Chapter for inclusion in the volume 'How Well Do 'Facts' travel', Howlett P and Morgan MS (eds.), Cambridge University Press, 2010.

Packaging Small Facts for Re-Use: Databases in Model Organism Biology

Sabina Leonelli
s.leonelli@exeter.ac.uk

Introduction

Model organism such as fruit-flies, mice and zebrafish are the undisputed protagonists of 21st century biology. Their prominent position as experimental systems has been further enhanced by the recent sequencing of their genomes, which opened up new opportunities for cross-species comparisons and inferences (the so-called 'post-genomic era'[1]). Such comparative research requires that facts about model organisms be able to travel across a multitude of research contexts. Indeed, the very idea of focusing on a limited set of organisms stems from the desire to bring together as many facts about these organisms as possible, in the hope to increase the scientific understanding of their biology and thus use them as representatives for the study of other species. Moreover, the high costs associated to the production of facts make their use beyond their context of production into an economic, as well as a scientific, priority.

Fulfilling this goal is complicated by the diversity of disciplinary approaches, methods, assumptions and techniques characterising biological research. Each research group tends to develop its own epistemic culture, encompassing specific skills, beliefs, interests and preferred materials.[2] Further, biologists tend to adapt their methods and interests to the features of their organism of choice, thus amplifying the existing diversity among research communities.[3] This pluralism in approaches makes it difficult to make facts travel to contexts other than the one in which they have been produced, as researchers do not share a common terminology, conceptual apparatus, tacit knowledge or set of instruments. The global nature of biological research makes travel even harder: not only facts need to cross disciplinary and cultural boundaries, but they also

[1] For information on genomics, see Dupré and Barnes (2008); on the epistemology of model organism research, see Ankeny (2007).

[2] The de facto pluralism characterising biology has been widely discussed in the social and philosophical studies of science (eg Mitchell 2003, Knorr Cetina 1999 and Longino 2002).

[3] It is common practice to name communities in experimental biology on the basis of the organism that they study (as in 'the worm community', denoting the ensemble of biologists using *Caenorhabditis elegans*, and 'the Arabidopsis community', using the plant *Arabidopsis thaliana*).

need to travel great distances, becoming accessible to biologists regardless of their geographical location.

Most of the current work in the field of bioinformatics is devoted precisely to resolving the tension between the local nature of facts about organisms and the need for them to circulate across widely different research contexts and locations.[4] Bioinformaticians, and particularly database curators, use digital technology to package facts for travel. Their work is defined by the need to serve a wide variety of database users across the globe, each looking for data fitting their own interests and methods. To a curator, successful travel is marked by the re-use of facts within new research contexts. However, making facts available online does not automatically involve making them usable: whether facts are adopted across contexts is the result of packaging strategies developed by curators through years of specialized training and dialogue with users.

This chapter examines these packaging strategies to address a key question in contemporary science: what counts as successful re-use of facts? This question lies at the heart of the study of travelling facts, which needs to discuss not only the conditions for the journey of facts from one realm to another, but also the conditions for their acceptance or rejection upon arrival in a new context. This is also the question that curators have to answer when packaging facts for travel. As I shall illustrate, 'good packaging' consists of developing labels that facilitate the retrieval and adoption of facts by prospective users. More specifically, facts need to be de-contextualised from their original locus of production, while at the same time retaining 'travelling companions' to facilitate their re-contextualisation into new research settings.[5] Balancing these two requirements against each other is not easy; nor, given the ever-changing nature of the facts and practices involved, are there universal and enduring ways to compromise between de-contextualisation and re-contextualisation. It is therefore curators' responsibility to constantly update their work to reflect the nature and potential destinations of travelling facts. The result is a dynamic process, whose functioning depends on the degree to which curators manage to capture the changing wishes and constraints of practicing biologists.

1. **Small and Big Facts About Organisms**

As a starting point for my discussion, I differentiate between two types of facts typically found in model organism biology. The first is what I call *big fact*. This is the type of fact that attracts the attention of most scholars of science, since it

[4] Mansnerus, this volume, offers an account of circulation of knowledge in the context of modelling which differentiates between the effects of such circulation at the receiving end, while Schneider, this volume, in discussing architectural style pays more attention to the ways in which receivers deal with facts.

[5] The problem users face in choosing a relevant information set parallels the problem of defining when comparisons are meaningful, or effective, to make facts travel, discussed by both Burkhardt and Ramsden, both this volume.

constitutes what is generally seen as the end-result of scientific research: knowledge about the world that help us to interact with it. Big facts in biology usually consist of a more or less general description of biological entity, one of its components or one of its functions (as in 'gene X regulates the development of trait Y'). They are expressed propositionally and travel mostly through publication in academic journals, though some of them reach vast non-academic audiences.[6] In this chapter I focus on the travel of the second type of facts, which I call *small facts*. These are the physical traces left by an experimental apparatus, such as images, numbers, dots on a slide and material objects (such as stains on an embryo resulting from in situ hybridisation). Small facts almost never travel beyond scientific circles and are not expressed in propositional form.

Especially in genomics, increasing quantities of small facts are produced in digital formats (i.e. XML files capturing DNA sequences) or reformatted so as to be exchangeable through the Internet (e.g. stained embryos are photographed and circulated as in the form of images) – indeed, these are the ones I shall be focusing on, as they are the ones that are most easily incorporated into digital databases.[7] Whether they are digital or not, small facts remain essentially material objects: they are the physical result of an interaction between a researcher working with specific instruments, and a biological sample like a tissue, a cell or a whole organism. Their physicality, which determines the ways in which they can be used within the context of experimental research, is a major factor setting them apart from big facts: small facts constitute the material grounds on which knowledge claims (big facts) about biological entities are extracted and validated. Another big difference between big and small facts lies in their ability to travel *solo*. Each big fact has a distinct individuality and can travel alone as well as with other big facts – as a headline in a newspaper or as a line in a textbook. Small facts tend instead to travel in groups (e.g. 'datasets'). One small fact does not usually have much evidential weight. To become significant in a research context, small facts find strength in numbers: the more small facts are grouped together, the strongest their identity and evidential value.[8]

My definition of small facts encompasses anything that biologists might refer to as data. It is similar to the definition of data as 'marks' provided by philosopher Ian Hacking,[9] insofar as it is a procedural definition: it characterises as a small fact anything that has been produced by an instrument under laboratory or field conditions, while differentiating these traces from the propositional statements

[6] Other chapters in this volume focus specifically on what happens to big facts that travel beyond the scientific community. See for instance the contributions by Adams, Ramsden and Oreskes.
[7] Communicating facts effectively is a considerable challenge in any science, see Merz, this volume, for how facts are communicated in images.
[8] This can be contrasted with a somewhat different pattern presented by Ankeny, this volume, whereby medical cases act as a vehicle to get particular facts to cohere together to make generic facts.
[9] 'Uninterpreted inscriptions, graphs recording variation over time, photographs, tables, displays' (Hacking 1992, p. 48)

(descriptions, explanations, hypotheses) made when trying to interpret their significance. To define small facts solely through the procedures through which they are produced might seem too broad, because it encompasses a huge variety of objects, as well as too restrictive, because it does not take into account the various degrees of preparation underlying the production of different types of small facts.[10] For the purposes of this paper, however, I gloss over the significant differences among types of experimental results, and focus instead on the common status that they enjoy among practicing scientists: that is, the status of 'raw data' used to validate or discredit hypotheses.[11]

As widely documented within the history, philosophy and social studies of science, the production of small facts is highly regimented and includes various types of interventions before, during and after any experiment. Small facts could not be produced without recourse to both tacit and articulated knowledge. Further, the conditions for the production of small facts are carefully engineered on the basis of specific expectations, interests, hypotheses, experimental settings and instruments. The production of small facts aims at the validation of big facts: strictly speaking, there are no such things as raw data. At the same time, however, small facts exhibit a biological significance that transcends their role as evidence for a specific experimental hypothesis, and justifies their treatment as raw data by researchers. This is because the experimental context in which they are produced does not wholly determine their evidential value. As intuited by Pierre Duhem (1974 [1914]) already a century ago, small facts are the result of experimenters' interactions with real entities. No matter how tightly controlled an experiment is, or how well-known the entities already are to scientists, what small facts end up revealing about those entities is not wholly predictable, nor can it be entirely captured by any single big fact. In Duhemian terms, the evidential value of small facts is underdetermined: small facts exhibit more or less significance depending on the context in which they are used.[12]

Model organism biology well exemplifies the underdetermined evidential value of small facts. In that context, the same set of small facts can often be used as

[10] For instance, Hans-Jörg Rheinberger (unpublished) provides an illuminating examination of the relation between traces and data in the case of sequencing, where he distinguishes between the sequence gel produced by radioactive tracing and the polished version of those marks (the chain of letters widely known to represent nucleotide sequences) regarded as the official and 'transportable' result of the experiment.

[11] As I discuss below, there are many cases in biology where small facts produced in an experiment remain unused. This does not affect my discussion: regardless of whether they end up being used or not, small facts are always produced in the hope that they may serve as evidence for one or more claims.

[12] As in the case of the seminal work by Bogen and Woodward (1988) on the relation between data and phenomena, philosophers have tended to forget some of Duhem's lessons and focus solely on the evidential value held by small facts within their context of production. It is assumed that small facts are always created to function as evidence for a given big fact (an hypothesis or claim in need of testing); and that their significance is tied to the context in which they were originally produced, as small facts can only be interpreted by researchers who are familiar with every detail of the setting in which they were created. I critique these views in Leonelli (2009a).

evidence for a variety of big facts. Even more strikingly, small facts about organisms are not always created to serve as evidence for a *specific* big fact.[13] Often they are created because biologists have acquired new instruments enabling them to obtain information about entities of interest ('high-throughput technologies', thus named because of their ability to produce vast datasets in a very short time). In these cases it is not obvious how that information should be interpreted once it is produced. An example of this type of data-driven research is the shot-gun technique used to sequence genomes, which produces billions of data points awaiting analysis and eventual interpretation in the form of big facts.

It is to capture the underdetermined evidential value of small facts that I wish to focus on the procedures, rather than on the theoretical lens, through which they are produced. In the eyes of researchers, these procedures are the most important characteristic of small facts. They define the type of intervention through which small facts are obtained, the type of entity on which such intervention is carried out and the physical appearance of small facts, which in turns determines the modalities through which they can be transported to new places and used as evidence for new claims. The procedures through which small facts are produced make them unique as a source of information. And indeed, as I show in section 4, it is information about these procedures that ends up functioning as their 'travelling companions'.

2. Packaging in Bioinformatics

Let us now examine the strategies used by database curators to make small facts travel. To place curators' work in context, I should note that traditional scientific institutions, including funding bodies and the peer review system, have hitherto favoured big facts as the preferred outcome of scientific research. Accordingly, the commonly accepted measure of a scientist's worth is her ability to discover and publish new big facts. Further, publications in scientific journals are not good vehicles for the travel of small facts: small facts are only included as proof that the big fact of interest has been empirically tested. This may seem logical, as small facts are only valuable insofar as they are used as evidence for big facts. However, this system does not take the underdetermined evidential value of small facts into account. Given the typically short length of scientific papers, most data obtained in any single experiment are not selected for publication and are discarded without any opportunity to be of use. Further, the small facts that are actually published are classified as evidence for a single big fact. This means that interest in that big fact becomes the only means to find and retrieve those small facts – a situation not exactly conducive to their adoption by different research contexts. Thanks to this publication system, most small facts are either thrown away or untraceable to anyone who has no direct interest in the project that used them first. As a result, the evidential value of small facts is not

[13] On data-driven research versus hypothesis-driven research, see Kell & Oliver (2003), Krohs & Callebaut (2007) and Rheinberger (unpublished).

maximised.

Biologists are well aware that this communication regime makes small facts unusable to anyone other than their producers and their closest peers. This is why databases are gaining attention as a system devoted exclusively to the disclosure of small facts, which complements vehicles for the travel of big facts (such as journals).[14] Making small facts travel requires apposite infrastructure: vehicles that physically store small facts and transport them outside of the context in which they have been produced. Given the sheer size and diversity of datasets to be circulated, these vehicles must be capable of storing and organising large amounts of small facts. They need to be accessible to users with disparate expertises and interests, which requires a user-friendly interface and the possibility to choose among several types of searches. Further, they should enable users to quickly scan through the available small facts for any given area, and perform comparisons among datasets to find possible correlations.

Online databases have the potential to meet all of these requirements. They are available through the Internet, which minimises the constraints imposed on the size and types of small facts travelling through them and eliminates the efforts and time involved in making them physically accessible around the world. Many databases, especially the ones developed through public funding, can be accessed free of charge or other restrictions. They can and often do provide differential access: thanks to the flexibility of digital interfaces, users can choose parameters for their queries depending on their interests and expertise. Moreover, their computational capabilities mean that they can incorporate tools for automated data analysis, which helps users to check for correlations and patterns across datasets – an indispensible help when needing to restrict one’s search from billions of small facts to a manageable sample.

When looking at databases as vehicles for small facts, it is easy to take the metaphor of ‘packaging’ seriously. This is because the process of packaging small facts for dissemination bears remarkable similarities to the process of packaging items to be dispatched through the mail. The tractability of travelling items is crucial in both cases: standard shape and dimensions help the packaging and circulation of the mail just as they help the packaging and circulation of small facts. Indeed, curators are involved in wider efforts within the biological community to standardise formats for different types of small facts.[15] Further, small facts are objects whose ability to travel depends on material aids and infrastructure designed for this purpose, as well as interventions by people other than their senders and receivers. Human activities and material

[14] For the idea of publications and databases as pertaining to two separate communication regimes, see Hilgartner (1995).

[15] The progressive standardisation of the format of small facts has greatly helped curators’ efforts, as standardisation simplifies the process of grouping different types of small facts within the same databases. For a detailed study of the issues involved in the standardisation of data formats, see Rogers and Cambrosio (2007) on the case of micro array data.

environments are equally important to the travel of small facts. Post offices, trucks, drivers, postmen and mail sorters play a similar role to databases and their curators. There would be no travel without the digital platform provided by databases and the work put in by curators to design and use them as a vehicle for small facts. And just as the mail is a service designed to satisfy senders and receivers, the need for small facts to travel is generated by the laboratory cultures in which data are produced and re-used in another setting.

3. Two Types of Labels

There are also important differences between packaging an object for express delivery and packaging a small fact for dissemination, and it is these differences that this chapter aims to explore. In both cases, whether travel is successful depends on whether what is packaged arrives at its destination without being damaged or lost; and which destination this will be depends on the way in which objects are labelled. However, in the case of small facts, labels *should not determine* the destinations to which the facts will travel. There is no doubt that the labels chosen by curators have a strong influence on the direction that small facts will take. This is unavoidable, since the function of labels is precisely to make small facts retrievable by potential users. Without labels, facts would not travel at all. Yet, for successful re-use to take place, the journey that small facts ultimately undertake should be determined as much by their users as it is by their curators. Curators cannot possibly predict all the ways in which small facts might be used. This would involve familiarity with countless research programmes around the world – as well as a degree of scientific understanding and predictive ability that transcends the abilities of one individual or group. Therefore, the best way to explore and maximise the evidential value of small facts is to enable as many researchers as possible to use small facts in their own way and within their own research context.

Given these premises, labelling becomes the most challenging component of the packaging process. Curators are required to create labels that, while making small facts retrievable by database users, do not prevent users from making their own selection of which small facts they wish to pick. These labels need to indicate the information content of small facts without adding indications – such as a mailing address – about where the facts could be delivered. Giving small facts the flexibility to travel wherever they might be needed constitutes a crucial characteristic of their packaging, which makes it much more sophisticated than the packaging of object for travel to an already well-defined destination.[16]

[16] In this respect the focus of my analysis differs from the analysis of circulation offered by historians Kapil Raj (2007) and Mary Terrall (2008). They focus on the mediation strategies used by early-modern travellers to exchange facts between Europe and the East; and on cases where wished-for objects (eg, specimens) are delivered to a scientific destination (a naturalist's collection). While this approach highlights a relocation of objects across space and time, it seems to capture the activity of collecting rather than the activity of circulating evidence, as it does not emphasise the possibility that travelling objects be used in unexpected ways depending on the

Choosing appropriate labels to classify small facts is crucial to their successful re-use. Within this section, I intend to demonstrate just how difficult a task this is for curators.

3.1 Relevance labels

Enhancing the facts' usability involves making them visible and accessible to as many researchers as possible. Curators thus label small facts in a way that make them attractive to users in new contexts: that is, according to their relevance to investigating biological entities. This labelling system, known as 'bio-ontologies', consist of a network of terms, each of which denotes a biological entity or process. Small facts are associated to one or more of these terms, depending on whether they are judged to be potentially relevant to future research on the entities to which the terms refer. For instance, gene VLN1 has been found to interact selectively with an actin filament known as F-actin (Huang et al 2005). This is an interesting finding given the crucial role played by the actin protein in several cellular processes, including motility and signalling. Still, the actual functions of VLN1 are still unknown: apart from its interaction with F-actin, there are no big facts yet to associate to the small facts about VLN1. Database curators tracked the available small facts about VLN1 and they classified them under the following terms: 'actin filament binding', 'actin filament bundle formation', 'negative regulation of actin filament depolymerisation' and 'actin cytoskeleton'. Thanks to this classification, users interested in investigating these processes will be able to retrieve the small facts about VLN1 and use them to advance their understanding.

Depending on which entities they aim to capture, there are many bio-ontologies in use in contemporary bioinformatics.[17] One of the most popular ones, from which I took the example above, is the Gene Ontology, which encompasses three types of biological objects: cellular processes, molecular functions and cellular components (Ashburner et al 2000). Since their introduction in the late 1990s, bio-ontologies have come to play a prominent role in databases of all types, ranging from genetic databases used in basic model organism research to medical databases used in clinical practice (Augen 2005, p. 64). One of the main reasons for this success is the way in which bio-ontology terms are chosen and used as labels for the classification of small facts.

Curators select these labels according to two main criteria. The first criterion is their intelligibility to practicing biologists, who need to use those labels as keywords in their data searches. In a bio-ontology, each biological entity or process currently under investigation is associated with one (and only one) term. This term is clearly defined so that researchers working in different areas can all

contexts through which they travels.

[17] I restrict the present discussion to the bio-ontologies listed in the Open Biomedical Ontologies consortium (Smith et al 2007).

understand what it is supposed to denote.[18] Often, however, different groups use different terms to formulate big facts about the same entity. This makes it difficult to agree on one term that could be used and understood by everyone interested in that entity – and in the small facts relevant to its study. Curators resolved this problem by creating a list of synonyms for their chosen label. This means that researchers can search for small facts associated to a given entity both by using the official label employed by the bio-ontology and by using one of the listed synonyms for that label: either way, they will be able to retrieve the small facts associated to the entity of interest.

The second criterion for the selection of labels is their association with datasets. The idea is to use only terms that can be associated with existing datasets: any other term, whether or not it is intelligible to bio-ontology users, does not need to be included as it does not help to classify small facts. Curators create an association between a dataset and a term when they have grounds for assuming that the dataset provides information about the entity denoted by that term. This happens mainly through consultation of data repositories, where small facts are categorised as resulting from the experimental manipulation of the entity denoted by the term (e.g. sequence data: small facts about the molecular composition of specific stretches of DNA); and of publications using data as evidence to establish a big fact about the entity denoted by the term (as for VLN1 data; I discuss this case further in the next section).

Thanks to bio-ontologies, researchers can check which small facts might be relevant to the object of their research. The focus on objects rather than methods or specific traditions makes it easier for researchers to bridge across the epistemic cultures in which small facts are originally produced. In this way, researchers with widely different backgrounds (in terms of methods and instruments used, discipline or even theoretical perspective) can access the same pool of small facts and assess their relevance to their research. This enormously increases the chance that database users spot small facts produced in other fields that are relevant for their own research purposes. It therefore becomes more likely that the same small facts are used as evidence towards the validation of various big facts about the same entity. Thus, labels such as bio-ontologies constitute a promising first step towards the packaging of small facts for successful re-use. They are not, however, sufficient for this purpose.

As I discussed in the previous section, the evidential value of small facts is underdetermined. This of course does not mean that any small fact can be used as evidence for any big fact. On the one hand, the successful re-use of small facts depends from the information that researchers manage to extract from

[18] For example, nucleus is defined as ‘a membrane-bounded organelle of eukaryotic cells in which chromosomes are housed and replicated. In most cells, the nucleus contains all of the cell’s chromosomes except the organellar chromosomes, and is the site of RNA synthesis and processing. In some species, or in specialised cell types, RNA metabolism or DNA replication may be absent’ (Gene Ontology website, February 2008).

them. For instance, consider the famous case of the DNA photographs taken by Rosalind Franklin in 1952 and reviewed by James Watson without Franklin's permission. From a quick glance at photograph 51, Watson was able to see evidence for his ideas on DNA structure, while Franklin, who did not share those ideas and was a more careful experimenter, did not interpret the image in the same way – a divergence that arguably led to Watson and Crick being credited with the discovery of the double helix in 1953. Franklin was not simply 'wrong': she used her own interpretation of the images as a guide to excellent work on viruses (Maddox 2002). This episode illustrates that there is no single 'right interpretation' of small facts. Interpretation depends on a users' background and interests, which again highlights the need for curators to package facts in ways that enable the emergence of local differences in interpretation.

3.2 Reliability labels

On the other hand, the emergence of such differences in interpretation, and thus the successful re-use of small facts, depends on the users' awareness of the experimental procedures through which the small facts were originally produced. As I stressed above, these procedures define several important characteristics of small facts. Their format, the actual organism used in the experiment, the instrument(s) with which they were obtained, the laboratory conditions at the time of production: all these elements are crucial in determining the quality, and thus the reliability, of small facts. This means that in order to re-use data found through a database, users need to be able to check, if they so wish, the conditions under which small facts have been obtained.

This is why curators devised a second type of label to classify information about the provenance of small facts. These labels are referred to as 'evidence codes' and they provide essential information about the procedures through which small facts are produced. They include categories for data derived from experimental research, as in IMP (Inferred from Mutant Phenotype), IGI (Inferred from Genetic Interaction) or IPI (Inferred from Physical Interaction); data derived from computational analysis, as in IEA (Inferred from Electronic Annotation) or ISS (Inferred from Sequence Similarity); and even information derived from informal communication with authors (TAS – traceable author statement) and intervention by curators (IC – inferred by curator). Evidence codes are associated to each set of small facts that shares the same provenance. Once users have found facts they are interested in, they can click on the related evidence code and start to uncover the procedures through which the facts have been produced.

Without this second type of labels, and the information retrieved through them, small facts could hardly be re-used. First, researchers who are interested in their evidential value would not necessarily be convinced of their reliability. The reliability of small facts is a function of who produced them, for which reasons and in which setting. Without access to this information, there is no justification for users to trust the small facts displayed in a database. Second, without knowing where data come from, users would not know how to align those facts with the evidence they already have. Adding a new set of data to an existing research project means having some means of comparing the new facts with the facts already produced, especially in case of non-overlapping or even conflicting information. Drawing such a comparison means, in turn, being able to evaluate the similarities and differences between the procedures through which the two sets of facts have been produced. Knowing that both sets have been obtained from the same type of organism (for instance, fruit-flies) would enhance a user's willingness to treat all facts at his disposal as compatible. Finding that one set of facts comes from experimental research, while the other is predicted through simulation, will instead warn the user that the two sets might not have the same evidential value.

4. Good Packaging: De-contextualisation for Re-contextualisation

What makes databases into good packages for small facts is the opportunity afforded to their users to evaluate both the *relevance* and the *reliability* of the facts in question. The two labelling systems enable users to disentangle the activity of searching and comparing data from the activity of assessing the reliability and significance of data. Thanks to bio-ontologies, researchers accessing a database can find out which existing datasets are potentially relevant to the study of the entities and processes in which they are interested. Once they have restricted their search in this way, they can use evidence codes to examine information about data production. This second type of labels enables them to assess the reliability of the data that they located through bio-ontologies, and eventually to discard data that are found wanting according to the users' epistemic criteria.

Remarkably, the consultation of evidence codes does not necessarily reduce the existing gap (if any) between the epistemic cultures of the producers and the users of the facts. Users get access to as accurate a report as possible about the conditions under which small facts were originally obtained. This does not mean that they need to know and think precisely what the producers know and think about those small facts. Rather, the consultation of evidence codes enables users to recognise disagreements with producers concerning suitable experimental conditions, to reflect on the significance of such disagreements and to form their own opinions on the procedures used to obtain small facts. Any judgements on the reliability of data necessarily depends on the user's viewpoint, interests and expertise – which is why curators abstain as much as possible from assessing the quality of small facts, and chose a labelling system allowing each user to form her own opinion.

On the basis of these insights, I argue that packaging small facts for successful re-use involves two complementary moves. The first move, for which database curators are entirely responsible, involves the *de-contextualisation* of small facts from their context of origin.[19] The labelling of facts through bio-ontologies ensures that facts are at least temporarily decoupled from the local features of their production, which enables users to evaluate their potential relevance to their research purposes without having to deal with a chaotic sea of information. When choosing and applying bio-ontology terms, database curators operate in ways similar to librarians when classifying books, or archivists when classifying documents: small facts are labelled so that users coming to the database can use those classificatory categories to search for a content relevant item and borrow it for their own purposes.

Identifying which facts to borrow from a database is a crucial first step for researchers interested in using them. Yet it does not help them to decide how to use the facts once they have borrowed them. In other words, while helping to de-contextualise small facts for circulation, bio-ontology labels do not help to re-contextualise small facts for use in a new research setting. This *re-contextualisation* is the second move required for the successful re-use of data, and it is achieved with the help of the second type of labels, the evidence codes. In selecting this second type of labels, the analogy between curators and librarians falls through, as libraries do not generally need to provide information about the circumstances in which a book or document was obtained in the first place. Indeed, the classification of information about the provenance of data is a rather different process from the classification of their potential biological relevance. It is a genealogical exercise in which curators investigate and reconstruct the sources and history of the small facts that they annotate. Small facts are material objects that need good travelling companions in order to be adopted and re-used across contexts. Evidence codes give access to the qualifications that endow small facts with what Mary Morgan, in her introduction to this volume, calls ‘character’.

Let us explore this idea in more detail. De-contextualisation through bio-ontologies is a way for small facts to lose the personality attributed to them in their original research context: the whole point of de-contextualisation is to make small facts extremely adaptable, which can only be achieved by stripping them of as many qualifications as possible, leaving them free to travel as objects in search of a new interpretation. By contrast, re-contextualisation through evidence codes enables users to evaluate the character of small facts by assessing their provenance. This second step is necessary to qualify the value of small facts as evidence, and thus building an interpretation of their biological significance in a new research setting. In this sense, the process of re-contextualisation is reminiscent of work conducted by curators in a very different setting: museum collections, whose visitors can best form an opinion about the cultural significance of the objects in display when they are given information about the

[19] Schell, this volume, offers a very different example of how de-contextualization helps facts travel.

history of those objects and their creators.[20]

By enabling users to access de-contextualised small facts, databases provide the differential access needed to make small facts travel across contexts. By providing evidence codes, databases facilitate the re-contextualisation of small facts, while at the same time making it possible for them to shift character and significance depending on their new location. This modality of re-use is particularly important in model organism biology, where the same small facts might acquire entirely different interpretations when examined by biologists working on different species and/or dissimilar research cultures. Through their vision of re-contextualisation, curators are attempting to enable biologists to pick up new small facts without necessarily having much in common in terms of their goals and expertise. For instance, researchers investigating the regulatory functions of specific genes are using databases to check what data are available on their gene of interest, how those data were produced and on which species. This enables them to compare what is known about the behavior of the gene across species, without having to become a specialist on each type of organism and experimental procedure involved.

5. The Role of Curators

While illustrating how databases make small facts travel, I already hinted at ways in which databases are challenging existing social structures and conventions governing the dissemination of scientific results.[21] They are giving visibility and usefulness to facts that, contrary to the big facts, used to be discarded by their producers or jealously kept in their laboratories for further research. In most scientific contexts, small facts used to be subject to public scrutiny only when providing crucial evidence for a big fact: they would not be made public before the big fact was published in a journal, and the small facts that did not serve an immediate purpose as evidence in that case would be discarded. Many factors have acted as an incentive to make small facts travel across contexts. Among them are the testimony of the few scientific communities that did exchange small facts at the pre-publication stage, a collaborative strategy which proved to be extremely successful especially in the case of model organism research[22], and the emphasis by funding bodies on the value of small facts as public goods, especially following the controversy on the importance of preserving open access to the results of the Human Genome Project (Sulston and Ferry 2002). Even in the presence of these factors, the opportunity of free exchange provided by databases challenges the competitive ethos prevalent within biological research,

[20] The ways in which small facts acquire character through evidence codes can be usefully compared to other cases of travelling objects in this volume, such as Wylie's 'archaeological facts' and Valeriani's building structures.

[21] Hine (2006) and Leonelli (2009b) discusses many of the challenges posed by databases to scientific social orders.

[22] One example is the vast community of researchers working on the plant Arabidopsis thaliana (Leonelli 2007).

thus making it more and more difficult for researchers to avoid donating their data.

An even more important challenge posed by databases concerns the role of curators as a new type of expert within biology, whose relations to existing experts are not yet clearly established. As I illustrated, curators are aware that their work of de-contextualisation is essentially at the service of the activity of re-contextualisation by the users. De-contextualisation is the means through which small facts are made fit to travel, and re-contextualisation is the ultimate aim of travel. Ideally, therefore, users should be able to re-contextualise small facts in ways that depend solely on their own backgrounds and interests. In practice, however, curators also need to develop a range of skills enabling them to choose labels (both bio-ontologies and evidence codes) that mirror as closely as possible the developments and expectations of the potential users of small facts.

Curators achieve a high level of fit between their work and the work of database users through a variety of interventions. For instance, take the activity of *extracting* small facts from publications and repositories. To do this, curators are forced to single out publications that they consider reliable, updated and representative for specific datasets. When gathering available data on a specific gene (such as the Unknown Flowering Object gene [UFO] in *Arabidopsis thaliana*), curators need to choose one or two publications that best represent data relevant to a given gene product for the purposes of classification. They cannot compile data from each relevant publication, as it would be too time-consuming: even just a keyword search on PubMed on ‘UFO Arabidopsis’ results in 35 journal articles, only one or two of which will be used as reference for an annotation. Thus, curators choose what they see as the most up-to-date and accurate publications on a specific gene product, which as a consequence become ‘representative’ publications for that entity.

Further, once curators settle on a specific publication, they have to assess which small facts therein contained should be extracted and/or how the interpretation given within the paper matches the terms and definitions already contained in the bio-ontology. Does the content of the paper warrant the classification of given data under a new bio-ontology term? Or can the contents of the publication be associated to one or more existing terms? These choices are impossible to regulate through fixed and objective standards. Indeed, bioinformaticians have been trying to automate the process of extraction for several years, without success. The very reasons why the process of extraction requires manual curation are the reasons why it cannot be divorced from subjective judgement: all the choices involved are informed by a curator’s expertise and his or her ability to bridge between the original context of publication and the context of bio-ontology classification.

Performing curation tasks such as extraction presupposes skills honed through specific training and years of experience. Curators are veritable ‘packaging experts', and their combination of skills is crucial to producing both bio-ontologies and evidence codes.[23] Their expertise includes, on the one hand, some familiarity various fields of biological research. This gives them the cross-disciplinary understanding necessary to recognize and respect the diversity characterizing epistemic cultures (and thus, terminologies and methods) within experimental research. On the other hand, curators need to have some experience 'at the bench'. This enhances their awareness of what users need to find through evidence codes (e.g. protocols and search parameters). Curators working on the Gene Ontology, for example, are biologists by training and motivation: their decision to extend their expertise towards computer science and bioinformatics was primarily due to their interest in improving data analysis tools for model organism research as a whole. The curators' hands-on knowledge of experimental work is reflected in the development of bio-ontologies and enhances their intelligibility to experimenters. At the same time, only through a more generalist expertise can curators assess which terms to use, how to define them and how to relate them with each other. Curation is no job for a specialist with a narrow experimental focus; nor is it a job for a computer scientist with no clue of how research at the bench is conducted.

6. User Perspectives

By taking upon themselves the task of choosing the appropriate package for small facts, curators take important decisions on what counts as relevant small facts for any specific research project. Most users are happy to trust them with this role, as they do not want to spare time and energy from their research to deal with choices about packaging. For this same reason, however, users are reluctant to invest effort in understanding the choices made by curators. Users want an efficient service thanks to which they access a database, type a keyword, get the relevant data and go back to their research. By so doing, users often do not understand the extent to which the packaging affects the travel of small facts and the ways in which they will be re-used.

Curators are well aware that their interventions influence where and how small facts will travel. They are willing to recognize that it is their professional duty to serve the user community as best as they can, and they feel both responsible and accountable for their packaging choices – indeed, they are actively seeking scientific recognition for their service as packaging experts (Howe, Rhee et al 2008). They are also aware that it is impossible to conform to the expectations and practices of rapidly changing fields, without being in constant dialog with the relevant user communities. This is also because, aside from one-to-one dialogue and website statistics on which parts of a database are most popular with users, there is currently no reliable way for curators to systematically evaluate how users are using information in the database. Many researchers are not yet used

[23] The expertise of the curators is critical here, as indeed, the expertise of the scientists in Howlett and Velkar's account of technology transfer, this volume.

to citing databases in their final publications – they would rather cite the papers written by the original producers of the data, even if they would have not been able to find those papers and associated data without consulting a database. Curators thus cannot assess which research projects have made successful use of their resources, unless researchers report their achievements to them directly.

Yet, many attempts to elicit feedback fail because of users' disinterest in packaging practices and their inability to understand their complex functioning. The gulf between the activities and expertises of curators and users tends to create a problematic system of division of labour. On the one hand, curators invite users to critically assess their work and complain about what they might perceive as 'bad choices'. On the other hand, users perceive curators' work as a service whose efficiency should be tested and guaranteed by service providers rather than the users. They thus tend to trust curators unconditionally or, in the absence of trust, simply refuse to use the service.

The tensions between database users and curators are exemplified by a recent attempt to package and re-use small facts about leaves. AGRON-OMICS is a European project sponsored by the Sixth Framework programme and bringing together plant scientists from a variety of laboratories and disciplines, including molecular, cellular and developmental biology. Its goal is to secure an integrated understanding of leaf development, by gathering and analysing data extracted from the model organism *Arabidopsis thaliana*. A crucial component of this project is precisely the search for efficient tools to circulate small facts among members of the group and to the research community at large. The question of labelling was uppermost in the minds of the group coordinators from the outset in 2006. What categories could be used to circulate data gathered by researchers so steeped in their own local terminologies and practice?

The very first meeting of the project, a two-day workshop titled "Ontologies, Standards and Best Practice", was devoted to tackling this question.[24] Participants included the main scientific contributors to AGRON-OMICS and the curators of the databases that were most likely to be of use, such as Geneinvestigator, the Arabidopsis Reactome, the Gene Ontology and the Plant Ontology. Curators did most of the talking, both through presentations explaining what their tools could do and through hands-on workshops teaching researchers to use them. Most questions raised concerned systems for tracking the relevance and reliability of small facts; users and curators certainly agreed on the importance of keeping the focus on these two factors. Overall, the workshop was very successful in alerting researchers to the importance of finding good packages to make their facts travel. Remarkably however, this lesson came with an increased awareness of the difficulties plaguing these efforts, and particularly of the problems associated with labelling small facts for re-use.

[24] The workshop, which I attended, took place at the Department of Plant Systems Biology, VIB-UGent (Gent, May 21-23, 2007).

Many of the scientists attending displayed mistrust for the work of curators, which they saw as far removed from actual biological research. The very need to de-contextualise facts was seen as potentially problematic, despite evidence for the necessity of this process to make facts travel. There were complaints that curators, in their tight collaboration with computer scientists, tended to favour a polished labelling system over one that would actually help experimenters; it was also remarked that the synonyms system devised by curators to accommodate terminological pluralism only works if curators are aware of all existing synonyms for a given label. Further, some researchers were dazzled by the multitude of tools available for labelling (well over 20 were mentioned at the meeting, most of which researchers were not yet acquainted with). While some labels, such as the Gene Ontology, are fairly well-established across a number of databases, there are many cases of databases developing their own labelling systems without regard for the ones already in place. This leads to a proliferation of labels which is confusing to most users, who feel they are wasting time in learning to use all those systems and in assessing each label's merits relatively to others. Although some scientists appreciated the idea of being able to choose among different labelling tools, this was often associated with an interest in developing those tools themselves.

Dialogue between users and curators over these difficulties resulted in both sides increasing their understanding of labelling processes. Curators walked away with a better idea of the needs and expectations of AGRON-OMICS researchers. Users however retained a degree of scepticism in curators' work. Indeed, precisely as they were learning to appreciate the scope and implications of curators' work, AGRONOMICS scientists saw the importance of selecting appropriate labels for their facts, as well as the power that this brings over the eventual re-use of those same facts. They therefore resolved to take over some of that work, to ensure that the labels used to package facts be perfectly suited to their research needs. One of the action points agreed upon at the end of the meeting was the creation of two new bio-ontologies: one for Arabidopsis phenotypes and one for Arabidopsis genotypes. The main rationale for this effort was the perceived absence of suitable labels dealing with these biological entities. Also, developing their own labels would ensure that scientists take over the packaging – and thus the modalities for future re-use – of small facts of particular importance to their project.

7. Regulating the Packaging Process

The successful re-use of small facts requires a highly dynamic system of labels. Curators have the crucial function of mediating between the needs of local research contexts and the need to devise standards that can be used by all.
In other words, curators are not simply responsible for making small facts travel: they are responsible for making small facts travel *well*, which involves communicating with users to make sure that facts are indeed being re-used.

In the case of AGRON-OMICS, the potential tensions between curators and users were resolved by making these two figures overlap. It is not clear, however, that this is a good solution. In the absence of a generalist curator aiming to serve the whole biological community, the labels used for packaging might end up serving the needs of the AGRON-OMICS group over and above the needs of other scientists, thus hampering the successful re-use of those same data in other quarters. Further, as I already mentioned, not all scientists are willing to invest time and effort towards the creation of good packages for small facts. Part of AGRON-OMICS funding is explicitly directed to the study and testing of packaging tools for small facts, which means that they can employ people to work on bioinformatics and they have resources for developing and maintaining communication with curators at the international level (thus preventing the danger of narrowing their vision to their own project). The same is not true of other projects, especially smaller projects with more specific goals.

A more general solution could be to enforce some mechanisms of communication between curators and users, so that curators receive frequent feedback from the widest range of users, thus ensuring that their packaging strategies are indeed serving the needs of users as they evolve through time. In other words, packaging – and particularly the de-contextualisation processes for which curators are responsible – is in need of external regulation. An example of such regulatory mechanisms is the requirement to submit data to databases in appropriate formats when publishing a paper. This has recently been implemented by Plant Physiology, a major journal in plant science, in collaboration with The Arabidopsis Information Resource (TAIR), the main database for Arabidopsis research. Researchers wishing to submit a paper to Plant Physiology are required to submit all the data created during their project to TAIR. This forces them to become acquainted with the labelling system adopted by TAIR (which includes both evidence codes and the Gene Ontology). The experience might encourage direct involvement by experimenters in the development and use of bio-ontologies.[25]

Yet another effective regulatory measure is the introduction of institutions that are responsible for implementing the packaging and setting standards for it. That the rise of regulatory institutions would support the development and maintenance of 'good packaging practice' will not come as a surprise to the readers of this volume, as the importance of such structures is emphasised by many other analyses of how facts are packaged for travel (e.g. the case of technology transfer in Northern India examined by Howlett and Velkar). In the case of bioinformatics, many prominent packaging efforts have been centralised in few loci, such as the European Bioinformatic Institute and the Gene Ontology Consortium in Hinxton, Cambridge.[26] This institutionalisation of packaging

[25] See Ort & Grennan (2008).

[26] Leonelli (2009b) discusses the institutional history and status of 'labelling centres' such as the Gene Ontology Consortium.

prevents the proliferation of labels and therefore enhances their power to cross contexts. It also helps to train packaging experts who can teach users how to deal with labels and vehicles; and it enables the creation of feedback mechanisms through which users can provide constructive critiques to curators, thus bettering their packaging strategies and the resulting re-contextualisation processes. A downside of institutionalisation is the centralisation of power on what counts as good packaging. This involves a potential loss of diversity in packaging strategies, as it gives particularly prestigious and well-funded groups the opportunity to shape the choice of labels according to their own preferences and interests. Another problem is that existing centres, despite the support they receive by funding bodies and user communities, are struggling to cope with the immense amounts of small facts to be curated.[27]

And yet, despite the efforts of so many highly qualified minds, it is still not clear whether this packaging system will end up working as desired. This is not because the system is not well-designed and maintained, but rather because we do not yet know whether the tensions between curators and users will be resolved in a way that is satisfactory to both.

Conclusions

My analysis has focused on the strategies devised by scientists to cope with the need to access and re-use the billions of small facts produced by contemporary research, and particularly high-throughput technologies. The case of bioinformatics illustrates the complexity of making facts travel even within the supposedly narrow boundaries of the scientific world. Travel across research contexts involves crossing large distances, both in geographic and in epistemic terms. This requires a lot of effort especially in the case of small facts, which are produced in extremely large numbers, travel only in groups and with the help of specific travelling companions (e.g. labels), and require purpose-made vehicles to move around.

What is most interesting for our purposes is precisely the purpose-made character of this new apparatus vis-à-vis its efficiency as a packaging tool. Given the care and thoughtfulness given to creating and perfecting packaging strategies using the latest technologies, the case of database curation constitutes an ideal case of travelling facts. This is a case where a whole system of labels, communication strategies, and digital vehicles has been explicitly created to make sure that facts travel well.

[27] Recourse to ‘crowdsourcing’ or ‘wikification’, i.e. user involvement in annotating databases, has been hailed as a solution to these problems (Leonelli 2009b). Yet, given the high level of specialised expertise required to curate small facts, it is difficult to know whether it would work. The degree to which good packaging requires centralisation remains an open question.

Yet the value of the packaging process is ultimately dependent on the efficiency with which curators and users communicate about their respective needs and interests. As I illustrated, scientists have recently had to acknowledge that a whole apparatus of innovative technologies, expertises and institutions was needed to make these facts travel: without curators, apposite labels and databases, it would be very difficult to enact the processes of de-contextualisation and re-contextualisation needed to make small facts travel.

Acknowledgments

This work was supported by the ESRC/Leverhulme Trust project, “The Nature of Evidence: How Well Do ‘Facts’ Travel?” (grant F/07004/Z) at the Department of Economic History, LSE. Special thanks go to the scientists who took time to be interviewed and provide feedback in the course of this research. I am particularly grateful to the co-ordinators of AGRON-OMICS, Fabio Fiorani and Pierre Hilson; Klaus Mayer and colleagues at the Munich Information Centre for Protein Sequences; Sean May and colleagues at the Nottingham Arabidopsis Stock Centre; Michael Ashburner, Midori Harris and colleagues at the Gene Ontology; Sue Rhee and colleagues at The Arabidopsis Information Resource; and participants to the BBSRC workshop ‘Data Sharing in the Biomedical Sciences’ (Edinburgh, June 2008).

Bibliography

Ankeny, R. (2007) Wormy logic: Model organisms as case-based reasoning. In Creager, A. N.H., Lunbeck, E. and Wise, N. eds. **Science Without Laws: Model Systems, Cases, Exemplary Narratives.** Chapel Hill, NC: Duke University Press, pp. 46-58.

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

Augen, J. (2005) **Bioinformatics in the Post-genomic Era. Genome, Transcriptome, Proteome and Information-Based Medicine**. Addison-Wesley.

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

Bogen, J. and Woodward, J. (1988) Saving the phenomena. **The Philosophical Review** 97, 3, pp. 303-352.

Bowker, G.C. (2000) Biodiversity Datadiversity. **Social Studies of Science** 30, 5, pp. 643-683.

Duhem, P. (1974 [1914]) **The Aim and Structure of Physical Theory**. New York: Atheneum.

Dupré, J. and Barnes, S.B. (2008) **Genomes and What to Make of Them**. John Wiley.

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

Hacking, I. (1992) The self-vindication of the laboratory sciences. In Pickering, A. ed. **Science as Practice and Culture**. The University of Chicago Press, pp. 29-64.

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

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

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

Huang et al. (2005) Arabidopsis VILLIN1 Generates Actin Filament Cables That Are Resistant to Depolymerization. **Plant Cell** 17, pp. 486-501.

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

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

Krohs, U. and Callebaut, W. (2007) Data without models merging with models without data. In Boogerd, F.C., Bruggeman, F.J., Hofmeyr, H.S. and Westerhoff, H.V. eds. **Systems Biology: Philosophical Foundations**. Elsevier, pp. 181-213.

Leonelli, S. (2007) Arabidopsis, the botanical Drosophila: From mouse-cress to model organism. **Endeavour** 31, 1: 34-38.

Leonelli, S. (2009a) On the locality of data and claims about phenomena. **Philosophy of Science** 76, 5, pp. TBA.

Leonelli, S. (2009b) 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.** London: Routledge, pp. 469-485.

Longino, H. (2002) **The Fate of Knowledge**. Princeton, New Jersey: Princeton University Press.

Maddox, B. (2002) **Rosalind Franklin: The Dark Lady of DNA**. Harper Collins.

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

Ort, D.R. and Grennan, A.K. (2008) Plant physiology and TAIR partnership. **Plant Physiology** 146, pp. 1022-1023.

Raj, K. (2007) **Relocating Modern Science: Circulation and the Construction of Knowledge in South Asia and Europe, 1650-1900**. Palgrave Macmillan.

Rheinberger, H. (unpublished) **From Traces to Data, From Data to Facts.**

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

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

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

Terrall, M. (2008) **Following Insects Around: Tools and Techniques of Natural History in Réaumur's World**. Talk delivered at the Sixth Joint Meeting of the BSHS, CSHPS, and HSS. 5 July 2008, Keble College, Oxford.

Zimmerman, A. (2007) Not by metadata alone: The use of diverse forms of knowledge to locate data for reuse. **International Journal of Digital Libraries** 7, pp. 5-16.

Online resources

AGRON-OMICS website http://www.agron-omics.eu/

Gene Ontology website http://www.geneontology.org/