Paper forthcoming in *Perspective on Science* (scheduled special issue edited by Henk de Regt and Wendy Parker, Spring 2014)

Sabina Leonelli, ESRC Centre for Genomics in Society, Department of Sociology and Philosophy, University of Exeter, St Germans Road, EX4 4PJ Exeter, UK, s.leonelli@exeter.ac.uk

Data Interpretation in the Digital Age

Abstract

The consultation of internet databases and the related use of computer software to retrieve, visualise and model data have become key components of many areas of scientific research. This paper focuses on the relation of these developments to understanding the biology of organisms, and examines the conditions under which the evidential value of data posted online is assessed and interpreted by the researchers who access them, in ways that underpin and guide the use of those data to foster discovery. I consider the types of knowledge required to interpret data as evidence for claims about organisms, and in particular the relevance of knowledge acquired through physical interaction with actual organisms to assessing the evidential value of data found online. I conclude that familiarity with research *in vivo* is crucial to assessing the quality and significance of data visualised *in silico*; and that studying how biological data are disseminated, visualised, assessed and interpreted in the digital age provides a strong rationale for viewing scientific understanding as a social and distributed, rather than individual and localised, achievement.

Keywords: data, computer, organisms, experimentation, understanding, automation, databases.

Introduction

Scientific knowledge production is currently affected by the dissemination of data on an unprecedented scale. Technologies for the automated production and sharing of vast amounts of data have changed the way in which data are handled and interpreted in several scientific domains, most notably molecular biology and biomedicine. In these fields, the activity of data gathering has become increasingly technology-driven, with machines such as next generation genome sequencers and mass spectrometers generating billions of data points within hours, and with little need for human supervision. Given the relative ease and low costs with which datasets can be produced (that is, once a laboratory has been able to afford these expensive machines in the first place), researchers often end up generating extremely large datasets in case any pattern of relevance to their investigations might emerge. At the same time, there is increasing pressure to make the data thus collected widely and freely available, and integrate them with other types of data, ranging from field observations to data produced through experimental research.[1] Data are seen as a resource of potential interest to all scientists

[1] For an overview of the challenges and opportunities afforded by the so-called ‘data deluge’, see the special issue of *Science* on ‘Dealing with Data’, volume 331, issue 6018, February 2011; and the special issue of Studies in the History and the Philosophy of the Biological and Biomedical Sciences on ‘Data-Driven Research in the Biological and Biomedical Sciences’, volume 43, issue 1, 2012.

working on the same phenomena, the result of large investments which need to be put to good use. As a consequence, scientific institutions and funding bodies have stepped up efforts to improve what Geoff Bowker has called 'memory practices' – techniques and technologies geared towards storing and retrieving facts (Bowker 2006).

Digital technologies such as online databases are widely believed to constitute the best available solution to the logistics of storing, disseminating, retrieving and analyzing data (Hey et al 2009). It is not hard to see why this should be the case: data sharing through the internet can happen in a matter of seconds, and software and hardware to make data travel online are becoming increasingly easier and cheaper to set up. Many research efforts are thus being devoted to the dissemination, modelling and visualisation of data online, in the hope that free and well-managed access to large datasets will enable scientists to use them to understand phenomena, thus generating new paths towards discovery. Some scientists go as far as claiming that the introduction of computational tools for data handling, such as databases and other digital infrastructures, heralds a new methodological paradigm in science, often referred to as data-intensive, or even data-driven, research (Kell and Oliver 2004; Hey et al, 2009).

What interests me here is one crucial assumption underlying these kinds of claims: the idea that computing and digital technologies for data handling are making it possible to automate not only the production and dissemination of data, but also their interpretation (see for instance Allen 2001). This paper aims to explore the philosophical significance and the practical feasibility of this idea by addressing the epistemic relation between research carried out *in vivo* and practices of data dissemination, visualisation and analysis through online databases. I consider the conditions under which researchers assess and interpret the scientific significance of data posted online: in other words, how scientists come to understand what those data tell them about the natural world. Such an understanding arguably underpins and guides the subsequent use of data to foster new discoveries; it is hard to imagine how data could be used as evidence for a claim, or as a reason to set up a research project, in the absence of intuitions about what those data tell scientists about specific entities or processes. Nevertheless, little philosophical reflection has addressed the problem of what gives meaning to data available online – what makes it possible for scientists to interpret them and assess their evidential value, so as to be able to use them to improve their understanding of phenomena. As a starting point to tackle this question, I shall discuss the very notion of scientific understanding and its relation to processes of data interpretation. I will then consider the idea of automated reasoning which underlies many of the claims made about the epistemic power of research carried out through digital databases and computational modelling. As I will show, this idea becomes problematic when one considers the amount of curation involved in making data available online and keeping databases useable in the light of new research advancements; and it is further undermined by the importance of embodied knowledge in assessing the evidential value of data disseminated in this way. Consideration of these two sets of problems with automated reasoning will lead me to formulate the following central claim: the ability to assess and interpret the evidential value of data, and thus to use them to generate new scientific understanding, is tied to familiarity with the target system(s) that data are taken to document, as well as with the conditions under which that system is studied *in vivo*.

In closing, I will consider two consequences of this view. First, I shall contest the idea that scientific discovery through the analysis of large datasets can ever be *fully* automated. My discussion of how data posted online help to understand the biology of organisms shows the complementarity between research conducted *in silico* and *in vivo*: no matter how accurately and efficiently databases help scientists to search and visualise datasets, the interpretation of data as evidence for a claim about a target system is intimately tied to knowledge about that system that can hardly be formalised within a computer system. Secondly, I shall point out that this does not mean that every researcher involved in the dissemination, curation and analysis of data *in silico* needs to be familiar with the organisms data are used to understand. The process of interpreting data is carried out collectively by a large group of scientists, often including bioinformaticians, experimental biologists, field biologists, computer scientists and even engineers, who may not even know each other, but who all contribute to building the infrastructures, experimental set-ups and research materials used to generate, disseminate, visualise and interpret data. Understanding organisms is a social achievement, obtained through the distribution and localised integration of specific cognitive abilities and types of knowledge. Recognising the distributed nature of scientific understanding as fostered by digital tools for data handling is an important step towards identifying what is new and exciting about data-intensive research in the digital age, and constitutes a more promising avenue for analysis than the emphasis on automation favoured by some commentators.

1. Interpreting data to understand phenomena

I shall start my analysis with a discussion of the relation between scientific understanding and processes of data interpretation, with an emphasis on how the very notion of data need to be conceptualised in order to make sense of data-intensive science in the digital age.

I define data as mobile pieces of information, which are collected, stored and disseminated in order to be used as evidence for claims about specific processes or entities. Thus any material product of research activities can be considered as a data point as long as (1) it is taken to constitute potential evidence for a range of phenomena, and (2) it is possible to circulate it across a community of scientists. These artefacts can be passed around in the form of pictures, graphs or numbers, and they can be manipulated to various degrees in order to create visualisations of large datasets. The opportunity to be 'passed on' that these objects afford, by virtue of their materiality, makes them ideally suited for travel from context to context. Scientists can share, exchange, donate datasets; data can be posted online and retrieved unchanged by whoever wishes to access them. This fact does not challenge the well-known philosophical contention that there are no such things as 'raw data' or 'theory-free' representations of data. There is no doubt that the visualisation and subsequent use of data is affected by choices made during their production - decisions about how to set up experiments or observations, which instruments to use and how to calibrate them, which data formats to adopt and which tools to use for collection, storage and dissemination (Bogen and Woodward 1988, Gooding 1990, Radder 2003 and 2009). However, the new ways in which data can be disseminated and retrieved do affect their epistemic value as evidence for claims about phenomena. As I have argued elsewhere, the evidential value of data is quintessentially dependent not only on how they have been produced, but also on the context in which they are adopted and used (Leonelli 2009a). Data can be re-used as evidence for several claims. Indeed, it is the possibility to provide different interpretations of their significance, all of which might turn out to be valid, that grounds and motivates the very idea of data-intensive science and the extensive computational resources allocated to it.

Another way to put this is to say that the dissemination and use of data across different scientific contexts is viewed as helpful to foster current scientific understandings of a variety of phenomena. Data are not viewed as part of one unique process of discovery, from which they are inextricable. Rather, data are viewed as potential components of more than one line of inquiry: the fact that they can be interpreted in different ways means that they contribute to the understanding of more than just one target system. The notion of understanding employed here is one that focuses on the processes through which an understanding of phenomena is achieved, rather than on a strict a priori definition of what counts as understanding. The only assumptions I make are that understanding phenomena is not the same as learning a set of claims about phenomena which account for some of their features (claims typically described by philosophers as propositional knowledge, or explanation). There is a difference between understanding a phenomenon and possessing an explanation of it, since it is perfectly possibly to have access to one such explanation without being able to understand it (see also de Regt 2009). My account focuses on the idea of understanding as a cognitive ability, which scientists acquire through three main types of experience: *intellectual*, involving reasoning through and developing concepts, theories and explanations for natural phenomena; *material*, involving learning and practicing ways of intervening in the world and particularly with the phenomena of interest; and *social*, involving learning and practicing how to contribute to one or more scientific communities with their own specific norms, goals, ways of thinking and ways of doing. These different types of experiences occasion scientists to pick up and exercise specific *epistemic skills* (the abilities to act in ways that are recognised by the relevant epistemic community as well suited to understanding a given phenomenon, e.g. when manipulating a mathematical model or calibrating a measuring instrument) and *research commitments* (bits of knowledge that are used as platforms for carrying out research, along the lines of the core of a research programme as envisaged by Lakatos). So, for instance, the ability to understand specific features of an organism will depend on the acquisition of background knowledge sanctioned as trustworthy by the biological community as well as expertise in handling the instruments, models, specimens and theories used to investigate and explain those features.[2]

The case of data is particularly interesting to consider when thinking about the conditions under which understanding is acquired, not least because few philosophers have yet done so (models and theory being the focus of the vast majority of current philosophical discussions of scientific understanding). I will argue that the ways in which scientists, and particularly biologists, assess the evidential value of data - and thus use them to develop and corroborate claims about phenomena - is intimately tied to skills and commitments formed through interactions with actual organisms, which in turn are extremely difficult to formalise and standardise so as to incorporate them into computational processes. Interpreting data to understand phenomena thus involves the iteration of computational analysis and decision-making processes grounded in the skills and commitments acquired by researchers through physical interaction with the systems they are attempting to understand. In what follows, I will thus focus on the material experiences that make it possible to interpret scientific data, and show how paying attention to these experiences as sources of knowledge, and to how they inform processes of data dissemination, clarifies how scientific understanding is obtained and socialised in contemporary data-intensive science. To that aim, I will now turn to what scientists actually mean by data-intensive science, and how advances in the computational analysis and online dissemination of data are fuelling visions of increasingly machine-driven discovery.

2. Dreams of automated reasoning

[2] For a detailed defence of this view on scientific understanding, see Leonelli (2009b).

‘Data-driven discovery’ is the idea that computer software can be assigned a prominent role in facilitating the extraction of scientifically meaningful patterns from data, either through statistical analysis or through search mechanisms in databases (such as, in the simplest cases, the use of keywords and related algorithms to retrieve data of interest to database users). The extent to which scientists are pushing the use of computers to interpret data is particularly evident within the life sciences, where the complexity of the entities under investigation and the related emphasis towards integrating diverse kinds of data (as in, for example, systems biology) are challenging biologists to find ever more sophisticated tools for data analysis (Stein 2008). Examples of databases used to this aim are ‘community databases’ in model organism research, which bring together information about several aspects of a specific model organism (such as The Arabidopsis Information Resource, which collects sequence, metabolic, physiological, morphological and expression data on the model plant *Arabidopsis thaliana*; Huala et al 2001); and ‘grids’ or ‘portals’ in biomedical research (such as the Cancer Biomedical Informatics Grid, which provides access to all sorts of data available on several types of cancer; Eschenbach and Buetow 2006).[3]

The overarching vision that drives at least some of these attempts is the pursuit of full automation in scientific inquiry, or, in the words of a recent and controversial commentary, ‘machine science’ (Evans and Rezhetsky 2010): the progressive elimination of human intervention (and thus, manual labour and subjective decision-making) from data analysis, resulting in the automatic generation of scientific hypotheses, findings and, ultimately, new discoveries. This vision of scientific research aims to make the involvement of humans in the selection, evaluation and interpretation of experimental data as limited as possible. The automation of reasoning processes has been the Holy Grail of computational science and AI

[3] For a detailed philosophical and historical analysis of how community databases have affected research on model organisms in biology, see Leonelli and Ankeny (2012).

since several decades, and the natural sciences constitute an ideal test case to probe the extent to which the limits and costs of human reasoning and intervention can be overcome through reliance on machines (King et al 2009).

Prima facie, automated reasoning seems to be growing increasingly plausible, with several examples of inference methods being successfully implemented to extract patterns from data. So-called 'random walks' through datasets, for instance, are algorithms devised to spot gene expression patterns from randomly assembled gene expression data (e.g. Noirel 2009). Another example is the use of robots to generate and test hypotheses by sifting through existing data – an approach used with remarkable success in the case of sequence data, as in the case of yeast (King et al 2009). Yet another form of automated analysis is carried out through the implementation of retrieval mechanisms to search databases in the first place: bio-ontologies such as the Gene Ontology are developed to structure the information contained within databases and the ways in which data are classified, visualized and modeled, so that scientists accessing these tools can gain access to useful data as quickly and efficiently as possible. These methods provide important insights for the investigation of the biology of organisms, for instance by enabling the integrated analysis of datasets that could not otherwise have been put in relation to each other (O'Malley and Soyer 2012; Leonelli forthcoming); and by directing researchers to specific research topics, questions and directions for future exploration (Wimsatt 2007; Krohs and Callebaut 2007).

Whether these methods can be regarded as promising substitutes for non-computational forms of inquiry, however, is disputable; and this paper aims to show how automated forms of reasoning, and particularly inference of meaningful patterns from data, cannot be expected to replace localised, physical interactions between human researchers and the target system(s) under investigation. In order to show the limits of automated reasoning, I shall focus on the role of the material experiences of scientists – which I shall refer to as 'embodied knowledge' - in assessing the evidential value of data found online, and the difficulties hitherto encountered in formalising this type of knowledge.

3. The reality of data curation

When considering how tools such as digital databases and computational models are developed and maintained, it becomes apparent that making and keeping data tractable for automated analysis requires considerable manual and conceptual labour. Data disseminated through digital databases, and thus made amenable to computational visualisation and modelling, need to be 'curated' by professionals whose expertise lies in making those data accessible to computational tools (Blake and Bult 2005, Buetow 2005). Professional curators play a crucial role in making it possible for researchers to retrieve and analyse data found on digital databases. Curation involves several complex tasks, including the selection of data to be assimilated into a database; their formatting into a standard that can be digitally tractable by the available software; their classification into retrievable categories, which makes it possible to 'mine' the data according to whichever biological question is asked; and their visualisation through modelling tools that display data in ways that make it possible to spot meaningful patterns (Howe et al 2008; Leonelli 2010). Further effort is put into ranking (or, in curators' own terms, 'cleaning') data as preparation for automated analysis; and in selecting information that is to accompany data on their digital journeys. Data cleaning alone is estimated to take 80% of the total time invested in preparing data for mining (Boerner 2010). Moreover, it is also important to stress the tight relation between data mining techniques and data visualisation tools. The latter are widely acknowledged as crucial ways to 'transform data into information', where information is interpreted as meaningful insight about specific phenomena (Fry 2008: 2). Data mining always involves devising ways to visualise and display results. Mining data for patterns can thus be understood as an exercise in visualising data, i.e. in finding ways to display the results of a search on a database performed with the help of data retrieval mechanisms. Visualisations are seen as 'revealing' patterns that would not be spotted unless data are adequately displayed. Visualisations also offer a potential solution to the problems posed by the quantities of data to be analysed, since a typical genomic experiment involves one million data points (Hey et al 2009). The extent to which data mining, visualisation and interpretation are intertwined points to the degree of responsibility that curators bear in setting up data for re-use, since it is clear that whoever chooses what counts as an adequate visualisation has a strong impact on how data will be interpreted. Curatorial processes form an integral part of the process of scientific inquiry through which data are analysed and interpreted. Indeed, experimental scientists are becoming increasingly aware of the significance of mining and visualising tools in affecting how data are eventually interpreted (Mariscal et al 2010). It has also been noted that the format given to data when they are processed for dissemination tends to drive the types of analysis that are then carried out - and thus the type of results obtained (Fry 2008).

The above glimpses into the work carried out by data curators, and its potential impact on how those data will be interpreted in the long term, suggest that their efforts are difficult to fully automate. When selecting data formats, visualisation tools, modelling techniques and classification systems for data, curators are making choices that partly determine the significance that those data can have when 'automatically' mined. These choices are informed by the curators' own knowledge of scientific research and by their assessment of the potential usefulness of specific datasets towards new insights.[4] Perhaps most importantly for my purposes here, data curation does not stop once curators have cleaned data, formatted them and developed mechanisms to retrieve them and visualise them. Curators need to maintain the databases that they have put online, and make sure that the choices made when first classifying and visualising them remain valid in the face of ever-shifting scientific developments. This process of updating requires difficult conceptual and practical decisions about how new discoveries are affecting structures, classifications and models previously set up to disseminate and visualise data. [5] The highly qualitative nature of these decisions, and their essentially unpredictable nature (by definition, they are responses to unforeseeable developments), make curatorial processes hard to formalise and automate as desired by supporters of 'machine science'.

4. Assessing the quality and reliability of data found online

The importance of curatorial decisions becomes even more relevant when we consider the knowledge that users of databases need to have in order to interpret the data that they find there. I shall thus turn to the ways in which scientists assess whether the data that they retrieve online are reliable; and how they decide to accept data as evidence for claims that those data were not originally produced to test. I will also consider the tools that curators provide to data users, in order to help them to assess the evidential value of data found online.

When finding potentially interesting data online, one of the first and most important questions that researchers need to ask concerns the reliability of those data. Can the data be trusted?

[4] For a more detailed discussion of the expertise of curators, see for instance Leonelli 2010 and 2012.

[5] For a study of how a specific database, the Gene Ontology, has evolved over time to take account of shifting biological knowledge, see Leonelli et al (2011).

Are they of good quality? Answering these questions usually involves evaluating the adequacy of the experimental conditions under which data have been produced (Bogen and Woodward 1988). How is this done when the researchers who use the data have not been personally involved in conducting those experiments? One solution to this problem consists of 'confidence rankings' set up by database curators. This constitutes an attempt to replace biologists' individual evaluation of the quality and reliability of datasets with standard rankings of data quality, indicating the degree of trust with which scientists should approach each dataset on a database. In particular, confidence rankings classify evidence as more or less reliable depending on the methods through which they were produced. For instance, data produced through knock-out experiments (which are seen as providing results that do not crucially depend on environmental conditions) tend to be ranked as more reliable than data produced through micro-array experiments (which are often critiqued as being highly susceptible to changes in environmental conditions; Rogers and Cambrosio 2007). A good example of confidence ranking are the 'evidence codes' used within The Arabidopsis Information Resource, according to which data 'inferred by direct assay' (IDA) are ranked higher than data 'inferred by electronic annotation' or computational prediction (IEA), because IEA has not been experimentally verified (Swarbreck 2008).

Reliance on confidence rankings involves delegating an important aspect of the evaluation of the evidential significance of data to curators. When constructing these rankings, curators are in charge of assessing the reliability of evidence and data-generating procedures in the first place. This is potentially problematic, as curators may bring their own biases and limited expertise to these classifications (particularly as they are often not familiar with many of the materials, including organisms, from which data are generated), thus generating a hierarchy of types of evidence that may not be dependable.[6] Even more questionable from a philosophical viewpoint is the idea of determining the quality of evidence through a ranking of its sources, i.e. the instruments and techniques used to produce data. As we have learned from the philosophy of experiment (e.g. Gooding 1990 and Radder 1993, 2009), experimental instruments - and well as of course other methods for data generation, such as field observations and specimen collection - do not have intrinsic reliability. The same experiment can be more or less reliable depending on the goals of the investigation at hand, the training and abilities of the scientists involved, and the circumstances and settings in which it is conducted. The tacit knowledge and specific conditions under which an experiment is carried out often determine the quality and reliability of its results, and cannot be fully captured by a mere description of the instruments and protocols used. The classification of experiments and other forms of data production by type, as proposed within confidence rankings, is therefore a dubious indicator of the quality of the data obtained.

Curators are aware of these difficulties and are trying to overcome them by giving database users the opportunity to consult 'meta-data', which consist of detailed information about the provenance of data – how they were obtained, where, on which materials, through which instruments, following which protocols and which research goals (see for instance Taylor et al 2008). The idea underlying the use of meta-data is that information about data provenance can be interpreted differently by each scientist interested in a specific dataset, depending on her own research experience. In other words, access to meta-data gives researchers the opportunity to assess the quality of data through the lenses of their own knowledge of their field and familiarity with (and opinion of) specific experimental set-ups. For instance, they might assess a dataset as reliable because they trust the instrument or laboratory or group that produced it, or because they see the materials on which data were obtained (a specific type of

[6] On this point, see also Cartwright's (2007) critique of evidence rankings in evidence-based medicine.

tissue from a standardised mutant specimen, for instance) as comparable with the materials used in their own research.

By appealing to researchers' personal experience of what counts as good data, the use of meta-data might come close to answering scientists' need to judge for themselves whether the data they use are reliable or not. In this way, the consultation of meta-data makes it possible for a scientist accessing data *in silico* to form her own opinion on their quality and reliability. Difficulties, however, abound also in this case, mostly due to the lack of standard terminology to describe data-collecting conditions across research contexts. How experimental practices are described, for instance, might be unintelligible to researchers coming from a research context other than the one in which data were originally produced (e.g. when shifting from biological to clinical research on human tissues). The ways in which instruments are calibrated and maintained might also differ; or the time-scale over which the measurement is carried out might vary (an important parameter when extracting data from living, and developing, organisms). Perhaps unsurprisingly, capturing processes of data production through descriptive and standardised tools such as evidence codes and meta-data constitutes a remarkable challenge.

Curators have several ways to cope with this challenge, often involving consultation with data users to determine which elements of an experimental system are most valuable in order to assess the quality of the result produced (Leonelli 2010). Still, curators are ultimately responsible for assembling information acquired through dialogue with researchers, and translating it into an adequate system for the classification of data and its provenance. And as noted before, this kind of work requires the constant updating and re-gearing of the classification systems in place to follow developments in scientific knowledge and practices, which is extremely difficult to automate.

5. Assessing the evidential value of data found online

Meta-data play an important role also when trying to establish the evidential value of data found online towards a specific claim about phenomena. This is because determining the evidential value of data requires knowledge of the organism in question and of the instruments used to explore it. Paraphrasing Evelyn Fox Keller (1983), it requires 'a feeling for' the material conditions under which phenomena are investigated.

Keller's expression has been critiqued for its lack of precision, a terminological vagueness that plagues arguments pointing to the importance of embodied, non-propositional knowledge in scientific research. Gilbert Ryle famously called this kind of knowledge 'knowing how', thus distinguishing the knowledge needed to carry out scientific research from the propositional knowledge used to devise experiments and interpret results ('knowing that'; Ryle 1949). Other philosophers, most famously Michael Polanyi, emphasised the 'tacit' nature of such knowledge, thus dismissing the very possibility that embodied knowledge could be articulated (Polanyi 1967). Meta-data become very interesting from the epistemological viewpoint when considered through the lenses of this philosophical literature, because they constitute an explicit attempt to articulate and formalise the embodied dimensions of scientific knowledge, and particularly the material experience of researchers involved in data production. By supplying as much information as possible about how data are produced, meta-data become a tool to express the 'knowing how' involved in the generation and use of data, thus demonstrating the extent to which such knowledge can be reported and assessed.

This does not mean challenging the idea that experimentation is a hugely localised, situated affair, each instance of which brings together a vast variety of skills, assumptions, materials, environmental conditions and goals. Both the curators who develop meta-data and the researchers who use them recognise that each method of and setting for data production has its own idiosyncrasies. Indeed, the selection of meta-data starts from the idea that each scientific inquiry is unique, due to extreme complexity of the parameters involved. At the same time, the selection of meta-data can be interpreted as involving three key assumptions about what researchers need to know about data in order to interpret their significance:

1. The belief that some of the characteristics of each experimental setting matter more than others when it comes to assess the quality and significance of the results obtained. For instance, it is impossible for any biologist to determine the evidential value of a given dataset, in the absence of information about what materials that dataset was taken from (which model organism and, if known, which specific mutant strain, including information about its phenotype and genotype).[7]

2. The belief that those characteristics can be singled out as pertaining to the same 'type' across several experimental settings. Any researcher wishing to interpret biological data found online will need information about what organism they were taken from, with which instrument(s), who carried out the experiment, where and when. So meta-data will need to include categories such as 'organism', 'instruments', 'authors', 'location of original experiment' and 'time of original experiment'.

3. The belief that at least some of these characteristics can be explicitly described through texts, graphs or other media. Textual descriptions can be useful in expressing

[7] Such information, which may seem trivial to assemble and record in a database, is actually hard to generate and document, as I discuss in the case of biomedical databases in Leonelli (2012).

> at least some aspects of embodied knowledge, as in the case of experimental protocols ('the pipette needs to be carefully inserted into the probe, so as not to shake the liquid inside'). Yet, as stressed by several authors within the 'tacit knowledge' tradition, propositions fare very poorly in capturing researchers' skills (how well a researcher can splice genes or photograph embryos) or the feelings and familiarity held by a researcher for an organism (how well a researcher knows a strain of mutant mice, thus allowing her to spot when their behaviour deviates from the norm). A partial remedy to this, increasingly used by database curators, is to capture aspects of embodied knowledge through graphs (for instance, when plant researchers illustrate how to intervene on a plant to keep it from dying) or video recording. The *Journal of Visualised Experiments* is one of the several online initiatives devoted to recording and distributing meta-data by filming whole experiments.

The development of meta-data on the basis of these assumptions does not challenge the idea that embodied knowledge relevant to the interpretation of data can only be obtained through actual, physical interaction with the target system(s) under investigation – in our case, with actual organisms. However, what meta-data seem to foster is a new level of reflexivity and communication across experimenters coming from different traditions. The principle underlying the use of meta-data is that the characteristics of the embodied knowledge involved in data production can and should be articulated, so that scientists not directly involved in that process can still form an opinion about how it was carried out. Researchers who do not share theoretical commitments and goals, but who do share a minimal amount of skills in laboratory practices, can thus form an opinion about the evidential value and scientific significance of each other's results.[8] Viewed in this way, meta-data aim to capture

[8] This process is also helped by the use of common standards, instruments and infrastructures, as shown by Rogers and Cambrosio (2007) in the case of micro-array data. The facts that a single company (Affimetrix)

and express the conditions under which researchers can interpret any dataset, no matter its origin, towards understanding phenomena of interest to them.

This is an important finding for philosophical research on the role of embodied knowledge in science. At the very least, it makes it clear how important this type of knowledge is to gaining scientific understanding of phenomena. The idea of carrying out research entirely *in silico*, through automated analysis and without complementing it with interactions with actual organisms, becomes untenable given this insight. Interaction with organisms *in vivo* is not important only as a validation of the results found online; it is required in order to be able to interpret the evidential value of data collected on those organisms in the first place.

At the same time, this reading of meta-data makes it possible to think of embodied knowledge as something which is not necessarily restricted to the boundaries of one specific research setting and to the experience of one individual. As I argued, a researcher's existing familiarity with experimental techniques, materials and instruments (his/her existing commitments and skills) is crucial to being able to assess the evidential value of data. This does not prevent such knowledge, once it exists, from being discussed and articulated. Differences between the skills and commitments favoured within different experimental settings can be identified and evaluated; and, most importantly, people with different training and experimental background can, through tools such as metadata, form opinions on each other's practices and use those opinions to interpret data found *in silico*. Of course, reliance on meta-data presupposes trust in the ways in which each researcher describes her own experimental practices. Data users have no means nor time to verify those descriptions – the information provided through meta-data needs to be accepted as correct, unless specific

acquired a monopoly on the technology used to store results on chips, and that minimal standards to describe a microarray experiment were introduced, have greatly enhanced the opportunities for scientists to exchange and interpret data coming from different labs.

reasons for doubt emerge (such as a public indictment for fraud, as in the recent case of Korean stem cell researcher Huang Woo-suk; Hong 2008). This however has arguably been the case throughout the history of science; a degree of trust in other researchers' accuracy and data has always been a requirement for the advancement of scientific knowledge, and seemed to have been largely vindicated overall.

A crucial insight emerging from this analysis is that the embodied knowledge, skills and commitments necessary to interpret data does not need to be harboured by each and every individual scientist involved in the complex process of producing, disseminating, assessing and interpreting data. As I illustrated, curators play an important role in making communication about embodied knowledge possible on the scale required by current data-intensive research. Researchers that are not involved in material interactions with organisms, such as modellers, computer scientists and statisticians, are also often involved in setting up digital databases and computational tools to retrieve, visualise and analyse data within them. Assessing the evidential value of data in relation to specific claims about organisms is greatly helped by reliance on the experience, expertise, skills and commitments of these scientists, many of whom will never have come into contact with the organisms that data are used to investigate. At the same time, the experience, expertise, skills and commitments of researchers who do interact with these organisms are indispensable to actually using data retrieved through databases in order to understand biological features.

Conclusion: Data analysis in the digital age, and the distributed nature of scientific understanding

I have shown that to assess the reliability, quality, relevance and significance of data found online, researchers greatly benefit from computational tools and standards such as meta-data, but also require embodied knowledge derived from material interactions with the systems under investigation. In order to yield fruitful scientific insights, the online consultation of databases needs to be embedded in a wider spectrum of scientific practices, particularly ones that enable researchers to understand what a specific dataset might signal for the purposes of their own investigations. Especially in the case of biological research on organisms, this means that at least some of the curators and users of databases need to be versed in some form of physical interaction with organisms, thus bringing the skills and commitments gained through those interactions to bear when classifying, assessing and re-using data on those organisms.

In closing, I want to discuss what I take to be two important implications of this line of argument. One is the claim that the full automation of reasoning about data, leading to data-driven discovery and 'machine science', remains highly unlikely despite the impressive advancement in data handling and computational technologies of recent years. This is due to the constant dialectic, continuously generating new insights and new standards for what counts as scientific understanding, between propositional and embodied knowledge ('knowing that' and 'knowing how'). Consider my analysis of the role of curators in processing data and making them available for online retrieval, for instance by developing meta-data; and the role of researchers in assessing meta-data to determine the evidential value of the data found online. The quality, reliability and accuracy of meta-data will certainly improve over years to come, but these developments are unlikely to lead to a completely automated set-up and updating of databases or to a fully automated interpretation of the data found therein. Let me stress again that I am not disputing the importance of e-science tools and automation in occasioning a methodological and epistemic shift in research practices. Rather, I am emphasising the crucial role of human expertise, and particularly of researchers' familiarity with methods of data production, in interpreting data found online. The possibility to provide multiple interpretations of the same dataset, so highly valued within contemporary science, is due to efforts to contextualise data within several different research situations. Data interpretation is at least partly a matter of understanding the circumstances in which data have been produced – and yet, there is no single (or, arguably, even a 'best') interpretation at stake. Depending on their research context and degree of familiarity with specific methods of data production, scientists may interpret the same data – the same travelling object – in different ways.

Far from detracting from the revolutionary power of digital technologies for data analysis, I take my argument to shed light on what is actually new and exciting about data-intensive science: that is, the ways in which digital technology is fostering new forms of collaboration and division of labour within the sciences. Researchers make important choices at all stages of data analysis. Individuals are called upon to decide how to set up experiments and calibrate instruments that produce the data in the first place; how data should be formatted, mined and visualised; how data should be interpreted and which evidential value they acquire in different research contexts. What is remarkable about the current situation is that, for any given datasets, several individuals, sometimes hundreds of them, are involved in making those decisions. Thanks to the unifying platform provided by computers and internet access, those individuals are increasingly likely to have little in common: they probably will not know each other, they might have very different expertises and priorities, and they might be working within a variety of epistemic cultures. Most importantly, each of those individuals might possess a different form of embodied knowledge, and thus make use of different skills and commitments when handling data.

In the past, individuals pertaining to such disconnected communities would rarely have crossed path. At least in part because of digital tools and the internet, the division of labour within science is becoming more fluid. The life of data is so long and unpredictable, that there is no way to control who is manipulating data, and how, as data journey across laboratories all around the globe. Data users need to trust data curators to have made the right choices when implementing tools for data visualisation. Data users also need to trust data producers to have accurately described their experimental context and the instruments and materials used to obtain their results. As I have shown, mechanisms are in place to enable users to check for themselves the quality and reliability of data posted online – but while these tools are crucial to the interpretation of those data, users still need to trust producers and curators in their descriptions of their decision-making processes. It is becoming increasingly clear that making sense of large datasets cannot be the task of one individual on his/her own. Rather, interpreting data in order to foster the scientific understanding of organisms is an achievement of a (sometimes very large) group of different individuals with diverse goals - and it is this harmonious mix of diversity and co-operation that makes it possible to extract several insights from the same datasets. This brings me to the second, and possibly most important, implication of the arguments made in this paper: the essentially distributed nature of scientific understanding as a collective cognitive achievement of the many scientists involved in contemporary data-intensive research.[9] Considering how data is disseminated and

[9] This reading of data-intensive science comes close to Ronald Giere's reading of research at CERN as a large distributed cognitive system (Giere 2006, pp.108ff.). We differ, however, in our emphasis. While Giere is interested in exploring the role played by artefacts in extending human cognition, I wish to stress the distributed nature of understanding itself as a cognitive achievement of scientific collectives.

interpreted in the digital age highlights the importance of distributed cognition within 21st century science, fostering an increasingly pluralistic understanding of data and, as a consequence, a richer understanding of the natural world.

Acknowledgments

Warm thanks go to the editors of this special issue, an anonymous referee, Maureen O'Malley and Miklos Redei, for their extremely useful and constructive feedback on an early draft; to Henk de Regt and James McAllister, who encouraged me to write this paper by inviting me to the workshop 'Understanding and the Aims of Science' (Leiden, June 2010); and the participants to the Leiden workshop, the ESF workshop 'Points of Contact Between the Philosophy of Biology and the Philosophy of Physics' (London, December 2010) and seminars at the University of Bologna and McGill, where this paper was presented. This research was funded by the UK Economic and Social Research Council as part of the ESRC Centre for Genomics in Society.

References

Allen, J. F. 2001. "*In silico veritas*. Data-Mining and Automated Discovery: The Truth is in There." *EMBO Reports* 2: 542—544.

Blake, Judith A., and Carol J. Bult. 2005. "Beyond the Data Deluge: Data Integration and Bio-Ontologies." *Journal of Biomedical Informatics* 39, 3: 314-320.

Bogen, James, and JamesWoodward 1988. "Saving the Phenomena." *Philosophical Review* 97, 3: 303–352.

Börner, K. 2010. *Atlas of Science: Visualising What We Know*. The MIT Press.

Bowker, Geoff C. 2006. *Memory Practices in the Sciences*. The MIT Press.

Buetow, Keith H. 2005. "Cyberinfrastructure: Empowering a 'Third Way' in Biomedical Research." *Science* 821-824.

Curry, Andrew. 2011. "Rescue of Old Data Offers Lesson for Particle Physicists." *Science* 331, 6018: 694-695.

von Eschenbach, Andrew C., and Kenneth H. Buetow. 2006. "Cancer Informatics Vision: caBIG." *Cancer Informatics 2*: 22-24.

Evans, James, and Andrey Rzhesky. 2010. "Machine Science." *Science* 329, 5990: 399-400.

Fox Keller, E. 1983. *A Feeling for the Organism*. New York: W.H.Freeman.

Fry, B. 2008. *Visualising Data: Explaining and exploring data with the processing environment*. O'Reilly.

Giere, R. 2006. *Scientific Perspectivism*. Chicago and London: The University of Chicago Press.

Gooding, D. 1990. *Experiment and the Making of Meaning: Human Agency in Scientific. Observation and Experiment.* Dordrecht & Boston: Kluwer.

Hacking, Ian 1992. "The Self-Vindication of the Laboratory Sciences." Pp. 29-64 in *Science as Practice and Culture*. Edited by Andrew Pickering. Chicago: University of Chicago Press.

Hey, Tony, Stewart Tansley, and Kristine Tolle. 2009. *The Fourth Paradigm. Data-Intensive Scientific Discovery.* Redmond, Washington: Microsoft Research. http://research.microsoft.com/en-us/collaboration/fourthparadigm

Hong, Songook. 2008. "The Hwang Scandal that 'Shook the World of Science'" *EASTS* 2(1): 1-7.

Howe, Doug, Seung Yon Rhee, et al 2008. "The Future of Biocuration." *Nature* 455: 48-50.

Huala, Eva et al 2001. "The Arabidopsis Information Resource (TAIR): a Comprehensive Database and Web-Based Information Retrieval, Analysis and Visualisation System for a Model Plant." *Nucleic Acids Research* 29: 102-105.

Kell, Douglas B., and Stephen G. Oliver. 2004. "Here is the Evidence, Now What is the Hypothesis? The Complementary Roles of Inductive and Hypothesis-Driven Science in the Post-Genomic Era." *BioEssays* 26, 1: 99-105.

King, R.D. et al. 2009. "The Automation of Science." *Science* 324: 85-89.

Krohs, Ulrich and Callebaut, Werner. 2007. "Data without Models Merging with Models without Data." Pp.181-213 in *Systems Biology: Philosophical Foundations*. Edited by Fred C. Boogerd, Frank J. Bruggeman, Jan-Hendrik S. Hofmeyr & Hans V. Westerhoff. Amsterdam, Reed-Elsevier.

Leonelli, Sabina 2009a. "On the Locality of Data and Claims About Phenomena." *Philosophy of Science* 76, 5: 737-749.

Leonelli, Sabina 2009b. "The Impure Nature of Biological Knowledge." Pp. 189-209 in *Scientific Understanding: Philosophical Perspectives*. Edited by Henk de Regt, Sabina Leonelli and Kai Eigner. Pittsburgh, PA: Pittsburgh University Press.

Leonelli, Sabina. 2010. "Packaging Data for Re-Use: Databases in Model Organism Biology" Pp.325-348 in *How Well Do Facts Travel? The Dissemination of Reliable Knowledge*. Edited by Peter Howlett and Mary S. Morgan. Cambridge, UK: Cambridge University Press.

Leonelli, Sabina. 2012. "When Humans Are the Exception: Cross-Species Databases at the Interface of Clinical and Biological Research." *Social Studies of Science* 42(2): 214-236.

Leonelli, Sabina (forthcoming) "Integrating Data to Acquire New Knowledge: Three Modes of Integration in Plant Science." *Studies in the History and Philosophy of the Biological and Biomedical Sciences*.

Leonelli, Sabina and Ankeny, Rachel A. 2012. "Re-Thinking Organisms: The Epistemic Impact of Databases on Model Organism Biology." *Studies in the History and Philosophy of the Biological and Biomedical Sciences* 43(1): 29-36.

Leonelli, Sabina, Alexander Diehl, Midori Harris, Karen Christie and Jane Lomax. 2011. "How the Gene Ontology Evolves." *BMC Bioinformatics* 12:325.

Mariscal, Gonzalo, Oscar Marban and Covadonga Fernandez. 2010. "A Survey of Data Mining and Knowledge Discovery Process Models and Methodologies." *The Knowledge Engineering Review* 25: 137-166.

Morgan, Mary 2003. "Experiments Without Material Interventions." Pp. 216-235 in *The Philosophy of Scientific Experimentation*. Edited by Hans Radder. Pittsburgh, PA: Pittsburgh University Press.

O'Malley, Maureen A. and Soyer, Orkun. 2012. "The Roles of Integration in Molecular Systems Biology." *Studies in the History and the Philosophy of the Biological and Biomedical Sciences* 43(1): 58-68.

Parker, Wendy. 2009. "Does Matter Really Matter? Computer Simulations, Experiments and Materiality." *Synthese* 169: 483-496.

Polanyi, M. 1967. *The Tacit Dimension*. London: Routledge.

Radder, Hans (Ed.). 2003. *The Philosophy of Scientific Experimentation*. Pittsburgh University Press.

Radder, Hans. 2009. “The Philosophy of Scientific Experimentation: A Review.” *Automated Experimentation* 1, 1: 2.

De Regt, Henk W. 2009. “Understanding and Scientific Explanation.” Pp.21-42 in Scientific Understanding: Philosophical Perspectives. Edited by Henk de Regt, Sabina Leonelli and Kai Eigner. Pittsburgh, PA: Pittsburgh University Press.

Rheinberger, H. 2010. *An Epistemology of the Concrete*. Duke University Press.

Rogers, Susan and Alberto Cambrosio. 2007. “Making a New Technology Work: The Standardisation and Regulation of Microarrays.” *Yale Journal of Biology and Medicine* 80: 165-178.

Ryle, G. 1949. *The Concept of Mind*. Chicago, Illinois: The Chicago University Press.

Stein, Lincoln D. 2008. “Towards a Cyberinfrastructure for the Biological Sciences: Progress, Visions and Challenges.” *Nature reviews. Genetics* 9: 678-88.

Swarbreck, David et al. 2008. “The Arabidopsis Information Resource (TAIR): Gene Structure and Functional Annotation.” *Nucleic Acid Research* 36: D1009-D1014.

Taylor, Chris F., Field, D., Sansone, S., Aerts, J., Apweiler, R., and Michael Ashburner. 2008. “Promoting Coherent Minimum Reporting Guidelines for Biological and Biomedical Investigations: The MIBBI Project.” *Nature Biotechnology* 26, 8: 889-96.

Wimsatt, William C. 2007. “On Building Reliable Pictures with Unreliable Data: an Evolutionary and Developmental Coda for the New Systems Biology.” Pp.103-120 in

Systems Biology: Philosophical Foundations. Edited by Fred C. Boogerd, Frank J. Bruggeman, Jan-Hendrik S. Hofmeyer, and Hans V. Westerhoff. Amsterdam, Reed-Elsevier.