Working Papers on The Nature of Evidence:
How Well Do 'Facts' Travel?
No. 25/08

Circulating Evidence Across Research Contexts: The Locality of Data and Claims in Model Organism Research

Sabina Leonelli

Department of Economic History
London School of Economics

March 2008

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

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

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

Series Editor:

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

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

Circulating Evidence Across Research Contexts:
The Locality of Data and Claims in Model Organism Research[1]

Sabina Leonelli

Abstract

In everyday scientific practice, facts come in two sizes: small facts (data acquired by researchers through experimentation or field work), and big facts (claims about phenomena for which data function as evidence). This paper explores the processes through which small and big facts are circulated and used across research contexts in model organism biology. This leads me to challenge the standard philosophical characterisation of data as embedded in the context in which they are produced (and hence "local") and of claims about phenomena as retaining their significance beyond that context (hence "non-local"). I argue that the degrees of locality of both small and big facts are not intrinsic to their epistemic status, but rather vary depending on the packaging used to make them travel. As illustrated in the case of bioinformatics, packaging processes include recourse to appropriate labels, vehicles and expertise. Facts about organisms travel well when they are temporarily liberated from information about their context of production, thus becoming non-local entities that can be recruited across new contexts. At the same time, information about provenance needs to be included in the packaging of facts, so as to enable prospective users to assess their reliability.

Introduction

Data[2] are the smallest, yet the most stubborn of scientific facts. They constitute the empirical backbone of scientific research: once they are

[1] **Acknowledgements**: Mary Morgan provided insightful suggestions on several drafts of this paper, which has greatly improved as a result. Warm thanks also to the "facts" group; to audiences in Twente and Madrid; to Chris Jarvis and his colleagues at the MRC Centre in London; and to the TAIR team, particularly Sue Rhee, for pictures and information. This research was conducted as part of the project "The Nature of Evidence: How Well Do "Facts" Travel?," funded by the Leverhulme Trust (grant number F/07004/Z) and the ESRC at the Department of Economic History, LSE.

[2] I here follow Ian Hacking's broad definition of data as any "marks" produced by a "data generator': "uninterpreted inscriptions, graphs recording variation over time, photographs, tables, displays" (Hacking 1992, 48). Biological data, for instance, include various types of marks, among which material objects (e.g. stains on an embryo resulting from an in situ hybridisation experiment), dots on a slide (e.g. micro arrays) and strings of letters (e.g. DNA sequences). Especially within genomics, increasing quantities of data are now produced in digital formats (i.e. XML files), to facilitate their dissemination through digital means (Hilgartner 2004).

adopted as reliable evidence for a given claim, data are generally trusted and used without being altered or questioned. But what is the relation between data and the claims for which they are taken as evidence? Can data be circulated independently of those claims, so as to be used in research contexts other than the one in which they have been produced? And in which ways and with what consequences does this happen, if at all? This paper tackles these questions by focusing on how biological data travel to research contexts other than the one in which they have been produced. I argue that data need to be appropriately packaged to be circulated and used as evidence for new claims; and that studying the process of packaging helps to understand the relation between data, claims, and the local contexts in which they are produced. My analysis leads me to challenge some of the conclusions drawn by Bogen and Woodward (B&W) on the evidential role of data and claims about phenomena. B&W characterise data as unavoidably embedded in a specific experimental context, a condition which they contrast with the mobility enjoyed by claims about phenomena, whose features and interpretation are alleged not to depend on the setting in which they are formulated. This view does not account for cases where data travel beyond their original experimental context and are adopted as evidence for new claims, nor for the extent to which the travelling of claims about phenomena depends on shared understanding across epistemic cultures.

In what follows, I argue that the capacity of data to travel constitutes a defining feature of contemporary biological science. This capacity is not intrinsic to data themselves, but needs to be enforced by scientists. Many efforts and resources are invested in creating tools and procedures through which data can be retrieved and used beyond the context in which they have been produced. The extent of these efforts is such that they often result in the birth of new types of expertises and infrastructures devoted explicitly to making data travel. My analysis focuses on one such case, namely: the use of digital databases to

gather, organise and distribute the heterogeneous mass of available data about model organisms. In the first half of the paper I introduce databases as means of overcoming the challenge presented to biologists by the accumulation of data on model organisms. I reconstruct the phases through which data are made to travel through databases, and the consequences of such travelling for biological research. In the second half of the paper (sections 2 and 3), I use this case to discuss B&W's claims. I examine three main components of the packaging through which databases allocate evidential value to data: labels, technological infrastructure and expert intervention. My analysis leads me to conclude that both data and claims about phenomena may be more or less embedded in a local research context: their degree of locality depends on the packaging strategies through which they are circulated.

1. Making Facts About Organisms Travel

Biology has yielded immense amounts of data in the last three decades. This is especially due to genome sequencing projects, which are yielding billions of data points about the DNA sequence of various organisms. Researchers in all areas of biology are busy exploring the functional significance of those structural data. This leads to the accumulation of even more data of different types, including data about gene expression, genes' position on the chromosomes and their mobility through time, morphological effects correlated to "knocking out" specific genes, and so forth (Kroch and Callebaut 2007). These results are obtained through experimentation on a small set of species, including fruit-flies (Drosophila melanogaster), worms (C. elegans), mouse cress (Arabidopsis thaliana) and mice (Mus Musculus), whose features are particularly tractable through available laboratory techniques. These are called "model

organisms," because it is assumed that results obtained on them will be applicable to other species with broadly similar features.[3]

Researchers are aware that assuming model organisms to be representative for other species is problematic. Cross-species transfers of knowledge are unavoidably a shot in the dark, as researchers cannot know the extent to which species differ from each other unless they perform accurate comparative studies. Indeed, reliance on cross-species inference is a pragmatic choice. Focusing research efforts on a few species enables researchers to integrate data about several aspects of their biology, thus obtaining a better understanding of organisms as complex wholes. Despite the dubious representational value of model organisms, the majority of biologists agree that cooperation towards the study of several aspects of one organism is a good strategy to advance knowledge, as results acquired on that one organism can be a starting point for the study of other species. The circulation of data across research contexts is therefore considered a priority in model organism research: cooperation can only spring from an efficient sharing of results.

The quest for efficient means to share data has recently become a lively research area in its own right, usually referred to as bioinformatics. One of the main objectives in bioinformatics is to exploit new technologies to construct digital databases that are freely and easily available for consultation (Rhee et al 2006).[4] Aside from the technical problems of building resources that would process, store and visualise huge amounts of data, bioinformaticians have to confront two main issues. One is the fragmentation of model organism biology into epistemic communities with their own local culture – that is, their own expertise, traditions, favourite methods, instruments and research goals (Knorr Cetina 1999). Making data accessible to all of these communities means finding a vocabulary and a format for the data that makes them retrievable by anyone according to her own research interests and

[3] For a detailed analysis of the characteristics of model organisms and the way in which they are used in research, see Ankeny (2007) and Leonelli (2007).
[4] Another strand of bioinformatics, which I will not discuss here, has to do with the mathematical modeling and analysis of populations.

background. The second issue concerns the characteristics of data coming from disparate sources and produced as evidence for various different claims, which make it difficult to assess the evidential scope of the data that become available (what are they evidence for?) as well as their reliability (were they produced by competent scientists through adequate experimental means?).

I examine bioinformaticians' responses to these demands by looking at three phases of data travel: (i) the disclosure of data by researchers who produce them; (ii) the circulation of data through databases, as arranged by database curators; and (iii) the retrieval of data by researchers wishing to use them for their own purposes. I illustrate my analysis of these phases through a specific case, that is the publication, circulation and retrieval of data concerning the expression of the UFO ("unusual flower organs") gene in the flowering of the model plant Arabidopsis thaliana.

i. *Disclosure by data producers*

The vast majority of experimenters disclose their results through publication in a refereed journal. Publications usually include a description of the methods and instruments used, a selected sample of the data thus obtained and an argument for how those data support the claim that researchers are trying to make. Data are selected on the basis of their value as evidence for the researchers' claim; their "aesthetic" qualities (e.g. clarity, legibility, exactness); and their adherence to the standards set in the relevant field.[5] Because of these strict selection criteria, a large amount of data produced in the course of experiments is discarded without being circulated to the wider community. Also, published data are only available to researchers who are interested in the claim made in the paper. Especially given the vast amount of publications currently issued in the biological sciences, there is little chance that a

[5] Each way to disclose data requires data producers to "polish" the data and standardise them according to the requirements imposed by the publication that they send data to.

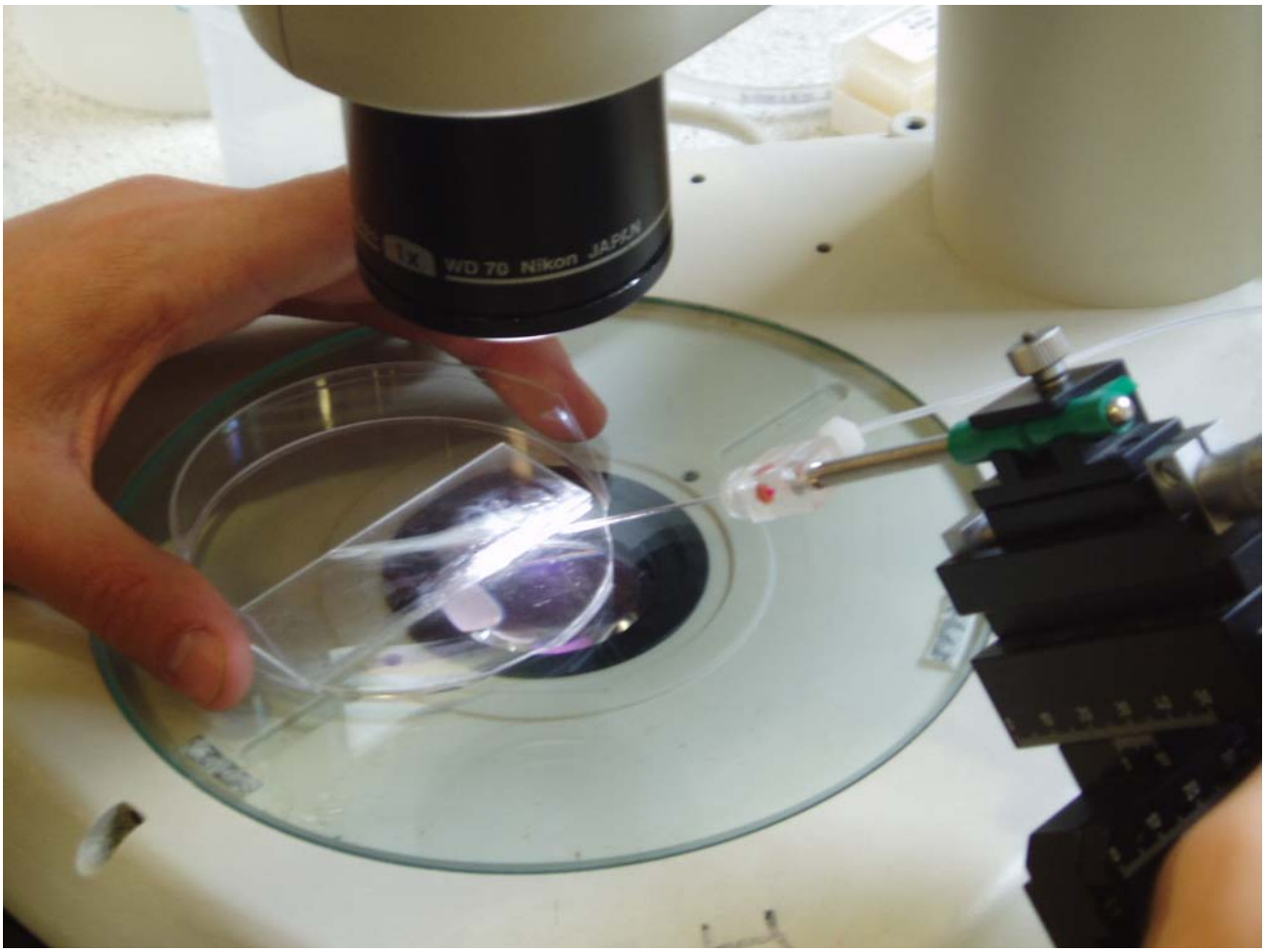
researchers working in a different area or on a different claim will read the paper, see those data and thus be in a position to evaluate their relevance to their own projects. As a consequence, there is little chance that those data will ever be employed as evidence for a claim other than the one that they were produced to substantiate. Disclosure through patenting, another popular means for circulating data in biology, has similar characteristics: data are reported in patents purely to add plausibility to the claims about phenomena (most often about how to intervene on organisms) made therein.[6]

Take the example of data obtained on the expression of the UFO gene in Arabidopsis, which plays a significant role in the development of the stamens and pistils of the plant. In situ hybridisation is a standard experimental technique used to establish whether a given gene regulates the development of specific morphological traits. It consists of hybridising a specific DNA molecule with a RNA probe, so that the mRNA produced by the DNA molecule will become stained and clearly visible under the microscope; gene expression, that is the mRNA signals sent by DNA outside of the cell nucleus, is thus visualised. In a plant like Arabidopsis, this technique is usually applied to the whole embryo: researchers grow the embryo to a desired stage, inject it with the probe (figure 1) and, via a series of procedures designed to help the probe hybridise with the target DNA, are able to observe the parts of the embryo in which the chosen gene is expressed at different stages of development (figure 2).

A group of Canadian researchers led by Alon Samach performed in situ experiments to visualise the expression of UFO in shoot apical meristem growth (that is, the development of the pluripotent cells located on the top of a growing plant into the specialised cells that make up flowering organs). The aim of their research was to prove the regulatory role played by UFO in the flowering stage of Arabidopsis development.

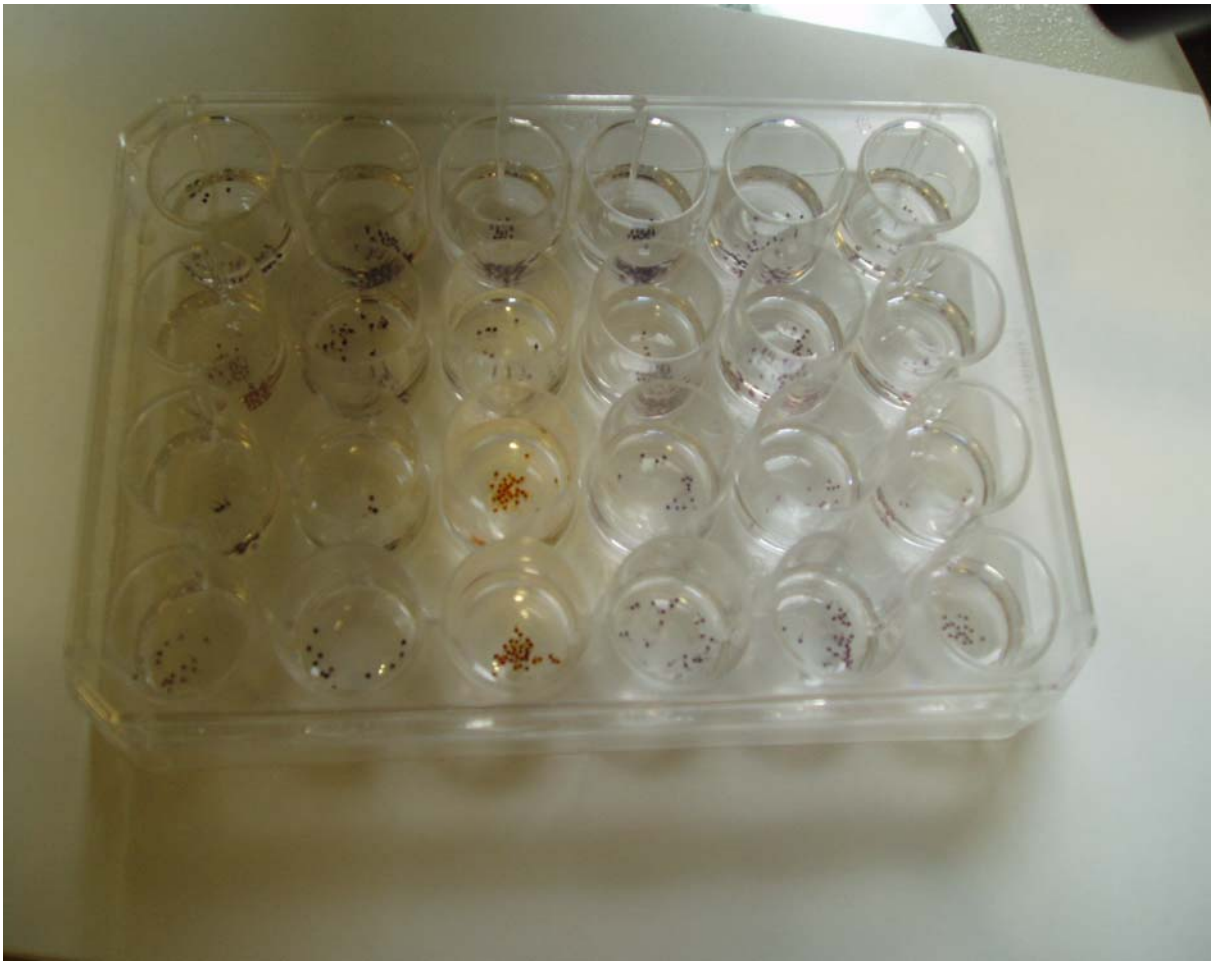
[6] Hilgartner (1995) coins the term “communication regimes” to analyse the complex sociotechnical systems supporting different forms of data disclosure, such as scientific journals, patents and databases.

Figure 1. Researcher injecting a probe into embryos. Picture taken by the author at the MRC Centre for Developmental Neurobiology, Kings College London.



The experiments performed by this group resulted in several hundreds stained tissues, most of which indeed display a strong expression of UFO in the late stages of shoot apical meristems development. This means that the group could publish a paper detailing their claims and the evidence on which they are based. The researchers therefore selected a few “good-looking” embryos (i.e. embryos where the staining effect is clearly legible) whose picture could be published as evidence for their claims. These pictures, as reported in figure 3, were published in 1999 in the prestigious *The Plant Journal*. The title of this publication “The UNUSUAL FLOWER ORGANS gene of Arabidopsis thaliana is an F-box protein required for normal patterning and growth in the floral meristem” makes the scope of the claims very clear. Only researchers interested in the development of the floral meristem will be

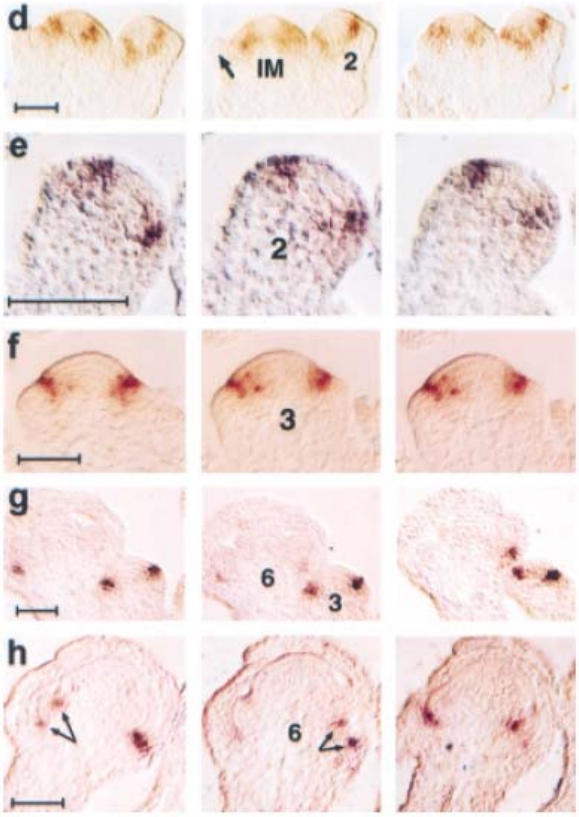
Figure 2: Each cylinder contains hybridised embryos at various stages of development. Picture taken by the author at the MRC Centre for Developmental Neurobiology, Kings College London.



interested in reading the paper and assessing the value of the data therein reported.

So far, this story confirms B&W's assessment of data as embedded in a specific research programme and unable to move across research contexts. If disclosure is left solely to publications and patents, data will travel as the empirical basis of specific claims about phenomena (e.g. "the UFO gene is expressed in apical shoot meristem development "). Data such as in figure 3 will rarely be examined independently of those claims, as they have been carefully selected to provide evidence, rather than to document the full spectrum of data obtained by researchers in their interaction with phenomena. In other words, publications aim to make big facts travel, and small facts such as data are treated as an indispensable part of the evidential baggage that

Figure 3: Photographs of top of developing embryo, showing stains from UFO expression, as published in (Samach et al, 1999).



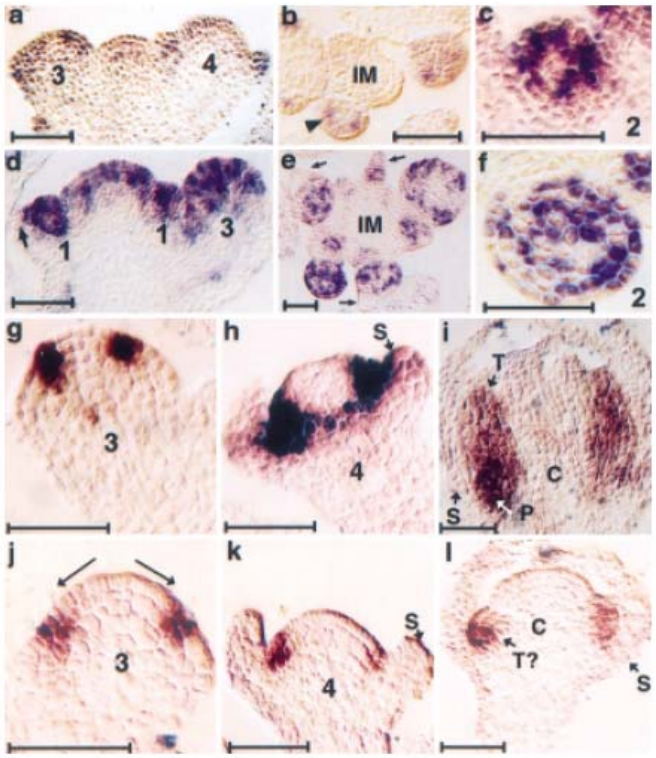


Figure 2. *In situ* hybridization of *PROLIFERA* and *AP3* probes to wild-type and *ufo-1* meristems.

claims need to bring with them. Both biologists and their funding agencies are however unhappy with this situation, since making data travel beyond their context of production allows their potential value to be exploited as evidence for other, possibly new claims about the same phenomena. Maximising the use made of each dataset means maximising the investments made on experimental research and avoiding the danger that scientists waste time and resources in repeating each other's experiments simply because they do not know that the sought-for datasets already exist.

One radical move to "liberate" data from their local context of production has been the construction of public repositories that are available online and collect all data produced in a digital format (e.g. in shot-gun sequencing, micro arrays or in situ experiments), regardless of which of them are used as evidence in publications. GenBank, one of the most famous such repositories, collects data gathered through

sequencing projects across a variety of model organisms.[7] Yet, even as it solves the problem of withholding data from public access, the disclosure of data through repositories presents another challenge: data are thrown into the repository with a minimal amount of standard formatting and classificatory criteria, thus making it difficult for users to locate the data that they may find interesting. The repositories are not the most efficient vehicle for travelling data: they do not provide means for users to retrieve and compare data quickly and efficiently according to their own research interests.

ii. *Circulation by data curators*

Community-based databases, that is databases devoted to collecting data about specific model organisms, are an attempt to solve the challenge presented by repositories without falling back to the inefficient disclosure format provided by publications (Rhee et al 2006). The curators responsible for the construction and maintenance of databases ground their work on one crucial insight. This is that what biologists consulting a database wish to see is first of all the actual "marks," to put it with Hacking, obtained through measurements and observations of organisms: the sequence of amino acids, the dots of micro array experiments, the photograph of an embryo taken after in situ hybridisation. These marks constitute unique documents about a specific set of phenomena. As I noted in the case of in situ experiments, they are produced under heavy constraints posed both by the experimental setting and by the nature of the entities and processes under scrutiny. These material and social constraints severely limit the representational value of the data: they certainly cannot be taken to document reality "as it is," regardless of human intervention, but are rather empirical traces gathered through highly situated human interaction with specific objects. This said, it is important to note that researchers with differing interests and expertises might – and often do – see the same data in different

[7] For a history of the construction of GenBank, see Strasser (2006) and Garcia-Sancho (2007). See also Hilgartner (1995) on its function as a communication regime.

ways and interpret them as evidence for a variety of explanations of the phenomena at hand. While Samach and his collaborators saw the pictures in figure 3 as telling something about the role of genes in development, researchers interested in physiology might see those same pictures as evidence for claims about the chronology of flowering processes; cell biologists might use them to explain of how pluripotent cells diversify; and biochemists might be interested in the variations of staining patterns across the pictures, which tells something about the susceptibility of tissues to the probe.

Curators have realised that researchers are unlikely to see data in the light of their own research interests, unless the data are presented independently of the claims that they were originally produced to validate. This does not mean that data can be used without accessing information about their sources and provenance; that is, the methods, instruments, protocols and setting through which they have been produced. As I shall illustrate, such information is key to evaluating the reliability of data: yet, it is not necessarily needed at the stage of research when biologists are trying to find out what work has already been done that could potentially inform their research goals. Acting on this insight, curators endeavour to insert both data and information about their production in their databases, but in a way that allows researchers to retrieve them separately if they so wish.

The idea that data can be separated from information about their provenance might seem straightforward. Most facts travelling from one context to another do not bring with them all the details about how they were originally fabricated. When spreading a rumour, for instance, we are interested in its content rather than its source (as in “they might be getting married”). The source only becomes important when adding credibility to the claim (as in: “the major said that they are planning to get married”). Even non-scientific cases show that the reliability of facts that travel widely is very hard to assess. This is precisely because to evaluate the quality of a claim we need to know how the claim originated. Knowing

whether to trust a rumour depends heavily on whether we know where the rumour comes from and why it was spread in the first place; a politician wishing to assess whether to believe a claim about, say, climate change, needs to go back to the available research on this topic and reconstruct the reasoning and methods used by scientists to validate the claim.

Thus, on the one hand, facts travel well when stripped of everything but their content and means of expression; on the other hand, the reliability of facts cannot be assessed except by reference to the way in which facts are produced. Curators are well aware of this seeming paradox. Their main challenge is to find ways to make data travel, without however depriving researchers who "receive" those data from the opportunity to assess their reliability for themselves, according to their own criteria. The realisation of this goal is everything but straightforward, and there are various schools of thought in bio-informatics concerning strategies to confront the challenge.[8] Here I will focus on a particularly popular approach among model organism databases, which can be characterised as involving three main activities: "data mining," "data annotation" and allocation of "evidence codes." As I go along, I shall illustrate how these processes work by describing how the UFO data published by Samach et al were added to The Arabidopsis Information Resource [TAIR], which is the best developed community database for genomic data on Arabidopsis thaliana.[9]

Data are first of all "mined" from all available sources, which usually include publications, repositories and direct communication from experimenters. Curators select data that they deem to be of high quality as well as representative for a given biological object. The process of mining starts not from the actual sources, but from the curators' list of

[8] See Leonelli (forthcoming B).
[9] TAIR personnel describes the goals of the database as follows: "to develop user-friendly tools that permit an individual working outside this model species to formulate a query based on their organism of interest, have that query directed to the relevant knowledge for the plant models, and present the information about the models in a way that can be understood by the plant biology community at large" (Rhee et al 2003).

items to be annotated, which can be divided into two main categories: basic objects in Arabidopsis biology, such as genes, markers, polymorphisms, chromosomes and so forth; and basic experimental techniques, whose results need to be centrally stored and organised so that they do not remain solely in the domain of the place where each experiment is performed (examples of this latter category are micro array experiments, sequencing techniques and in situ experiments). Data mining from publications does not present problems in terms of quality assessment, as curators trust that referees for the journal in question would have done the job for them. In the case of the 1999 paper by Samach et al, TAIR curators selected all information therein contained that is relevant to the UFO gene (see figure 4). Data mining from repositories is trickier, as curators can often choose between different datasets – for instance, sequence data gathered by different labs on the same part of the chromosome.

Figure 4: Data annotated in the databases on the basis of the paper by Samach et al.

TAIR Annotation Search [Help]

Your query for annotations based on the publication The UNUSUAL FLORAL ORGANS gene of Arabidopsis thaliana is an F-box protein required for normal patterning and growth in the floral meristem. resulted in 4 records.

Displaying 1 - 4 of 4 records on page 1 of 1 pages.

Gene	Relationship Type	Keyword Category	Keyword	*Evidence Code*: Evidence Description: Evidence With: Reference	Annotated By/ Date Last Modified
UFO	required for	biological process	meristem organization	*inferred from mutant phenotype*: analysis of visible trait: none: Samach, et al. (1999)	The Arabidopsis Information Resource/ 2003-03-27
AT1G30950.1	located in	cellular component	ubiquitin ligase complex	*inferred from physical interaction*: none: none: Samach, et al. (1999)	The Institute for Genomic Research/ 2003-04-17
ASK2	functions in	molecular function	protein binding	*inferred from physical interaction*: yeast two-hybrid assay: UFO: Samach, et al. (1999)	The Arabidopsis Information Resource/ 2006-05-10
UFO	expressed in	plant structure	floral meristem	*inferred from direct assay*: in situ hybridization: none: Samach, et al. (1999)	The Arabidopsis Information Resource/ 2003-03-27

Once they are mined, data are labelled with what curators call a "unique identifier." This is a standard label given to the item with which the dataset has been associated during data mining, so that there can be no confusion as to which item datasets are supposed to document (see figure 5). Importantly, the unique identifier is machine readable, which makes it possible for data to be computed and analysed through machines (when possible, in automated ways). The labelling of datasets with unique identifiers marks the beginning of the process of annotation, which involves the packaging of datasets to be retrieved through the database. This is a process of classification and of standardisation at the same time: data are modified to fit the machine-readable formats adopted within the database, and they are classified according to the conceptual categories chosen by curators to order the database and make it intelligible to practicing biologists.[10] The step of classification is the one that interests me the most here, as it implies the association of each dataset with keywords. These keywords, which are part of so-called "bio-ontologies," signal the phenomena to which data are presumed to be potentially associated: that is, the phenomena for which they might be used as evidence. In the case of the UFO data, the chosen keywords are "meristem organisation" and "floral meristem." This implies that the data could be relevant to research performed on any aspect of meristem organisation, above and beyond the narrow claim that Samach et al have published in conjunction with the disclosure of those data.

Performing data mining and annotation does not involve reference to the context of data production. Data are taken to speak for themselves as "marks" that are potentially applicable, in ways to be specified, to the range of phenomena indicated by keywords. The layered structure of databases, which exploits a hierarchical organisation of data allowing scientists to peruse different classes of information by clicking on sections of the screen, enables curators to store as much information

[10] Leonelli (forthcoming A) discusses in detail the epistemic significance of the annotation process, including the advantages of machine-readable identifiers for the classification of data and the importance of keywords used in bio-ontologies.

about how data have originally been produced as can be put on a screen. “Evidence codes” are categories classifying any given set of data according to the method with which it was obtained (e.g. table 1). By clicking on these codes, the database user can access information about the methods and protocols used to obtain data; the model organism, often down to the specific ecotype, used in the experiment; the instruments and techniques used; the publications or repository in which the data have first appeared; and the names and contact details of the researchers responsible for that bit of research, who can therefore be contacted directly for any question concerning information not directly reported in the database. In short, evidence codes allow database users to access information detailing the methods, instruments, goals, protocols and people who made up the context of data production.

Figure 5: The annotation of gene UFO in TAIR. The unique identifier is what is here signalled as “locus,” while the bio-ontology terms as referred here as “keyword terms.”

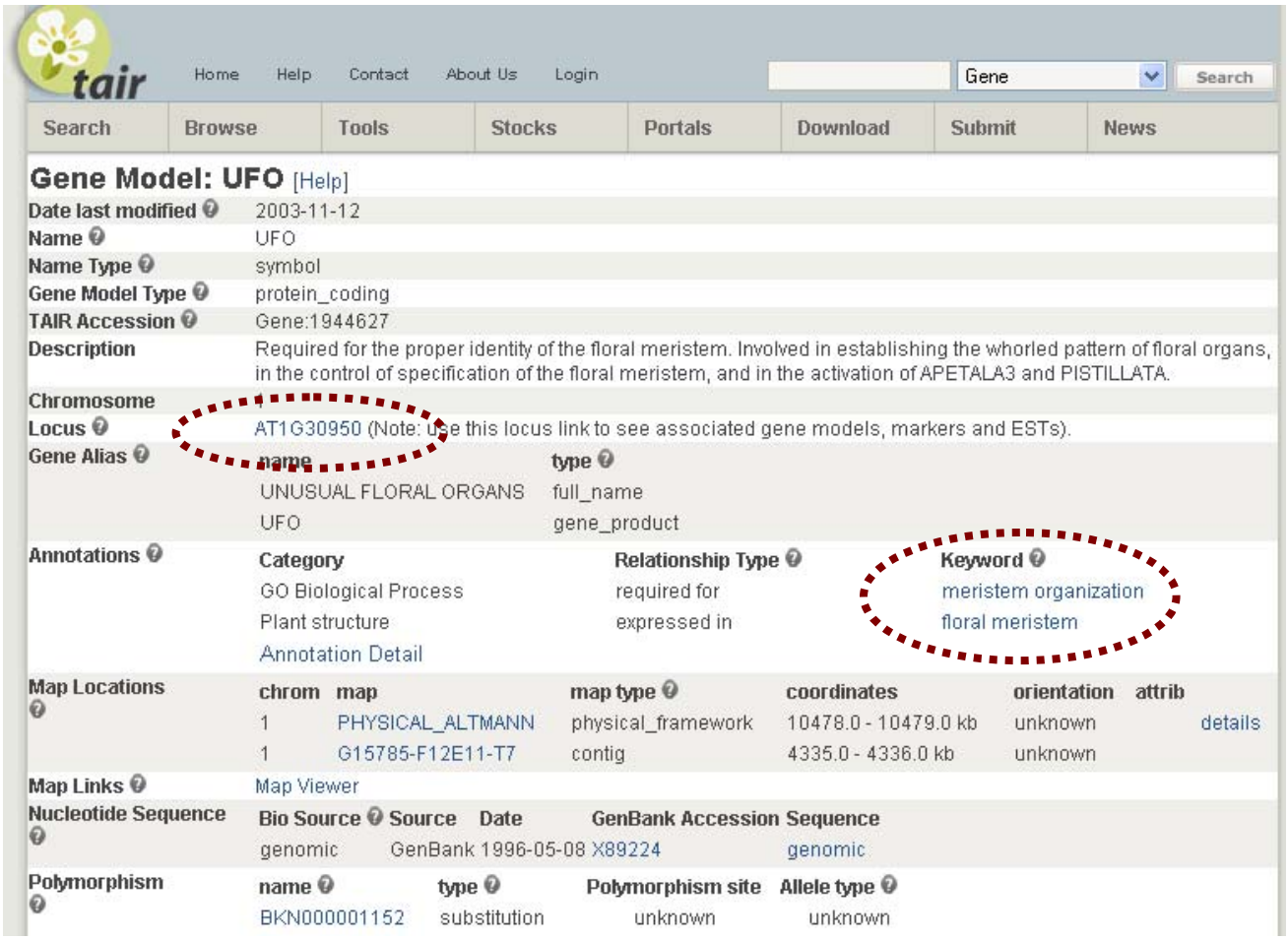


Table 1: Evidence Codes used in TAIR

Experimental evidence codes

- Inferred from Mutant Phenotype
- Inferred from Direct Assay
- Inferred from Genetic Interaction
- Inferred from Physical Interaction
- Inferred from Expression Pattern

Computational analysis

IEA - Inferred from Electronic Annotation
RCA - Reviewed Computational Analysis
ISS - Inferred from Sequence Similarity

Author statement

TAS - Traceable Author Statement
NAS - Non-traceable Author Statement

Curatorial statement

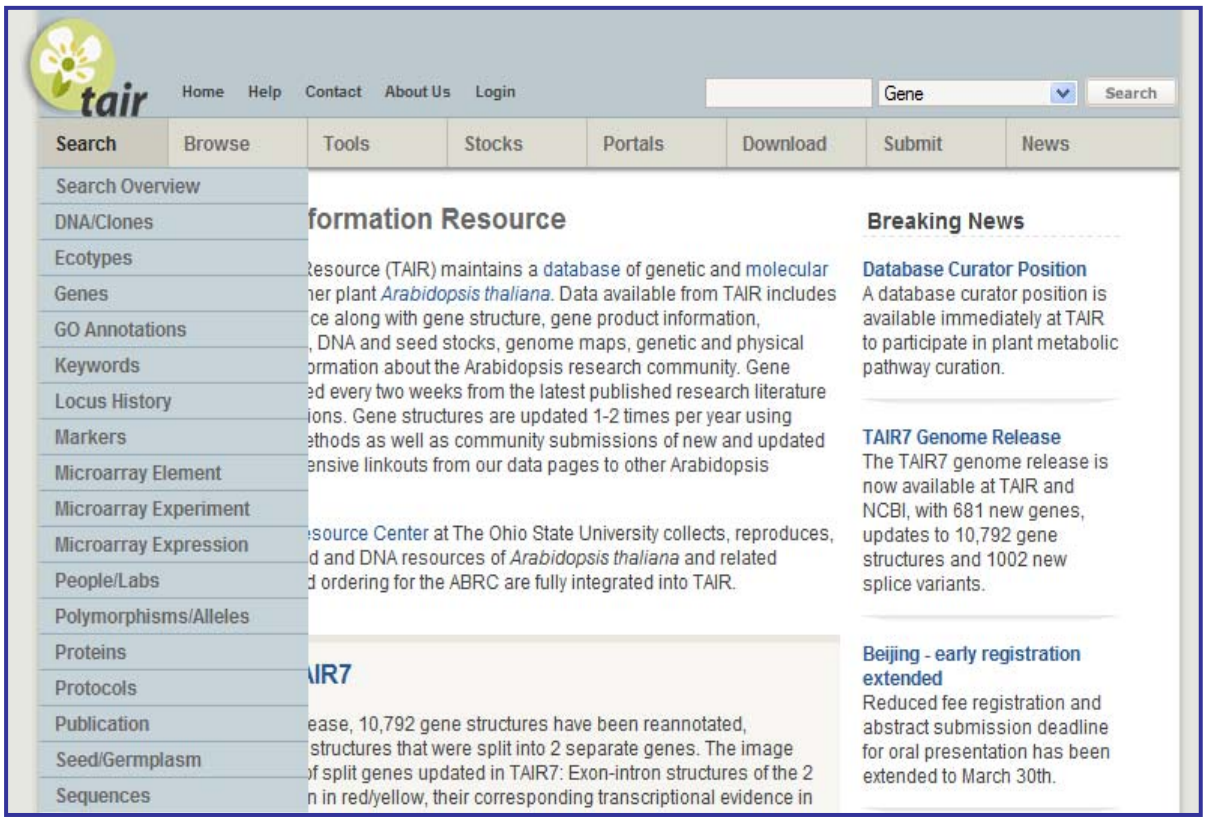
IC - Inferred by Curator
ND - No biological Data available

iii. *Retrieval by data users*

Data packaged through the processes of mining and annotating are ready to be retrieved by the database users. Retrieval happens through so-called “search tools.” TAIR, for instance, contains around 30 tools, each of which allows the submission of queries about one type of items: e.g. genes, metabolic cycles, ecotypes, proteins, microarray experiments, markers, polymorphisms and so forth (see figure 6). The display of search results happens through a series of digital models

developed by curators, which allow users to visualise data according to the parameters they request.

Figure 6: A list of the principal search tools available on the TAIR homepage, obtained by clicking on “search” in the main options bar.



For the purposes of this paper, what is most remarkable about these search and display tools is their flexibility to the demands and expertise of the researchers using them. Users can look up appropriate keywords in which to phrase their query (thus establishing the range of phenomena of interest, as when looking up “floral meristem”) or specific gene names (e.g. “UFO gene” or, to use its unique identifier in TAIR, “AT1G30950”; see figure 7). Once they have successfully formulated their query, they will get a series of links to all the data associated with either the keyword or the unique identifier. By clicking on one such link, they will narrow down their search to the items that they are looking for. Also, they will be able to choose among ways to display those results, so as to fit the data that they receive as much as possible to the methods and standards characterising their own research context.

Figure 7: Display of results of query “UFO gene” in TAIR.

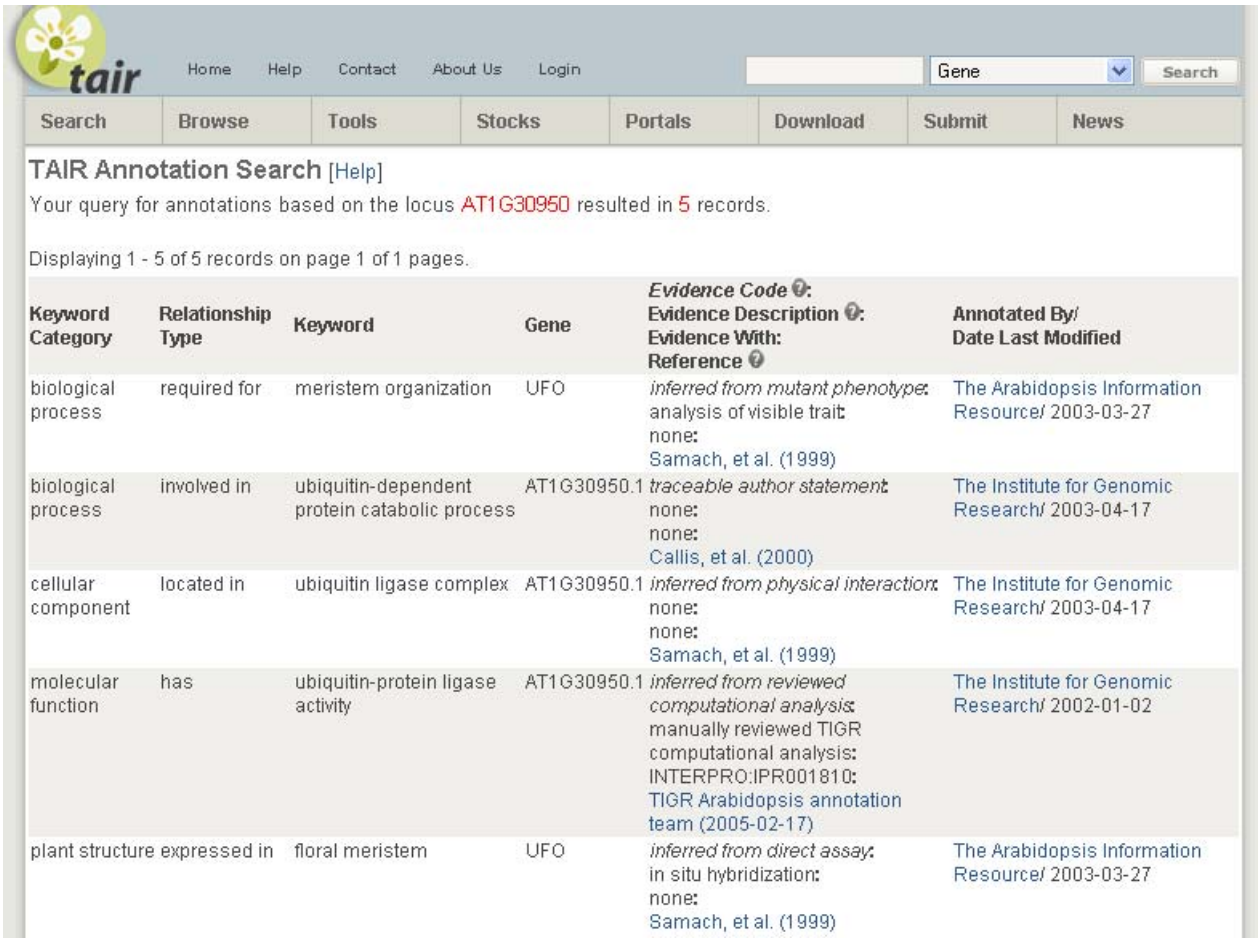
Locus	Gene Model	Description	Other Names	full-length cDNA	Keywords
AT1G30950	AT1G30950.1	Required for the proper identity of the floral meristem. Involved in establishing the whorled pattern of floral organs, in the control of specification of the floral meristem, and in the activation of APETALA3 and PISTILLATA.	F17F8.16	yes	ubiquitin ligase complex, ubiquitin-protein ligase activity, ubiquitin-dependent protein catabolic process
	UFO	Required for the proper identity of the floral meristem. Involved in establishing the whorled pattern of floral organs, in the control of specification of the floral meristem, and in the activation of APETALA3 and PISTILLATA.	UNUSUAL FLORAL ORGANS UFO	yes	meristem organization, floral meristem

On the basis of this array of representations, database users are able to peruse countless sets of data and eventually to correlate them with each other. This process quickly generates precious information concerning the quantity and types of available data that are classified as relevant to a given phenomenon; alternative ways of ordering the same dataset, for instance in relation to two different but biologically related phenomena; and even new hypotheses about how one dataset might correlate with another, or about how one phenomenon might be causally linked to another, which researchers might then go on to test experimentally. Researchers are able to gather this type of information from such an extensive dataset because data are presented in formats that do not take into account the differences in the ways in which they originated. Without de-contextualisation, it would be impossible to consult and compare such a high number of different items (remember that we are talking about several billions of data stored in each database), not to speak of distributing such information across research communities.

Once researchers have found data of interest to them, they can narrow their analysis down to that dataset and assess its quality and reliability. It is at this stage that they need to know more about how the data have been produced, by whom and for which purpose. The first step in that direction is the consultation of evidence codes. Information about the data producers, sources and materials used is also readily available

(see figure 8). This background knowledge is crucial to deciding whether to pursue the investigation further, and how.

Figure 8: Display of information about the provenance of a specific dataset in TAIR.



tair

Home Help Contact About Us Login Gene Search

Search Browse Tools Stocks Portals Download Submit News

TAIR Annotation Search [Help]

Your query for annotations based on the locus AT1G30950 resulted in 5 records.

Displaying 1 - 5 of 5 records on page 1 of 1 pages.

Keyword Category	Relationship Type	Keyword	Gene	*Evidence Code*: Evidence Description: Evidence With: Reference	Annotated By/ Date Last Modified
biological process	required for	meristem organization	UFO	*inferred from mutant phenotype*: analysis of visible trait: none: Samach, et al. (1999)	The Arabidopsis Information Resource/ 2003-03-27
biological process	involved in	ubiquitin-dependent protein catabolic process	AT1G30950.1	*traceable author statement*: none: none: Callis, et al. (2000)	The Institute for Genomic Research/ 2003-04-17
cellular component	located in	ubiquitin ligase complex	AT1G30950.1	*inferred from physical interaction*: none: none: Samach, et al. (1999)	The Institute for Genomic Research/ 2003-04-17
molecular function	has	ubiquitin-protein ligase activity	AT1G30950.1	*inferred from reviewed computational analysis*: manually reviewed TIGR computational analysis: INTERPRO:IPR001810: TIGR Arabidopsis annotation team (2005-02-17)	The Institute for Genomic Research/ 2002-01-02
plant structure	expressed in	floral meristem	UFO	*inferred from direct assay*: in situ hybridization: none: Samach, et al. (1999)	The Arabidopsis Information Resource/ 2003-03-27

The possibility to consult data coming from other contexts makes an enormous difference to the range of claims that those data can be brought to bear upon. Databases are efficient in making data travel in both a geographical and a scientific sense: data are used by researchers working in places very distant from the place where the data were first produced, as well as possessing scientific interests and goals other than the ones characterising the original context of data production.[11]

[11] In the case of UFO data, geographical travel can be traced through publications by Japanese and Australian labs that retrieved the Canadian data online and used them

2. Packaging for Travel

As I illustrated in the previous section, sciences such as experimental biology have a different way to deal with data and their connection to claims about phenomena than sciences such as physics or astronomy. This is not purely due to the size of datasets handled by scientists in these different areas, which tend to be massive across the board. Rather, the difference is due to the variability in the types of data available, and the purposes for which they are produced and used. Experimental biology is characterised by very diverse sets of data, obtained on different organisms and by appeal to varying theoretical perspectives, instruments, protocols and goals. Unpredictability of use is a prime characteristic of biological data, especially at the genomic level. Data are not produced to validate or refute a given claim about phenomena, but rather because biologists possess new technologies to extract data from organisms (such as micro array experiments and sequencing machines), and it is hoped that the potential biological significance of those data will emerge through comparisons and correlations among different datasets. This type of research is data-driven as much as it is hypothesis-driven: the characteristics of available data shape the questions asked about their biological relevance, and at the same time existing open questions about biological entities and processes (that is, claims about phenomena that need to be verified and validated) shape the ways in which data are produced and interpreted.

The inter-dependence between data types and claims about phenomena is made possible by packaging processes such as the ones used by database curators. Packaging here serves the purpose of connecting data to claims about phenomena in ways that differ from the one-way evidential connection depicted by B&W. Curators are required

---

for their own purposes (e.g. Ikeda et al 2007, Taylor et al 2001). Also, several publications that make reference to UFO data found in TAIR have epistemic cultures that differ widely from the one that produced the data: some are concern with different topics, such as leaf senescence (Ryun Woo et al 2001) and stress signalling (Devoto and Turner 2005); others are working with species other than Arabidopsis, such as rice (Ikeda et al 2007) and peas (Taylor et al 2001).

to find ways to connect data with claims about phenomena in a way that will not fix the value of data solely as evidence for specific claims, but rather will allow users to recognise the potential relevance of data for as many claims as possible. In this section, I focus on three elements of packaging and on the epistemic consequences of adopting these measures to make data travel across contexts.

Labels

The allocation of labels involves the selection of terms apt to classify the facts being packaged, thus making it possible to organise them and retrieve them. In the case of datasets in databases, the labels are the keywords used by curators to indicate the range of phenomena for which data might prove relevant.[12] For example, the pictures of stained Arabidopsis tissue resulting from the in situ experiment mentioned above are labelled "floral meristem" and "UFO gene," thus indicating that the pictures might provide evidence for claims about these phenomena (such as "the UFO gene regulates the development of the floral meristem"). Labels indicating phenomena are the only element used both for data classification and for the formulation of claims about phenomena for which data might provide evidential support. Indeed, the labelling system devised by curators has the crucial epistemic role of connecting the terms used to indicate which phenomena data can be taken to document with the terms used to formulate claims about those phenomena.

It is not at all obvious that the terms used to classify data and formulate claims about phenomena should or even could be the same. Indeed, scientists tend to keep these two labelling processes separate from each other, since they satisfy different demands arising from different circumstances. Labelling data for prospective retrieval and reuse means choosing terms referring to phenomena that are easily observed in the lab, either because they are very recognisable parts of an

[12] In what follows, I refer chiefly to databases that make use of Open Biomedical Ontologies, that is an ensemble of labelling systems sanctioned by a consortium of specialists as the most reliable and highly standardised in contemporary bioinformatics (see http://obo.sourceforge.net).

organism or a cell (e.g. “meristem,” “UFO gene,” “ribosome”), or because they are processes whose characteristics and mechanisms are widely established across research contexts (e.g. “mitosis”). By contrast, labelling phenomena for the purposes of formulating claims about them has the primary aim of ensuring that the resulting claims are compatible with the background knowledge and interests of the scientists adopting the labels. Any label that is seen as loosely compatible with observations will be accepted, as long as it fits the epistemic culture of the research context in question. This means that there will be as much variation among labels adopted to formulate claims about phenomena as there is variation across research cultures and beliefs. For instance, both immunologists and ecologists use the term “pathogen” as a label in formulating their claims; however, they use it to refer to two different phenomena (for immunologists pathogens are a specific type of parasites, while ecologists tend to view them as potential symbionts). Another example is the term “bud,” which is used by botanists to describe a protuberance on a stem or branch; other biologists, however, do not recognise that definition and use the term to indicate an asexual reproductive structure. Equally common are cases where the same phenomenon is described through different terms at different locations.

The multiplicity of definitions assigned to the same terms (and of terms assigned to the same definition) limits the power of the label to carry information across contexts. Indeed, the use of “local” labels is one of the reasons why journal publication is an inefficient means to disseminate data. Since journals in biology address very specific communities with their own unique epistemic culture, the keywords used to designate the phenomena discussed in each article are the ones which are most familiar to the audience that the journal is meant to address. Databases have ways to confront this issue. For a start, curators assign a strict definition to each term chosen as a label. These definitions are as close as possible to the definitions used by scientists working on the bench when formulating claims about phenomena. Yet,

they have two additional functions: to make definitions explicit, so that scientists can access them and critique them when needed, and to standardise those definitions that vary depending on the research contexts in which the corresponding term is used. Once specific labels are chosen and defined, curators examine cases where a different label is given the same definition or where several labels are proposed as fitting one definition. To accommodate the former option, curators create a system of synonyms associated with each chosen label in the database. For instance, the term “virion” is defined as “the complete fully infectious extracellular virus particle.” Given that some biologists use the term “complete virus particle” to fit this same definition, this second term is listed in the database as a synonym of “virion.” Users looking for “complete virus particle” are thus able to retrieve data relevant to the phenomenon of interest, even if it is officially labelled “virion.” Curators use another strategy for cases of substantial scientific disagreement on how a specific term should be defined. This is the use of the qualifier “sensu,” which allows them to generate sub-terms to match the different definitions assigned to the same term within different communities. This is especially efficient when dealing with species-specific definitions of terms: the term “cell wall” is re-labelled “cell wall (sensu Bacteria),” which is defined as peptidoglycan-based, and “cell wall (sensu Fungi),” which contains chitin and beta-glucan. As long as curators are aware of differences in the use of terms across communities, that difference will be registered and assimilated so that users from all communities will be able to query the database for data.

Database users cannot extract data without using the labels chosen by the curators for the purposes of their query. Thanks to the use of synonyms and “sensu,” this is true even in cases where the terms and definitions used by the user do not match the ones used to classify data in the database. As a consequence, users are not only “taught” the official labelling system preferred within the database, but they are also invited to accept those labels and definitions, at least for the purposes of

retrieving data and assessing their value as evidence (for example, a user looking for “complete virus particle” has to accept that this term is equivalent to the term “virion” in order to retrieve data thus classified). In this way, the packaging of data pulls phenomena along: users accessing data through the database do not only get the prospective evidence that they need, but also a specific interpretation of how the terms used as labels in the database refer to objects and processes in the world. When extracting data from a database, users implicitly agree to use the labels found in the database when formulating claims about phenomena for which those data serve as evidence.

The result of adopting this labelling system is to successfully implement an inter-dependence between the characteristics of data, which determine the terms used for their classification, and the characteristics of claims about phenomena, which determine the terms used to indicate phenomena. Using such a labelling system is crucial to the process of packaging data. The labels with which it refers to phenomena are recognised across scientific contexts as applying both to the classification of datasets and to the formulation of claims about phenomena. This fit between labels used to refer to phenomena allows data to travel efficiently across a wide spectrum of communities, even in the face of large disparities in epistemic cultures.

Vehicles

A second key component of packaging is vehicles such as the journals used to make claims travel and the databases disseminating data. Already from these two examples, it is clear that the technological infrastructure used to package facts makes a big difference to how efficiently they will travel and thus become non-local. The possibility to delegate efforts to computers has changed the speed, breath and accuracy with which data can be analyzed. Through tools such as unique identifiers, data are made recognizable to the available software, which means that they are searchable according to the parameters set in the

dabatase. A high degree of automation in the handling of data means a fast and reliable access to those data, as well as the possibility for curators to insert vast amounts of data without investing too much time and manual labor. It is true, as argued by Strasser (2006), that databases are in many ways continuing the natural history tradition of collecting, and as such they inherit the conceptual and practical problems of how to classify disparate objects. However, the software used in databases, combined with the flexibility of virtual space in the world wide web, makes it possible to order and retrieve data in ways unimaginable until three decades ago. The Web Ontology Language (OWL), for instance, is one of many schema languages developed to facilitate interoperability between controlled vocabularies. Many of these technologies are developed as part of the Semantic Web, which is broadly defined as

> an extension of the current Web that enables navigation and meaningful use of digital resources by automatic processes. It is based on common formats that support aggregation and integration of data drawn from diverse sources. (Ruttenberg et al 2007, 3)

The system of labels and synonyms devised by database curators only works thanks to software and HTML interfaces geared to be widely accessible, extendable, and as flexible and decentralized as possible. This point has been recognized by most historical and sociological work devoted to databases.[13] Its enormous epistemic implications, however, have yet to be explicitly discussed. Without the layered structure and immense capacity for storing information characterizing databases, it would be impossible for curators to classify data separately, and yet in relation to, information about their provenance. Further, languages such as OWL make it technically possible to align labels used to classify data with labels used to formulate claims about phenomena. Finally, the multiple search tools developed through XML imply that databases can

[13] See for instance Hilgartner (1995), Bowker (2000) and Zimmerman (2007).

adapt to the needs and expertise of their users, allowing them to phrase their queries in the terms most familiar to them (as I explained above in the case of labels and their synonymous terms). Information technology provides novel tools through which to implement classificatory criteria for the purpose to circulate information.

Agency

Technology by itself does not offer ready-made solutions to the problem of devising criteria through which to order, store and distribute data: the classificatory work is primarily a conceptual effort and it remains the responsibility of curators, who need a fitting expertise to accomplish this task. The third element required in packaging is thus appropriate agency, i.e. the exercise of skills geared towards making facts travel.[14] When mining and annotating data from papers, curators have to bridge a conspicuous gap between (A) the information available in and about the publications and (B) the information required by databases to classify the potential relevance of data as evidence. (A) encompasses the names and affiliations of the authors and the actual text of the publication. (B) includes a description of the data included in the publication, an estimate of its relation to other datasets in the database, its classification through an appropriately chosen label, and eventual comments by the curator as to how the data should be interpreted given their characteristics (such as their format or the organism on which they were obtained). Information of type (B) is often not displayed in the text of the relevant papers. Partly this is because authors are writing for an audience of specialists who do not need to have every detail of the adopted technique and preferred evidence spelled out. Further, it is because data are packaged as evidence for one specific claim. Curators are thus not just cutting information from the papers and pasting it into their databases. They need to interpret the content of the papers in the light of their own

[14] This is why curators are sometimes referred to as “intermediaries” (e.g. Markus 2001): as recognised by scholarship in information technology management, “it takes organisational work to develop local knowledge for broader use” (Brown et al 1998, 99).

familiarity with the techniques and methods used in that field, so as to be able to extract the (B) type of information needed to package data in the database.

This means, on one hand, that curators need to be acquainted with as many fields, experimental methods, and epistemic cultures as possible, so as to be able to understand and combine results arising from different expertises without committing mistakes or misrepresenting research projects. On the other hand, curators need to be aware of what it is like to experiment with model organisms on a daily basis. Indeed, curators often come "from the bench": they have been experimenters before entering bioinformatics and their contribution to databases is heavily informed by their awareness of what it takes to manipulate model organisms in the laboratory.[15] The apparently simple information about a gene's location, expression and (especially) biological functions are actually very difficult to extract from a bunch of papers written by different authors for a variety of mixed purposes. The curator's ability is not so much to annotate the data so that they match the chosen keywords: it is to comprehend the experiments described in the papers (including how to handle the instruments, prepare the organisms and compare results obtained from different experimental techniques) so as to be able to extract the relevant data and information about the sources of the evidence. In this sense, curators should be seen as specialists in packaging: their expert knowledge of how to intervene is both generalist and informed by details, and this mix of generality and specialization is necessary to bridge between the differing expertises of experimenters.

3. On the Locality of Data and Claims

I now consider the consequences of successful packaging on the locality of data and claims about phenomena. I take a fact to be local, when its evidential scope depends on the context in which it has been produced; non-local, when its evidential scope can be determined without reference

[15] This finding is based on several interviews to curators working in a variety of databases which I conducted between 2004 and 2007.

to that context. By evidential scope I mean the value of the fact as evidence for a specific set of claims, which in the case of data will be claims about phenomena, and in the case of claims about phenomena will be more general theoretical claims. For instance, a dataset is local when researchers need to be acquainted with the method, tools, materials and background knowledge originally used to produce it so as to understand which claims it can be taken to support. A claim about phenomena is local when its interpretation as evidence for theories depends on the background and tacit knowledge possessed by researchers using it.

The distinction between the role of data as evidence for claims about phenomena, and the role of claims about phenomena as evidence for theories, is one put forward by B&W (1988, 306). In their contribution to the debate on the evidential value of data, B&W use this distinction to argue that data and claims about phenomena have intrinsic degrees of locality. Data are “idiosyncratic to particular experimental contexts, and typically cannot occur outside of those contexts” (B&W 1988, 317). This is because their characteristics are “heavily dependent on the peculiarities of the particular experimental design, detection device, or data-gathering procedures an investigator employs” (ibid.). Claims about phenomena are non-local or at least intrinsically less local than data, since they “can occur in a wide variety of different situations or contexts” (ibid.). B&W see phenomena as real objects in the world, which they characterise as having “stable, repeatable characteristics which will be detectable by means of a variety of different procedures, which may yield quite different kinds of data” (ibid.). Data by contrast are the product of human interactions with the world. They carry information about what the world is like, but such information is expressed in ways that can only be properly understood and interpreted by scientists who are familiar with the functioning, output and setting of the instruments through which data are acquired. Knowledge about phenomena thus needs to be freed from its embedding in data (and thus in the local practices through which data

are obtained) in order to be shared among scientists irrespectively of their familiarity with the means through which it has been acquired. "Liberation" comes through the formulation of claims about phenomena: data help scientists to infer and validate those claims, yet ultimately it is the claims about phenomena that travel around the scientific world and are used as evidence for general theories. Hence B&W see data as local evidence for non-local claims: data serve to produce reliable descriptions of the world which form the evidential basis for scientific theories and explanations. Once data have been used as evidence for a claim, their only remaining scientific use is that of guaranteeing the credibility of that claim: as observed by Ronald Giere, data are "something available for public scrutiny," bearing witness to the scientists' powers of observation and analysis (Giere 2006, 31).

As I showed in the case of UFO data, data published through a paper are indeed selected and formatted to fit their role as evidence for a claim about a phenomenon ("gene X regulates developmental process Y"). In line with B&W's arguments, what readers are required to take away from the paper is not the data themselves, but rather the empirical interpretation of those data provided by the authors (and approved by peer reviewers) in the form of a claim about the phenomenon of flowering. Disclosure through publication is, however, increasingly complemented (and in some cases supplanted) by disclosure through databases or other tools geared towards making data, rather than claims, travel. This signals a change in how biological knowledge is constructed[16] which Bogen and Woodward, writing in 1988, could hardly have anticipated. Yet, it also signals biologists' uneasiness with the system of disclosure through publication and their willingness to develop alternative systems of data sharing, which I believe stems precisely from the scientific need to circulate and use data far beyond their context of production.

[16] See for instance the analyses offered by Gilbert (1991), Hilgartner (1995), Bowker (2000), Strasser (2006), Garcia-Sancho (2007) and contributions to Rheinberger and Gaudillere (2004).

My analysis of packaging shows that transport through databases expands the evidential scope of data in several ways. It makes data accessible to other research contexts and therefore potentially re-usable as evidence for new claims; and it associates data with a broader range of phenomena than the one to which they were associated in the context of production. This brings me to contest B&W's idea that data are intrinsically local: data can in fact be made non-local through the use of appropriate packaging processes. Data that travel through databases become non-local. They travel in a package that includes information about their provenance, but they can be consulted independently of that information. This is a way to "free" data from their context and transform them into non-local entities, since the separation of data from information about their provenance allows researchers to judge their potential relevance to their research. This is different from judging the reliability of data within a new research context. This second type of judgment requires researchers from the new context to access information about how data were originally produced and match it up with their own (local) criteria for what counts as reliable evidence, as based on the expertise that they have acquired through their professional experience in the lab. What counts as reliable evidence depends on scientists' familiarity with and opinion of specific materials (for instance, the model organism used), instruments, experimental protocols, modelling techniques and even the claims about phenomena that the evidence is produced to support. Thus, data judged to be reliable become once again local: what changes is the research context that appropriates them. The successful packaging of data for travel involves both the de-contextualisation of data, without which it would be impossible to retrieve them online according to their relevance to phenomena, and the re-contextualisation of data for use in a new research context. A journey has to start and finish in a specific place: the temporary de-contextualisation achieved through packaging processes ensures that the trajectory of the journey – and thus the

evidential scope of data across contexts - is determined by the receiving context rather than by the production site.

The second point in B&W's account that I wish to critique is the idea that claims about phenomena are intrinsically non-local, or at least less local than data themselves.[17] My analysis of packaging shows how scientists' interpretation of a claim about phenomena is always situated by their specific background knowledge, skills and interpretive framework or perspective. As evident from my discussion of labels, the classification and definition of phenomena depends on the interests and expertise of the scientists who investigate them. Phenomena are always described on the basis of a specific epistemic culture – a combination of the concepts used to classify, share and reason upon data, the skills used to interact with the world, past experience and socialisation. This holds also for the curator's attempt to develop non-local labels for phenomena, which requires them to nurture a cross-disciplinary expertise mediating between the local epistemic cultures that characterise research at the bench. Like data, claims about phenomena only acquire non-local value through apposite packaging. The non-locality of claims is an important scientific achievement, requiring the selection and efficient implementation of packaging elements such as the labels, vehicles and agency outlined above.

B&W appeal to the intrinsic non-locality of claims about phenomena in order to defend their main argument about the evidential value of these claims: "facts about phenomena are natural candidates for systematic explanation in a way in which facts about data are not" (1988, 326). This argument is compatible with the view I propose. Claims about phenomena do have a different epistemic role from data. I contest the idea that this difference can be accounted for as a question of locality.

[17] McAllister rightly points out that B&W do not clearly distinguish claims about phenomena (in the sense of patterns in data sets) from phenomena themselves (in the sense of investigator-independent constituents of the world). I agree with McAllister's characterisation of phenomena as "labels that investigators apply to whichever patterns in data sets they wish so to designate" (1997, 224): here I focus on the epistemic status of those labels rather than the ontological status of the notion of phenomenon.

Claims about phenomena are privileged candidates for systematic explanations because they are propositions, rather than dots on a slide, photographs or numbers generated by a machine. Both claims about phenomena and data aim to carry information: claims are formulated to make a specific interpretation of data available to others and data are produced through standardised procedures so as to be legible by a community of scientists. The difference between data and claims does not consist in the degree of locality of the elements of which they are made, but rather on the ease with which they can be integrated into formal structures such as theories or explanations. Claims about phenomena are expressed so as to be tractable as evidence for a general formula or statement. For instance, the claim “gene X regulates the development of trait Y” can be used as evidence for a general statement about gene regulation, such as “genes regulate the development of morphological traits.” Data about the UFO gene, whether in the form of gene sequences or photographs of embryos, are not directly useable as evidence for such a statement – nor were they produced to fulfil this aim. To be useful for such purpose they need to be interpreted: first through association with a specific set of phenomena, then through a proposition expressing what they are taken to demonstrate.

My conclusion is that the effectiveness of claims about phenomena as mediating between data and theoretical statements does not come from the supposed constancy and stability of such claims across contexts, as argued by B&W (1988, 326), but rather from their tractability for the purposes of producing formal theories and propositional explanations. As illustrated in the case of data travelling through databases, data are most tractable for other scientific purposes, such as the discovery of correlations (through statistical analysis or direct comparison of datasets, depending on the format of the data in question) and the formulation of hypotheses to be tested.

Conclusion: Degrees of Locality

There are many cases in scientific research where data function as non-local entities that are shared and used across a wide range of research contexts. In the case of biological databases, data travel well when packaged separately from information about their production; at the same time, access to such information is needed to assess the reliability of data, so as to be able to use them in a new research context. Through the process of labeling, database curators select and define the range of phenomena to which data could be associated. The labeling system is then implemented in a vehicle – the database -capable of storing vast quantities of diverse types of data and retrieving them according to the specific interests and competences of each user. As a result, database users can quickly assess the evidential scope of data, that is their potential relevance to specific research areas and projects. Databases thus help scientists to form new hypotheses that might guide future research. For instance, spotting that the same gene participates in the regulation of two different developmental processes might lead to an investigation of the evolutionary links between the two; or finding a correlation between gene expression levels and its metabolic function might suggest that the genes involved play a regulatory role. When the time comes to actually test those hypotheses, biologists can examine information about the original context of data production. This helps them to assess the reliability of the data within the new context and thus eventually re-use them: data become local again, but in a context other than their production site.

One implication of this analysis is that non-locality is a scientific achievement, obtained through complex processes of packaging.[18] This holds for data as well as for claims about phenomena, whose non-locality depends on the extent to which the terms and tacit knowledge used in

[18] This argument is closely associated with and inspired by Bruno Latour's work on "immutable mobiles" and the circulation of references (Latour 1987, 1999). While Latour has focused his analysis on the social conditions for and implications of the scientific achievement of non-locality, however, I am here interested in the epistemic import of packaging processes.

their formulation are standardised and widespread. Indeed, both data and claims about phenomena can have varying degrees of locality. This contrasts with B&W's characterisation of data as local and claims as non-local. It does not, however, betray B&W's fundamental concerns with the relation between local and non-local knowledge: I am taking those concerns one step further, by noting how both claims and data are made in very specific and possibly unique contexts, and need to comply with precise requirements to be granted significance outside that context. Model organism biology provides a particularly apt setting to investigate this issue, as the applicability of results (whether data or claims) beyond the species or even the individual organism on which they are obtained needs to be evaluated in every single case of "travel." The same dataset can have very different evidential scope depending on the research context and especially the organism on which it is brought to bear: a specific gene sequence might be used as evidence for a variety of claims in Drosophila melanogaster, while it might only be useful as evidence for one claim in Mus musculus and might have no evidential value at all in Arabidopsis thaliana.

This brings me to conclude that, at least in some scientific realms such as model organism biology, it is not only possible but even desirable for data and claims about phenomena to have a variable evidential scope and, as a consequence, variable degrees of locality. This variability depends on the scientific, material and social circumstances in which facts are packaged for travel. The efficient packaging of genomic data, whose functional significance is still so far from clear and whose production requires large efforts and resources, is now a priority in biological research: both researchers and their sponsors aim to make genomic data travel as widely as possible, so as to exploit their large potential as evidence for future discoveries. At the same time, it is vital for researchers to keep in mind that genomic data might turn out to be more local than initially assumed, given the immense variations in

the relevance of these data across species and our increasing knowledge about the genetic basis for biodiversity.

In closing, the emphasis on the variable degrees of locality of data and claims about phenomena brings me back to a well-known argument put forward by Pierre Duhem a century ago. Duhem introduced his thesis about the underdetermination of theories by data[19] by observing that

> An experiment in physics is the precise observation of phenomena accompanied by an interpretation of these phenomena; this interpretation substitutes for the concrete data really gathered by observation abstract and symbolic representations which correspond to them by virtue of the theories admitted by the observer (Duhem 1974 [1914], 147)

Duhemian underdetermination has long been a source of headaches to naïve realist and strict inductivist philosophers. Yet, it is often far from being a problem for scientists, many of whom need to access data gathered through experiment without necessarily buying the interpretation of those data proposed by the original "observer." The study of how data travel in biology shows that the more claims data can be used as evidence for, the more science may develop. Data are a resource that scientists need to maximize: biologists try to extract as much knowledge as possible from the same datasets. Underdetermination in this case is a strength of scientific research. Data produced with high efforts and costs can and arguably should be used as evidence for a variety of claims about phenomena. This will not happen by itself: as I have shown, data travel requires effort and appropriate means. Further, in many scientific disciplines it is not yet clear precisely how and with which consequences data can be made to travel. This is a

[19] "The physicist can never subject an isolated hypothesis to experimental test, but only a whole group of hypotheses; when the experiment is in disagreement with his predictions, what he learns is that at least one of the hypotheses constituting this group is unacceptable and ought to be modified; but the experiment does not designate which one should be changed" (Duhem 1974[1914],187).

place where philosophy of science can effectively cooperate with science itself, by contributing to a better understanding of procedures, strategies and implications of data travel.

Bibliography

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

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

Bowker, GC (2000) Biodiversity Datadiversity. *Social Studies of Science*, 30, 5: 643-683.

Brown, JS and Duguid, P (1998) *Organising Knowledge*. California Management Review 40, 3: 90-111.

Devoto, A and Turner, JG (2005) Jasmonate-regulated Arabidopsis stress signalling network. *Physiologia Plantarum* 123:2, 161–172

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

Garcia-Sancho, M (2007) *Sequencing as a Way of Work: A History of Its Emergence and Mechanisation – From Proteins to DNA, 1945-2000*. PhD Dissertation, Imperial College London, UK.

Gaudilliere JP and Rheinberger, HJ (eds.) (2004) *From Molecular Genetics to Genomics*. Routledge.

Gilbert, W (1991) Towards a Paradigm Shift in Biology. *Nature* 349, 6305: 99

Hacking, I (1983) *Representing and Intervening*. Cambridge University Press.

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: 240-263.

Hilgartner, S (2004) Making Maps and Making Social Order: Governing American Genome Centers, 1988-1993. In Gaudilliere JP and

Rheinberger, HJ (eds.) *From Molecular Genetics to Genomics*. London: Routledge, pp. 114-128.

Ikeda, K, Ito, M, Nagasawa, N, Kyozuka, J and Nagato, Y (2007) Rice ABERRANT PANICLE ORGANIZATION 1, encoding an F-box protein, regulates meristem fate. *The Plant Journal* 51, 6: 1030–1040

Knorr Cetina, KD (1999) *Epistemic Cultures: How the Sciences Make Knowledge*. Cambridge: Harvard University Press.

Kroch, 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.

Latour, B (1987) *Science In Action: How to Follow Scientists and Engineers Through Society*. Harvard University Press, Cambridge MA.

Latour, B (1999) Circulating reference: Sampling the soil in the Amazon forest. In *Pandora's hope: Essays on the reality of science studies*. Cambridge, MA: Harvard University Press, pp.24-79.

Leonelli, S (2007) Arabidopsis, the Botanical Drosophila: From Mouse-Cress to Model Organism. *Endeavour* 31, 1: 34-38.

Leonelli, S (forthcoming A) Bio-Ontologies: Integrative Perspectives to Make Facts Travel.

Leonelli, S (forthcoming B) When Data Travel Well: Strategies and Skills in Database Curation.

Markus, ML (2001) Toward a Theory of Knowledge Reuse: Types of Knowledge Reuse Situations and Factors in Reuse Success. *Journal of Management Information Systems* 18, 1:57-93.

McAllister, JW (1997) Phenomena and Patterns in Data Sets. *Erkentniss* 47: 217-228.

Rhee, SY et al (2003) The Arabidopsis Information Resource (TAIR): a Model Organism Database Providing a Centralised, Curated

Gateway to Arabidopsis Biology, Research Materials and Community. *Nucleic Acid Research* 31, 1: 224:228.

Rhee, SY, Dickerson, J and Xu, D (2006) Bioinformatics and Its Applications in Plant Biology. *Annu. Rev. Plant Biol.* 57: 335-360.

Samach et al (1999) The UNUSUAL FLOWER ORGANS Gene of Arabidopsis thaliana is an F-box Protein Required for Normal Patterning and Growth in the Floral Meristem. *The Plant Journal* 20, 4: 433-445.

Strasser, BJ (2006) Collecting and Experimenting: The moral economies of biological research, 1960s-1980s. *Preprints of the Max-Planck Institute for the History of Science* 310: 105-123.

Taylor, S, Hofer, J and Ian Murfet (2001) Stamina pistilloida, the Pea Ortholog of Fim and UFO, Is Required for Normal Development of Flowers, Inflorescences, and Leaves. *The Plant Cell* 13, 1: 31–46

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: 5-16.

Hye Ryun Woo, Kyung Min Chung, Joon-Hyun Park, Sung Aeong Oh, Taejin Ahn, Sung Hyum Hong, Sung Key Jang and Hong Gil Nam (2001) ORE9, an F-Box Protein That Regulates Leaf Senescence in Arabidopsis. *The Plant Cell* 13: 1779-1790

LONDON SCHOOL OF ECONOMICS
DEPARTMENT OF ECONOMIC HISTORY

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

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

2005

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

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

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

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

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

06/05: Market Disciplines in Victorian Britain
Paul Johnson

2006

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

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

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

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

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

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

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

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

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

2007

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

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

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

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

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

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

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

2008

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

24/08: Travelling in the Social Science Community: Assessing the Impact of the Indian Green Revolution Across Disciplines
Peter Howlett

25/08: Circulating Evidence Across Research Contexts: The Locality of Data and Claims in Model Organism Research
Sabina Leonelli