960s Cambridge (Ankeny 2001), and *Arabidopsis* as a 'botanical Drosophila', the harbinger which brought genetics into plant science, in the late 1970s (Leonelli 2007). While it is not the purpose of this paper to reflect extensively on the diverging histories of these model organisms, we use this background to highlight the relation between current developments and the historical trajectories of the communities involved. In addition to historical literature, our sources and methods include archival research (publications and archives of databases), participant observation at several relevant conferences and interviews with curators and users carried out between 1996–9 (RA on *C. elegans)* and 2000–9 (SL on *Arabidopsis*).

2. Curating Information about Organisms

Community databases store as much information about their target organism as possible through the efforts of their curators and the contributions of the research communities involved.[3] Information is usually donated by researchers, manually extracted from

[3] This is of course within the limits of what each community considers to be 'useful' information. So depending on the history and disciplinary leanings of each model organism community, certain types of data might be excluded: for example, extensive available data from agricultural research is not included in either WormBase

publications by curators or automatically ‘mined’ from repositories or literature through parameters and software set up by curators. Curators work in relatively small groups, usually between five and ten per database, with areas of expertise ranging from experimental biology to bioinformatics. In contrast with other types of databases, particularly data repositories that tend to be managed by software engineers, most of the curators of community databases are biologists by training. Some have experience with studying organisms in the lab, which gives them a deep awareness of the concerns, interests and skills characterising their user community. Familiarity with the everyday needs of database users is considered to be a crucial feature of curators’ work, which is often described as a ‘service’ to the community. Indeed, despite the small size of the curator groups compared to the thousands of researchers populating the communities associated with the relevant databases, curators’ activities are central to the good functioning of community databases, to their long-term maintenance, and to their popularity, perceived trustworthiness and reliability among experimenters.[4]

Curators’ responsibilities include gathering data from publications and/or repositories; annotating these data through appropriate classification systems to facilitate their retrieval by users; creating software for online data retrieval and, in some cases, for automated analyses of its quality and relation to other existing data; linking datasets to information about their provenance, and particularly to stock centres which can provide users with the actual specimens which were used to produce the data; providing general information about the organism (such as techniques for growing and maintaining the organisms) and the people studying it; developing tools for the integration and modelling of disparate types of data; and,

orTAIR, while the Mouse Genome Informatics does not include data collected under restrictive intellectual property agreements (e.g. public-private partnerships with pharmaceutical companies).

[4] How curators came to play such a central role in model organism communities is a complex piece of history which we discuss at length in what follows and yet cannot fully capture within the scope of this paper. Some important elements of this story, such as the birth of ‘biologically competent’ bioinformaticians that accompanied the development of bio-ontologies and community databases, can be found in Garcia-Sancho (this volume) and Leonelli (2010a).

last but not least, building awareness among database users of the importance of sharing information and supporting community infrastructure.

Our analysis focuses on two main types of curatorial activities: (1) the choice of *terminology* to classify data and (2) the selection and provision of information about *experimental settings* in which data are produced, including information about specimens and protocols. Exploring these aspects shows how cyberinfrastructure has become central to defining what a model organism is and how it is to be used, and in turn provides a microcosmic view of changing practices in experimental biology. At the same time, the analysis clearly highlights the continuity between the vision characterising model organism communities from their early histories and the current development of cyberinfrastructure: in many ways, the availability of digital technologies facilitated the implementation of ideas and values that have long characterised this type of research, while also giving them a precise and concrete shape (asnever before possible using earlier forms of communication).

3. Standardising Terminology

The absence of shared terminology has long been a basic obstacle to communication and to the sharing of materials among scientists even when they are interested in the same organism. All four model organism communities on which we focus employed newsletters in the early days of their histories, so as to exchange information, but also to make certain that researchers acquired common terminologies to describe their research, materials and findings. Almost as soon as he arrived in Morgan's laboratory, Calvin Bridges was put in charge of labelling its mutant types of *Drosophila* and started to promulgate his classification system through an informal newsletter which came to be known as the *Drosophila* Information

Service in 1934 (Kohler 1994, pp.73ff). The classification of Arabidopsis ecotypes preceded its adoption as a model organism in plant science, dating back to the 1920s when it attracted the interest of German ecologists. Even before World War II, plant biologists began to circulate the Arabidopsis Information Service (AIS), a newsletter reporting advances in the classification and standardisation of Arabidopsis wildtypes (Meinke and Scholl 2003). The first publication usually associated with the use of *C. elegans* as a model organism was an extensive inventory of mutants (Brenner 1974), and major handbooks to using the organism later became an essential way of communicating and maintaining standards in the field (e.g., Wood and the Community of *C. elegans* Researchers 1988). Finally, as previously discussed, the Mouse Club newsletter was initiated in 1922 by a group of mouse geneticists headed by Little with the intent to share information about mutant stocks and breeding experiments; the publication was renamed the 'Mouse Newsletter' in 1941, and thus officially recognised as a key form of communication among mouse scientists (Rader 2004, pp. 54–56; Rader 1998).

All four newsletters survived well into the 1980s. However, they failed to keep up with the growth in size and scale experienced by these model organism communities over the last two decades. A growing population of scientists located across the globe meant a growing body of information to be exchanged, and created pressing needs for information to travel quickly and efficiently, requirements that could not be met by publications that had to be manually updated, typewritten and sent off through the mail. Even the advent of email communication and online journal availability was not enough to offset these problems: the quantity and variety of information produced by these model organism communities, whose membership in the 1990s was reaching the tens of thousands of researchers, simply could not be disseminated solely through journal publications and manually compiled newsletters. There are clear pragmatic advantages to this form of digital technology, which include ease of access on a global basis, the ability to maintaining and update them dynamically and at relatively low cost, the ability to simultaneously access various types of information for comparison, the open access afforded to all interested researchers and so on, advantages shared by many computerized scientific resources. Hence globalised scientific networks have welcomed databases as an excellent alternative vehicle for capturing and sharing information.

But an additional and key advantage to the new community databases lies in their systematic structures consciously imposed by those who constructed them. The exponential increase in the scale of the research enterprise in fact could have resulted in decreased efficiency of access without the logical structures and other tools which were integrated into the databases from the start. The scientists in charge of compiling the newsletters of old clearly recognized the need for standardized nomenclature as discussed previously. The new database curators built on this principle, as they knew from the start that the key to facilitating knowledge exchange was to categorise information in ways that were accepted and recognised across the community of users, but they also knew that it was essential to agree on a format for presentation and access that would be viewed as reliable and consistent, and that would contribution to the accretion of information and knowledge. In the case of model organism communities, these requirements meant enforcing not just standard (or traditional) terminological choices but also a language that would be intelligible across several disciplines, ranging from genetics to cell biology and physiology, and which would allow translation data from further afield, such as from ecology. The challenge of finding adequate terminology in this new research environment was complicated by the necessity, felt by all four groups of curators, to make their systems interoperable across a wide variety of species, that is, to allow users to compare findings across databases with ease.

This cross-disciplinary, cross-species vocabulary is what the Gene Ontology Consortium set out to achieve with regard to gene products. The Gene Ontology was created in the late 1990s by the curators of community databases as a ‘controlled vocabulary’ to classify data about gene products of *Drosophila*, the mouse and the yeast *Saccharomycae cerevisiae* (Consortium 2000). TAIR and WormBase curators were early members of this group, joining in 2001, which resulted in the development of the Gene Ontology as a common terminology to classify, exchange and compare data about a wide variety of species (Consortium 2008), as a specific instance of what have subsequently come to be known as ‘bio-ontologies’. This system has been very successful both in terms of the number of key databases that adopted it, and the number of species that it encompasses – now numbering in the hundreds. The success of the Gene Ontology has helped the development of several other bio-ontologies within other biological domains, ranging from foundational anatomy to proteins. Bio-ontologies have become basic tools to homogenise terminology so as to make it possible to retrieve and compare data across databases, species and research contexts (Rubin, Shah and Noy 2008).

What is most interesting for our present purposes about the Gene Ontology as a classificatory tool is that curators make critical interpretive choices when selecting bio-ontology terms and their associated datasets. These choices are unavoidable, given the diversity of terminologies adopted in the several fields interested in the available pool of data. Curators have to choose labels that have the highest likelihood of being recognised and unambiguously interpreted by users of their databases. However these choices, which are dictated by the desire to strengthen communication across fields, do at a minimum affect the ways in which researchers learn to present their results to their peers (and are likely to have much more systemic effects in relation to how they come to understand their results). The more popular that these databases have become, the more researchers have needed to make certain that their data are securely and recognisably stored in them. As a consequence, they often come to systematically prefer nomenclature used within the databases to other terminologies, including terms that are used and understood only within their own local contexts. This means that terms privileged by database curators are likely to become predominant as communication tools in cross-disciplinary discussions. These phenomena have been reinforced by the increasing pairing of bio-ontologies and community databases with high-profile journals (as exemplified by the pairing of *Plant Physiology*, a prominent journal in plant science, with TAIR), requiring all submissions of papers for consideration for publication to be accompanied by the submission of data to the appropriate public database. Scientists wishing to submit a paper must annotate the relevant data through the keywords (bio-ontology terms) selected by database curators.[5]

An example from the Gene Ontology illustrates how terminological decisions taken by curators can impact what counts as 'accepted knowledge' within model organism communities. Before *Arabidopsis* was added in 2001, the Gene Ontology had only one term for gamete formation, 'gametogenesis', which was defined as the "generation, maintenance, and proliferation of gametes". Originally the curators had not specified what a 'gamete' was, yet the term and definition were generated with animal gamete formation in mind. Plant biologists, however, use 'gametogenesis' to refer to the generation of a gametophyte, (i.e., a plant in the haploid phase that can produce gametes). Prolonged discussions among and interventions by at least three curators led to an extensive set of changes that both removed ambiguity about the usage of 'gametogenesis' and added terms to accommodate processes specific to plant biology. The definition was altered to specify that "a gamete is a haploid reproductive cell" and to remove all mention of proliferation; for plant processes, the new

[5] An extreme consequence of this process is that claims represented through bio-ontologies become background knowledge for future research and hence may become invisible to future users, particularly those new to the field. This idea is further explored in Leonelli 2010a.

term (‘gametophyte development’) was added. A few years later, the term was renamed ‘gamete generation’, and ‘gametogenesis’ became a related synonym for the term when used to refer to animals as well as plants. It is apparent that these changes, while motivated by the desire to accommodate both plant and animal biology, resulted in shifts in the ways in which animal biologists themselves talked about gametogenesis.[6]

Through classification systems such as the Gene Ontology, databases foster implicit terminological consensus within model organism communities, thus strengthening communication across disciplines but also imposing epistemic agreement on how to understand and represent biological entities and processes. Well aware of the epistemic power of their classification systems, database curators continue to campaign to raise awareness among biologists of how bio-ontologies are built and how experimenters can contribute to that process (Garcia-Hernandez and Reiser 2006; Rhee et al. 2008). Attempts to involve experimental biologists in the building of controlled vocabularies have not been very successful to date, because of the time, effort and expertise involved in understanding how bio-ontologies work. In addition it may well be the case that biologists do not yet fully realise the significance of choosing terms within bio-ontologies and their implications for their research practices.

4. Standardising the Experimental Setting

Terminology is only one of several elements of research practice which must be captured within any functional and useful community database. Tacit knowledge, and the expertise that comes from daily engagement with experimental systems as well as with familiarity with

[6] For more details on this and other examples, see Leonelli et al. (forthcoming). Thanks in particular to Midori Harris for providing SL with this example.

laboratory ‘ways of doing’, somehow must be incorporated in databases, despite the obvious difficulties in formalising and standardising knowledge that is not propositional. The reasons for wishing to incorporate at least some elements of tacit knowledge have to do with the conditions under which scientists can retrieve and use the data found in databases. To interpret data produced by someone else, researchers need to acquire as much awareness as possible of the conditions under which data were originally produced, including the goals of the data collection, instruments and protocols utilized and so on; in other words, they must be able to access information about the data’s provenance (Leonelli 2010b). This requirement means that databases need to incorporate information about the whole life cycle of research, focusing on data production but including experimental methods, goals, materials and instruments through which such production has occurred. When model organism communities were nascent and very small in size, with only one or a few laboratories training all workers, this information floated informally in a reasonably effective manner, at least within any one community. Nowadays, curators must select and synthesise this information through their choices of ‘meta-data’, which are bits of information that reveal those aspects of the provenance of datasets that are considered to be of key relevance for their future use.

Unsurprisingly, there is no consensus within research communities on which meta-data are the most relevant for the proper interpretation of data. The choice depends in part on the particular research contexts and types of data in question, which makes reaching agreement on which meta-data should be included in a community database, and how they should be represented, very difficult to reach. Some basic agreement is possible: for instance, no researcher in the life sciences would deny that knowing from which organism data have been generated is of paramount importance to deciphering the biological significance of those data. Following in this vein of reasoning, several groups of curators have become involved in establishing standards for what should count as ‘essential’ meta-data – a subset of meta-data that are recognised and agreed to play important roles across research contexts. A well-known example of this type of standard setting can be seen in The Minimum Information about Biological and Biomedical Investigation (MIBBI), a project established to help standardise what counts as essential information about an experimental set-up. This project illustrates the impacts that database use, and the related necessity to file meta-data, are having on how researchers think about experiments. Indeed, the curators working on the MIBBI project recommend that in order to be able to describe and classify their experimental procedures within an online format and in ways that are intelligible to other users, researchers should standardise the very procedures that they employ in the lab in the first instance (Taylor et al. 2008). Standardising experimental protocols themselves is seen as the best way to ensure that researchers across the globe can understand descriptions of data provenance even without being familiar with the local setting in which data have been produced.

Of course if this effort is successful, it is likely to have a huge impact in how researchers conceive of local know-how, and the degrees of freedom allowed in exploratory experimentation.[7]At the very least, the implementation of standard ways to describe experimental procedures is likely to result in changes to the experimental set-ups commonly utilized within the community of users, which also might affect the rules of the ‘scientific game’ and what counts as evidence in it. One instance of this happening is the Minimal Information About Microarray Experiments project (MIAME), which was set up as part of the MIBBI initiative to help standardise the format and experimental procedures used for micro-array data. MIAME explicitly targets both the ways in which micro-array experiments are described and the ways in which they are conducted. Its explicit purpose is to set

[7] On the epistemic role and importance of exploratory experimentation, see Burian (1997) and O’Malley (2007).

conditions for the ways in which researchers carry out micro-array experiments, particularly in view of the raging controversy over the reliability and replicability of the data thus produced (for details of the development and impact of MIAME, see Keating and Cambrosio in this same issue as well as Rogers and Cambrosio 2007).

While it is too early to provide an empirical assessment of the effects of MIBBI on model organism research more broadly, the use of one specific type of meta-data in several community databases provides another telling example of how the choice of various types of standards can affect daily laboratory practice. In databases such as TAIR, evidence codes are being used which imply a categorisation of types of evidence on the basis of (1) the ways through which it has been obtained (e.g., through specific experimental techniques, such as high throughput technologies, computer simulation or even word-of-mouth) and (2) their degree of reliability (i.e, the extent to which the evidence is credible and can be assumed to be valid without further verification). This latter index is called the Gene Confidence Rank which aims to give users some measure of how reliable data are by giving different scores to different types of data (e.g., experimentally-obtained data are represented as more reliable than computationally-obtained data), thus establishing a hierarchy of levels of evidence. The implication of applying this system (whose implementation is fully driven by curators) is that experimenters will come to privilege certain forms of evidence over others and strive to provide experimentally-produced evidence according to the ranking established by the Gene Confidence Rank. Curators defend this decision by pointing out that they are simply formalising an evaluative system that is already informally or implicitly active in the community, since most people would, in the case above, find computationally-generated evidence less convincing than experimentally-obtained data. Even if this claim is true, the evidence code system formalises and systematises evidence rankings in much the same way as a set of guidelines: anyone who wants to publish their data in a community database, and have other researchers build on their work, will need to make efforts to produce data of the highest possible ranking.[8]

Another element of experimentation that is affected by the use and structures of community databases is the standardisation and use of biological materials, that is, the actual specimens of organisms on which experiments are conducted. The four communities on which we focus in this paper on are cases in point. The Morgan laboratory started to classify and standardise *Drosophila* specimens and distribute them to other labs in the 1920s; JAX mice have been produced and disseminated by the Jackson lab since the 1930s; *C. elegans* stocks have been available via a formal strain centre since 1978; and *Arabidopsis* collections were established in the 1930s in Germany, where the Laibach collection still constitutes the core of *Arabidopsis* stocks. In all four cases, access to specimens initially was available via paper catalogues, newsletters and informal contacts among research groups. These efforts were usually coordinated by one of the most powerful and well-funded nodes of the budding network, such as the Morgan lab for *Drosophila* and the Jackson lab for mice. However, these ways to exchange materials proved to be unsustainable in the long term, and certainly became unwieldy as soon as model organism communities grew in the 1980s–90s from encompassing a few recognisable groups to large conglomerates of researchers with no direct geographical, personal or disciplinary ties. At that point, a more formalized system was needed that would centralise access to specimens and keep researchers updated on what materials were available from whom, without having to rely on one-to-one contacts and word-of-mouth between laboratories. The *Arabidopsis* and *C. elegans* communities tried to overcome these problems by establishing stock centres that would collect, store and distribute stocks to the whole

[8] And as a result it potentially has many of the same implications (some of which are negative) as have been recognized in the active debates over hierarchies of evidence in medicine, details of which are too lengthy to provide in this context.

community on demand. This type of system did not resolve problems associated with granting access to the ever-growing and ever-changing catalogues of stocks available in each centre. Databases became crucial, since they facilitate the posting of accurate and up-to-date information on stocks and thus promote the selection and obtaining of the ‘right (strain of the) organism for the job’ by interested researchers.

Through direct online links to stock centres, community databases provide access to information on, as well as actual specimens of, the strains of model organisms that have been developed and subjected to experimentation across the globe. Community databases typically have no direct responsibility for how specimens are collected and distributed by the stock centres. Nonetheless they play key roles in supporting the work done at stock centres by offering centralised online access to specimens (Rosenthal and Ashburner 2002). This service requires tight coordination between the ways in which stock centres describe their specimens and the information reported online about them within the databases. Further, database curators have to manually align information about each strain of mutants available in stock centres with the online data actually available in relation to those strains. Because of these collaborative activities, which are essential to the coordination of stock centres and databases to permit systematic choice and use of strains by researchers, the curators of community databases clearly influence the ways in which specimens are described, stored and disseminated to users.

C. elegans and *Arabidopsis* are the most successful examples of close collaborations between stock centres and databases: the *Caenorhabditis* Genetics Center is directly accessible through WormBase, while the two existing *Arabidopsis* stock centres (the Nottingham *Arabidopsis* Stock Centre and *Arabidopsis* Biological Resource Centre) were developed and expanded in the late 1990s in collaboration with TAIR. The fruitfly and mouse communities have generally been less efficient in aligning database development with the standardisation of stocks, primarily because stocks of these organisms have not yet been successfully centralised. In the Drosophila community, collections are greatly diversified and some are privately held. The mouse situation is even more diffuse, as stock collections are highly diverse and mainly in the hands of collections at individual laboratories or institutions. Even (one might say especially) in these situations, community databases play a key role in guaranteeing access to stocks. FlyBase lists all existing *Drosophila* collections, which can then be contacted individually by users for orders (http://flybase.org/static_pages/allied-data/stock_collections.html), while mouse collections can be obtained through a portal called the International Mouse Strain Resource (http://www.findmice.org/index.jsp). The absence of a centralised stock centre with a direct link to Mouse Genome Informatics is the object of heated debates within the mouse community, several members of which have argued that this lack of common access is delaying, and in some cases impeding, research progress (Anonymous 2009; Schofield et al. 2009).

5. Cyberinfrastructure and Research Ethos: Shaping Practices

The two sets of practices analysed above illustrate the nature of the links between model organism communities, their practices, their ethos and their cyberinfrastructure, particularly community databases. As we discussed in the case of terminology and experimental settings, community databases have played important roles in making criteria and values explicit that were previously well-established in the relevant communities, and yet did not previously need to be formalised and standardised due to the size of the communities and their relatively local nature, and the scale of the research, among other factors. In so doing, community databases also are creating new standards and guidelines for what counts as reliable evidence, intelligible nomenclature and acceptable (or perhaps ideal) experimental practice within model organism communities.

Up to the 1990s, limited coordination within each model organism community was achieved among small groups, through newsletters, meetings, personal contacts and so on. The ethos of sharing was strongly established early on in the absence of digital communication technologies. For instance, *C. elegans* researchers published handbooks on worm biology which were authored by 'the Worm Community' (Wood et al. 1988), while the founders of the Arabidopsis community, particularly George Redei and Chris and Shauna Sommerville, enforced the sharing of results at the pre-publication stage of research from the very start of molecular work on the plant in the 1970s (Leonelli 2007). The use of community databases has made it possible to dramatically increase the quantity of information on model organisms that can be stored and integrated, as well as the number of researchers with access to such information. This quantitative shift has brought about a series of qualitative changes in the nature of the community interested in any given model organism and the ways in which members of such communities can communicate with each other.

The most striking of those changes is the fact that membership of any model organism community in some sense has become in principle completely open and inclusive, as any interested biologist can access the database and use its contents. On the one hand, this fact challenges the identity politics which has hitherto characterised model organism work (see Ankeny and Leonelli 2011): researchers do not need to define themselves as 'worm people' or 'mice people' any longer, and have personal contacts or have been trained in key laboratories, in order to be able to access and work with worm or mouse data respectively.

Furthermore, comparative research across different species is fast becoming the norm rather than the exception for users of community databases. On the other hand, the existence and rich contents of community databases also reinforce the need for keeping some model organisms as privileged reference points for comparative research. As argued by plant biologists in defence of the importance of TAIR in their community (Ledford 2010), the wealth of information available on organisms such as *Arabidopsis* and *C. elegans*, as well as the sophistication of retrieval and modelling tools offered within community databases, may provide the strongest incentive to keep carrying out research on these organisms. Thus, by giving visibility to model organism research, as well as supporting its integrative nature, community databases strengthen the need for keeping a tight research focus on the organisms about which we happen to know the most (such as the four organisms discussed in this paper), while at the same time reducing the need for individual researchers to identify with a research community focused only on one organism.

Databases also have become critical mechanisms for division of labour and the fostering of collective trust within and between model organism communities. As previously argued, there was a strong ethos of open sharing of specimens, experimental techniques and data within many model organism communities even prior to the advent of formal databases. The development of standards for terminology and experimental settings as captured in databases reflects a form of ceding of responsibility for these types of activities away from individual researchers or particular laboratories, and even away from the communities as previously conceptualized as informally-organized entities, to the databases as the recognized, formal levels of organisation which promote key community functions. In turn, curators are now the authorities who have the recognized expertise to make decisions about a range of critical issues, for instance about terminology, nomenclature and standards for experimental settings. There are at least two reasons why these databases (and the curators who operate them) have succeeded in taking up this role: first, many curators were previously practicing laboratory scientists, often within the community with which their database is associated, and hence are viewed as members of the community with its interests at the centre of their activities. Far from being outsiders imposing external standards or making scientists conform to certain values for the sake of efficiency or some other goal, curators are viewed as facilitating a project which is explicitly collective in nature. Second, facilitating open access to data through the databases is also fundamentally underpinned by developing various standards which allow this data to be usable both by members of the community and interested others who might join the community in its expanded form. Hence these databases often are seen as public resources (particularly because they are created and maintained largely through governmental funding), in direct opposition to more commercialized genomic undertakings which insist on limiting access to data. The underlying conflict of ethos and perceived attitudes towards public science was of course exemplified by the debate on open access to human genome data which took place in the 1990s – not uncoincidentally the same era in which model organism community databases began to flourish.

The case of research on mice may be seen as an interesting exception to these trends because its funding structure and links with medical research have made it much more competitive and less prone to the sharing of resources than the other three communities examined here. *Arabidopsis* and *C. elegans* research has been lavishly funded by governmental and other public funding bodies, which also supported efforts towards building common, open access stock centres and cyberinfrastructure. By contrast, mouse research has been heavily sponsored by private industry and medical institutions, all of which are under the commercial, competitive pressures of the medical/pharmaceutical sphere and thus typically more interested in having priority in research through patenting and publishing than in collaborative efforts. Largely as a result of these different institutional and disciplinary structures and pressures, the mouse community has been less efficient (as noted above) in establishing formal mechanisms to foster a collaborative ethos than the other three cases we have considered. The difficulties in building and maintaining a collaborative ethos is reflected in the problems that have been encountered by mouse people in establishing and maintaining a cyberinfrastructure to serve the whole community (Schofield et al. 2009).

One final aspect to consider when assessing the qualitative shift brought about by the advent and growth of community databases is the ways in which they have affected the disciplinary landscape in biology. It is undeniable that the speed with which communication happens has changed radically through digital technologies, enabling users to upload updates in real time and access information whenever needed merely with the help of an internet connection, which has clear impacts for discipline definition and building. However, this outcome is not specific or unique to community databases. More importantly, these databases were built to foster integrative, interdisciplinary research; although these values were implicit previously within these research communities,[9] the structures within the databases made them explicit and implementable. Databases now enable coordination of research among huge numbers of researchers in diverse disciplines and locations. They function as a showcase of interdisciplinary research for outsiders by making results and discoveries highly visible to anyone interested in them, and by improving the ease with which collaboration among model organism groups (and other researchers, including those doing biomedical or clinical research) can be established. By centralising resources and access to data, and fostering a focus on specific organisms, they actually facilitate the decentralisation of research across different disciplines and species.

[9] And not necessarily espoused by any one individual researcher, but oftentimes underlay broader mandates or rationales for research given by founding members of that community on behalf of the community as a whole.

Databases thus are a key link between the tradition and the future of model organism research, a natural consequence of the collaborative ethos that characterised work on model organisms from its inception. At the same time they provide an ongoing powerful incentive to cultivate that ethos and reap its fruits through integrative and comparative research across species.

6. Conclusion

Within the last two decades, community databases have acquired foundational roles within model organism communities. They have become indispensable providers of access to available knowledge about each organism in the form of data, publications, specimens and tools. As aptly stated in a review on model organism community databases: 'The efforts of model organism database groups ensure not only that organism-specific data are integrated, curated and accessible but also that the information is structured in such a way that comparison of biological knowledge across model organisms is facilitated' (Bult, 2006, 28). Community databases thus play pivotal roles in defining the epistemic roles of model organisms in contemporary biology. They greatly facilitated the fulfilment of several of the premises on which model organism communities were founded: the collaborative ethos and the willingness to share results so as to understand organisms as 'wholes', the related need to exchange materials and specimens and the resulting efforts to standardise nomenclature and experimental protocols so as to make exchanges as seamless and global as possible. Occupying this niche allowed community databases to demonstrate how fruitful collaboration around and across model organisms could be, thus effectively establishing and reinforcing the very notion of a 'model organism' as a protagonist of late 20th and early 21st century science. This process was a continuation of the history of model organism research since the turn of the 20th century, and resulted in an acceleration of the building of these communities and their expansion into a global research force (and a considerable share of the funding available for biological and even biomedical research).

On the one hand, model organism databases certainly reinforced the power of popular model organisms over other organisms with less well-organized communities. Indeed, the popularity and usefulness of particular model organisms has tended to grow incrementally with the scale and organisation of their community databases , a principle readily recognised by biologists wishing to promote new organisms as ‘biology’s next top models’, according to whom obtaining funding to build a community database is a crucial step in the process (Maher 2009; Behringer et al. 2008), and by researchers wishing to highlight the usefulness of model organism research for the future study of human disease (Spradling et al. 2006) and evolutionary developmental biology (Sommer 2009). At the same time, however, it is important to note how the expansion of community databases has shifted attention away from research on single species to comparative, cross-species research. Model organism databases are taken as reference points for the investigation of other species in the same family or kingdom about which less is known. For instance, TAIR also is a very popular tool among researchers focused on crops or trees, because it provides a reference point for how specific processes (such as vernalisation, cell metabolism or root development) might work and which genes and biochemical pathways might be involved. The WormBase is increasingly used by researchers working on nematodes other than *C. elegans,* such as those which are significant agricultural or human parasites.

Hence community databases have made essential contributions to the development of an understanding of model organisms as tools which permit *comparison* across species: research on model organisms began in part to provide reference points for such comparison, and the use of databases has enhanced their capabilities to act as such reference points. Of course, the strength of model organisms as comparative tools lies in their capacities to *represent* specific groups of organisms, as well as to enable cross-disciplinary research programmes exploring several different aspects of their biology, with the ultimate goal of reaching an *integrative* understanding of the organisms as intact wholes. Cyberinfrastructure, in the form of databases and thus of the communities, specimen collections and information to which they provide access, is the platform over which model organisms can now define themselves as comparative, representative and integrative tools. Without cyberinfrastructure, the exchange of information about model organisms and their use for comparative purposes would be impossible to realise on the appropriate scale, given the integrative goals of many contemporary biological research programs. By bringing results, people and specimens together using infrastructure, community databases have come to play a crucial role in defining what counts as knowledge of organisms in the post-genomic era. Thus, we argue, they are an integral part of what defines what counts as a ‘model organism’.

Acknowledgments

This research was funded by the ESRC and the British Academy (SL) and by the Faculty of Humanities and Social Sciences at the University of Adelaide (RAA). The authors wish to thank numerous scientists who spent time talking with them about their work, the participants at the workshop ‘Data-driven research in the biological and biomedical sciences’ (Exeter, April 2010), and the attendees at presentations at the Society of Philosophy of Science in Practice biennial conference (Minneapolis, June 2009) and the International Society for the History, Philosophy, and Social Studies of Biology Biennial Meeting (Brisbane, July 2009) for their comments and suggestions.

References

Ankeny, R.A. (2001). The natural history of *Caenorhabditis elegans* research. *Nature Reviews Genetics 2,* 474–8.

Ankeny, R.A. & Leonelli, S. (2011). What’s so special about model organisms? *Studies in the History and the Philosophy of Science 42*(2), 313-323.

Anonymous. (2009). The sharing principle (editorial). *Nature, 459*, 752.

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

Behringer, R. R., Johnson, A.D. & Krumlauf, R.D. (Eds.). (2008). *Emerging model organisms: A laboratory manual. Volume 1.* Cold Spring Harbor Laboratory Press.

Brenner, S. (1974). The genetics of *Caenorhabditis elegans*. *Genetics, 77*, 71–94.

Bult, C. J. (2006). From information to understanding: the role of model organism databases in comparative and functional genomics. *Animal genetics,* 37 *Suppl 1*, 28–40.

Burian, R. M. (1997). Exploratory experimentation and the role of histochemical techniques in the work of Jean Brachet, 1938–1952. *History and Philosophy of the Life Sciences*, *19*, 27–45.

Garcia-Hernandez, M., & Reiser, L. (2006). Using information from public Arabidopsis databases to aid research. *Methods in molecular biology (Clifton, N.J.), 323*, 187–211.

Gene Ontology Consortium. (2008). The Gene Ontology project in 2008. *Nucleic Acids Research*, *44*(0), 1–5.

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

Kohler, R.E. (1994). *Lords of the fly. Drosophila genetics and experimental life*. University of Chicago Press.

Jonkers, K. (2009). Models and orphans; concentration of the plant molecular life science research agenda. *Scientometrics*.

Ledford, H. (2010). Plant biologists fear for cress project. *Nature, 464*(11), 154.

Leonelli, S., Harris, M., Diehl, A., Chris, K. & Lomax, J. (forthcoming). How the Gene Ontology evolves. BMC Bioinformatics.

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

Leonelli, S. (2010b). Packaging data for re-use: Databases in model organism biology. In Howlett, P. & Morgan, M. S. (Eds.). *How Well Do Facts Travel? The Dissemination of Reliable Knowledge*. Cambridge, MA: Cambridge University Press.

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

Maher, B. (2009). Biology's next top model? *Nature, 458*(9), 695–698.

Meinke, D. & Scholl, R. (2003). The preservation of plant genetic resources. Experiences with Arabidopsis. *Plant Physiology, 133*, 1046–1050.

O'Malley, M.A. (2007). Exploratory experimentation and scientific practice: Metagenomics and the proteorhodopsin case. *History and Philosophy of the Life Sciences*, *29*, 337–360.

Rader, K. (1998). 'The mouse people': Murine genetics work at the Bussey Institution of Harvard, 1910–1936. *Journal of the History of Biology*, *31*(3), 327–354.

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

Rhee, . Y., Wood, N., Dolinski, K., & Draghici, S. (2008). Use and misuse of the gene ontology annotations. *Nature Reviews Genetics,* 9(7), 509–515.

Rhee, S. Y., & Crosby, B. (2005). Biological databases for plant research. *Plant Physiology, 138*(1), 1–3.

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

Rosenthal, N. & Ashburner, M. (2002). Taking stock of our models: The function and future of stock centres. *Nature Reviews Genetics, 3*, 711–717.

Rubin, D. L., Shah, N. H., & Noy, N. F. (2008). Biomedical ontologies: A functional perspective. *Briefings in Bioinformatics, 9*(1), 75–90.

Schofield, P. et al (2009). Post-publication sharing of data and tools. *Nature, 461*(10), 171–173.

Sommer, R.J. (2009). The future of evo-devo: Model systems and evolutionary theory. *Nature, 461*(10), 416–422.

Sommerville, C., & Koornneef, M. (2002). A fortunate choice: The history of *Arabidopsis* as a model plant. *Nature Reviews: Genetics*, *3*, 883–889.

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.

Taylor, C. F., Field, D., Sansone, S., Aerts, J., Apweiler, R., Ashburner, M., Ball, C. A., et al. (2008). Promoting coherent minimum reporting guidelines for biological and biomedical investigations: The MIBBI project. *Nature biotechnology*, *26*(8), 889–896.

Wood, W. B., & the Community of *C. elegans* Researchers (Eds.). (1988). *The nematode* Caenorhabditis elegans. Cold Spring Harbor: Cold Spring Harbor Press.