Published in Philosophy of Science, 2009.

On the Locality of Data and Claims about Phenomena*

Sabina Leonelli † ‡

† ESRC Centre for Genomics in Society

University of Exeter

Byrne House, St Germans Road

EX4 4PJ Exeter, UK

s.leonelli@exeter.ac.uk

‡ Mary Morgan's insightful suggestions and advice have been crucial to the development of this paper. Warm thanks also to Maureen O'Malley, the 'facts' group, Julian Reiss, David Teira and audiences at the SPSP 2007 and the PSA 2008 for very helpful discussions. This research was funded by the Leverhulme/ESRC project "The Nature of Evidence: How Well Do 'Facts' Travel?" (grant number F/07004/Z) at the Department of Economic History, LSE; and by the ESRC Centre for Genomics in Society at the University of Exeter.

Bogen and Woodward (1988) characterise *data* as embedded in the context in which they are produced ('local') and *claims about phenomena* as retaining their significance beyond that context ('non-local'). This view does not fit sciences such as biology, which successfully disseminate data via packaging processes that include appropriate labels, vehicles and human interventions. These processes enhance the evidential scope of data and ensure that claims about phenomena are understood in the same way across research communities. I conclude that the degree of locality of both data and claims about phenomena varies depending on the packaging used to make them travel and on the research setting in which they are used.

1. Introduction

Can data[1] be circulated independently of the claims for which they are taken as evidence, so as to be used in research contexts other than the one in which they have been produced? And in which ways and with which consequences does this happen, if at all? This paper tackles these questions through a study of how biological data travel across research contexts. 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 about phenomena 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 one 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.

2. Making Facts About Organisms Travel

Biology yielded immense amounts of data in the last three decades. This is especially due to genome sequencing projects, which generated billions of datasets about the DNA sequence of

[1] I follow Hacking's broad definition of data as any 'marks' produced by a 'data generator' (1992, 48). Biological data, for instance, include material objects (stains on an embryo), dots on slides (micro arrays) and strings of letters (DNA sequences).

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, morphological effects correlated to ‘knocking out’ specific genes, and so forth (Krohs and Callebaut 2007). These results are obtained through experimentation on a relatively small set of species whose features are particularly tractable through the 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 (Ankeny 2007).

Researchers are aware that assuming model organisms to be representative for other species is problematic, 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 few species enables researchers to integrate data about different 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. The circulation of data across research contexts is therefore considered a priority in model organism research: the aimed-for cooperation can only spring through an efficient sharing of results.

The quest for efficient means to share data has become a lively research area in its own right, usually characterised as a branch of bioinformatics. One of its main objectives is to exploit new technologies to construct digital databases that are freely and easily available for consultation (Rhee et al 2006). Aside from the technical problems of building such resources, bioinformaticians have to confront two main issues: (1) the fragmentation of model organism biology into epistemic communities with their own expertise, traditions, favourite methods, instruments and research goals (Knorr Cetina 1999), which implies finding a vocabulary and a format that make data retrievable by anyone according to their own interests and background; and (2) the diverse characteristics of data coming from disparate sources and produced for a variety of purposes, 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 how bioinformaticians tackle these issues 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.[2]

i. Disclosure

The majority of experimenters disclose their results through publication in a refereed journal. Publications include a description of the methods and instruments used, a sample of the data thus obtained and an argument for how those data support the claim that researchers are trying to make. Data disclosed in this way 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. Because of these selection criteria, a large amount of data produced through experiment 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. There is little chance that researchers working in another area and interested in different claims will see those data and thus be in a position to evaluate their

[2] For an exemplification of how these phases work in biological practice, see Leonelli 2008a.

relevance to their own projects. Hence, those data will never be employed as evidence for claims other than the one that they were produced to substantiate.[3]

If disclosure is left solely to publications, data will travel as the empirical basis of specific claims about phenomena and will rarely be examined independently of those claims. Biologists and their sponsors are unhappy with this situation, since maximising the use made of each dataset means maximising the investments made on experimental research, by avoiding the danger that an existing dataset is overlooked by scientists who could use it towards new discoveries.

ii. Circulation

Community-based databases, devoted to collecting data about specific model organisms, constitute an alternative to the inefficient disclosure provided by publications. The curators responsible for the construction and maintenance of databases ground their work on one crucial insight. This is that biologists consulting a database wish to see 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 hybridised embryo. These marks constitute unique documents about a specific set of phenomena. Their production is constrained by the experimental setting and the nature of the entities under scrutiny; still, researchers with differing interests and expertises might find different interpretations of the same data and use them as evidence for a variety of claims about phenomena.

---

[3] There are of course many examples of scientific breakthroughs made on the basis of borrowing data from another field. What I argue here is not that data travel via publications is impossible in principle, but that it is difficult in practice given the current proliferation of specialised journals and increase in publication rates.

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 produced to validate. 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. The source becomes important, however, when adding credibility to the claim: *to evaluate the quality of a claim we need to know how the claim originated*. Knowing whether to trust a rumour depends on whether we know where the rumour comes from and why it was spread; a politician wishing to assess a claim about climate change needs to reconstruct the reasoning and methods used by scientists to validate it. 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 can only be assessed by reference to how they are produced. Curators are aware of this seeming paradox and strive to find ways to make data travel, without however depriving researchers who 'receive' those data from the opportunity to assess their reliability according to their own criteria.

To confront the challenge, data are 'mined' from sources such as publications; labelled with a machine-readable 'unique identifier' so as to be computed and analysed; and classified via association with keywords signalling the phenomena for which each dataset might be used as evidence. Data are thus taken to speak for themselves as marks that are potentially applicable, in ways to be specified, to studying the range of phenomena indicated by keywords. At the same time, the structure of databases enables curators to store as much information as possible about how data have been produced. This is done through so-called *meta-data*, including 'evidence codes' classifying each dataset according to the method and protocol through which it was obtained; the model organism and instruments used in the experiment; the publications or repository in which it first appeared; and the contact details of the researchers responsible, who can therefore be contacted directly for any question not directly answered in the database.

iii. Retrieval

Retrieval happens through a variety of *search tools*, each of which accepts queries about one type of items: e.g. genes, markers, ecotypes, polymorphisms. Search results are displayed through digital models that visualise data according to the parameters requested by users and thus to their own expertise and demands. A query phrased through appropriate keywords typically yields links to datasets associated with the keywords. By clicking on these links, users can quickly generate information concerning the quantity and types of data 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, which researchers might then test experimentally. Researchers would not be able to gather this type of information unless data were presented in formats that do not take into account the differences in the ways in which they were originated. Without de-contextualisation, it would be impossible to consult and compare such a high number of different items, not to speak of distributing such information across research communities.

Once users have found data of particular 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 will need to know how the data have been produced, by whom and for which purpose. The first step in that direction is the consultation of meta-data such as evidence codes, which provide access to the information needed to decide whether and how to pursue the investigation further.

3. Packaging for Travel

Due to the variability in the types of data available, and in the purposes for which they are used, sciences such as experimental biology have specific ways to deal with data and their connection to claims about phenomena. Especially in contemporary genomics, many datasets are not produced to validate or refute a given claim about phenomena, but rather because biologists possess new 'high-throughput' technologies to extract data from organisms and it is hoped that the biological significance of those data will emerge through comparisons 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, while theoretical questions about biological entities and processes shape the ways in which data are produced and interpreted.

This inter-dependence between data types and claims about phenomena is made possible by packaging processes such as the ones used by database curators, which connect data to claims about phenomena in ways that differ from the one-way evidential relation depicted by B&W. Curators aim to enable users to recognise the potential relevance of data for as many claims as possible. I now focus on three elements of packaging and on the epistemic consequences of adopting these measures to make data travel.

Labels

The allocation of labels involves the selection of terms apt to classify the facts being packaged, thus making it possible to organise and retrieve them. In the case of datasets in databases, the labels are the keywords indicating the range of phenomena for which data might prove relevant.[4] These labels are used to classify data *as well as* to formulate claims about phenomena for which data provide evidential support. Indeed, the labelling system devised by curators has the crucial epistemic role of *streamlining the terms used to indicate phenomena for which data function as evidence with the terms used to formulate claims about those phenomena*.

In research circumstances such as the ones examined by B&W, the labelling procedure underlying the formulation of claims is kept quite distinct from the labelling procedure underlying the classification of experimental evidence: these two processes satisfy different demands arising from different circumstances and scientists thus tend to keep them separate from each other. Labelling data for prospective retrieval and re-use means choosing terms referring to phenomena that are easily observed in the lab, either because they are very recognisable parts of an organism or a cell ('meristem', 'ribosome'), or because they are processes whose characteristics and mechanisms are widely established across research contexts ('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, and will therefore fit existing theories and explanations: any label seen as somehow compatible with observations will be accepted, as long as it fits 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, the term 'bud' is used by botanists to describe a protuberance on a stem or branch; however, other plant biologists do not recognise that definition and use the term to indicate an asexual reproductive structure. Basic terms such as 'pathogen' and 'cell wall' have different meanings within different model organism

---

[4] Further details about this classification system, named 'bio-ontology' by the biologists who created it, can be found in Leonelli 2008b.

communities. 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) within experimental biology limits the power of these terms 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. Databases have ways to confront this issue. First, curators assign a strict definition to each term chosen as a label. These definitions are as close as possible to the ones used by scientists working on the bench; yet, they are meant to be understood in as many research contexts as possible. To this aim, once the labels are chosen and defined, curators examine the cases where a different label is given the same definition or several labels are proposed as fitting one definition. To accommodate the former option, curators create a system of synonyms associated with each label. 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 the phenomenon of interest, even if it is officially labelled 'virion'. Curators use another strategy for cases of substantial disagreement on how a specific term should be defined. They use the qualifier '*sensu*' 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: the term 'cell wall' is labelled 'cell wall (*sensu Bacteria*)' (peptidoglycan-based) and 'cell wall (*sensu Fungi*)' (containing chitin and beta-glucan). As long as curators are aware of differences in the use of terms across communities, that difference is registered and assimilated so that users from all communities are able to query 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, but they are also invited to accept those labels and definitions for the purposes of retrieving data and assessing their value as evidence. Users accessing data get a specific interpretation of how the terms used as labels for data refer to objects and processes in the world, which can be used when formulating claims about phenomena for which those data are serve as evidence. The result is the sought-for inter-dependence between the terminology used to classify data and the terminology used to formulate claims about phenomena. Thus the packaging of data pulls phenomena along.

Vehicles and Agency

A second key component of packaging is *vehicles*. Already from the examples of journals and databases, it is clear that the technological infrastructure used to package facts makes a big difference to how efficiently they travel. The possibility to delegate efforts to computers has changed the speed, breath and accuracy with which data are analyzed. Through tools like unique identifiers, data are made recognizable to software. A high degree of automation in the handling of data means a fast and reliable access to those data, as well as the possibility to curate vast amounts of data with minimal investments in time and labor. 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 system of labels and synonyms devised by database curators could not work unless it was supported by appropriate software and widely accessible HTML interfaces. This point has important epistemic implications. Without the layered structure and immense capacity for storing information characterizing databases, it would be impossible for curators to classify data separately from, and yet in relation to, information about their provenance. Also, the multiple search tools developed through XML imply that databases can adapt to the needs and expertise of their users, allowing them to phrase their queries in the terms most familiar to them (as in the case of labels and their synonyms).

Technology by itself does not offer ready-made criteria through which to order and distribute data: the classificatory work remains the responsibility of curators, who need a fitting expertise to accomplish this task. The third component of packaging is thus appropriate *agency,* i.e. the exercise of skills geared towards making facts travel. 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 text of the publication. (B) includes a description of the data included in the publication, an estimate of their relation to other datasets, its classification through an appropriately chosen label, and eventual comments as to how the data should be interpreted given 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: authors are writing for an audience of specialists and data are packaged as evidence for one specific claim. To extract (B) type of information, curators need to interpret the content of the papers in the light of their own familiarity with the techniques and methods used in that field.

This means that curators need to be acquainted with as many 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. 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.

4. 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 to that context. By *evidential scope* I mean the range of claims for which a fact can be taken as evidence, 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 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). Claims about phenomena have 'stable, repeatable characteristics' and can thus 'occur in a wide variety of different situations or contexts' (ibid.). Data 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 setting in which data are acquired. Knowledge about phenomena need to be freed from its embedding in data (and thus in the local practices through which data are obtained) so as 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. To B&W, data are local evidence for non-local claims.

Data published through a research article are indeed selected as evidence for one claim about phenomena. In line with B&W's arguments, what readers are required to take away from a paper is not the data themselves, but rather the empirical interpretation of those data provided by the authors in the form of a claim. Disclosure through publication is, however, increasingly complemented (and in some cases supplanted) by disclosure through databases geared to make data travel. This signals a change in how biological knowledge is constructed which Bogen and Woodward could hardly have anticipated.[5] At the same time, it also signals biologists' uneasiness with disclosure through publication and their willingness to develop alternative systems of data sharing, which stems precisely from the need to circulate and use data beyond their context of production.

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 production context. 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

[5] See Gilbert 1991 and Krohs and Callebaut 2007.

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 local once again: what changes is the research context that appropriates them.

Another questionable idea in B&W's account is that claims about phenomena are intrinsically non-local, or anyhow less local than data themselves. My analysis of packaging shows how scientists' interpretation of the terms used in claim about phenomena is always situated by their specific background knowledge and skills.[6] 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. This also holds for the curators' attempts to develop non-local labels for phenomena, which requires a specific expertise mediating between the local cultures of research at the bench. Like data, claims about phenomena only acquire non-local value through apposite packaging of the terms used to refer to phenomena. The resulting non-locality of claims is an

[6] McAllister defines phenomena as 'labels that investigators apply to whichever patterns in data sets they wish so to designate' (1997, 224).

important scientific achievement, requiring the selection and efficient implementation of packaging elements such as the labels, vehicles and agency discussed 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). I certainly agree with the intuition that claims about phenomena have a different epistemic role from data. What I wish to contest is the idea that this difference can be accounted for as a question of locality. 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 data and claims about phenomena aim to carry information. The difference between them consists in 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. Data are not useable as evidence for such statements: they are tractable for other scientific purposes, such as the discovery of correlations (through statistical analysis or comparison of datasets) and the formulation of hypotheses to be tested; further, they need to fulfil empirical demands, such as reproducibility and precision, which cannot be placed on claims about phenomena.

5. Conclusion: Degrees of Locality

The effectiveness of claims about phenomena as mediating between data and theoretical statements does not come from the supposed constancy and stability of the meaning of claims across contexts, but rather from their tractability for the purposes of producing formal theories and propositional explanations. Indeed, I have argued that both data and claims about phenomena can have varying degrees of locality. This variability depends, on the one hand, on the circumstances in which facts are packaged for travel. The efficient packaging of genomic data, whose functional significance is far from clear and whose production requires large efforts, is currently construed to allow for the greatest variability in degrees of locality: from entirely non-local (when first extracted from a database) to local according to the parameters set by the research contexts in which data are re-used. Thanks to this variability, researchers can exploit the potential of data as evidence for future discoveries. On the other hand, variability also depends on the research settings in which data and claims are used. When travelling across diverse research contexts, data and claims are interpreted in different ways depending on the context which adopts them. In the case of databases, the interpretation of data in a new context depends both on the available information about their production and on how such information is assessed by the context in which data are re-used. The significance of claims about phenomena is similarly dependent on the theoretical perspectives and interests of the contexts in which they are evaluated.

The epistemic role played by such variability is especially striking in model organism biology, where the applicability of results (whether data or claims) beyond the species or even the organism on which they are obtained needs to be evaluated in every single case of 'travel'. This is a case where it is *desirable* for data and claims about phenomena to have a variable evidential scope and, as a consequence, variable degrees of locality. This variability enables researchers to reconcile two characteristics of this field that are potentially at odds with each other: the need to re-use data across research contexts, and the awareness of how the significance of any one dataset may vary dramatically depending on the species and biological phenomenon to which it is brought to bear.

References

Ankeny, Rachel (2007) "Wormy Logic: Model Organisms as Case-Based Reasoning", in Angela N. H. Creager, Elizabeth Lunbeck and M. Norton Wise (eds.) *Science without Laws: Model Systems, Cases, Exemplary Narratives*. Chapel Hill, NC: Duke University Press, 46-58.

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

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

Hacking, Ian (1992) "The Self-Vindication of the Laboratory Sciences", in: Pickering, A. (ed.) *Science as Practice and Culture*. The University of Chicago Press, 29-64.

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

Krohs, Ulrich, and Werner Callebaut (2007) "Data Without Models Merging with Models Without Data", in Fred C. Boogerd, Frank J Bruggeman, Hans S. Hofmeyr and Hans V. Westerhoff (eds.) *Systems Biology: Philosophical Foundations.* Elsevier, 181-213.

Leonelli, Sabina (2008a) *Circulating Evidence Across Research Contexts: The Locality of Data and Claims in Model Organism Research*. Working Papers on The Nature of Evidence: How Well Do 'Facts' Travel?, Number 25/08, Department of Economic History, London: London School of Economics.
http://www.lse.ac.uk/collections/economicHistory/pdf/FACTSPDF/2508Leonelli.pdf

Leonelli, Sabina (2008b) “Bio-Ontologies as Tools for Integration in Biology”, *Biological Theory* 3, 1: 8-11.

McAllister, James W. (1997) “Phenomena and Patterns in Data Sets”, *Erkentniss,* 47: 217-228.

Rhee, Sue Y., Julie Dickerson and Dong Xu (2006) “Bioinformatics and Its Applications in Plant Biology”, *Annu. Rev. Plant Biol.,* 57: 335-360.