British Academy Conference – 'Enquiry, Evidence and Facts'

Sabina Leonelli
s.leonelli@lse.ac.uk

Circulating Facts About Organisms: Biological Databases

Data[1] are the smallest, yet the most stubborn of scientific facts. They constitute the empirical backbone of scientific research: once they are 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 that they are taken as evidence for? 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 which consequences does this happen, if at all? This paper tackles these questions by focusing on what happens to biological data after they have been obtained and interpreted within a specific experimental setting: in other words, 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 this process of packaging can help to understand how science involves, but is not limited to, the accumulation of facts about the world.

My philosophical starting point is the view on the relation between data and phenomena espoused by Bogen and Woodward in 1988. The core of their position is that 'we need to distinguish what theories explain (phenomena or facts about phenomena) from what is uncontroversially observable (data)' (1988, 314). Phenomena are the interpreted outcomes of the modeling of, or abstracting from, data: 'phenomena are detected through the use of data, but in most cases are not observable in any interesting sense of that term' (1988, 306). They are the object of scientists' most general claims about what the world is like, as expressed in theories and explanations. Claims about phenomena are used as evidence for these general claims: for instance, when defending the Weinberg-Salam theory (which attempts to unify the electromagnetic and weak forces) on the basis of claims about the behaviour of weak neutral currents. By contrast, data cannot be used as evidence for general explanatory claims; rather, they constitute evidence for claims about phenomena, for instance when measurements taken by electronic particle detectors at CERN serve as data for the existence of weak neutral currents. Data help scientists establish what the world is like, thus fixing an ontology on which they can construct their theories. In Bogen and Woodward's words:

> "with respect to their evidential role, what distinguishes data from phenomena is not that only facts about data may serve as evidence, but rather that facts about data and facts about phenomena differ in what they serve as evidence for (claims about phenomena versus general theories)' (1988, 306)

[1] 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).

While I find Bogen and Woodward's distinction between data and phenomena very useful, I take issue with their characterisation of how data and phenomena are used as evidence by practicing scientists. In their view, data are the result of measurements taken in very specific settings, whose features and interpretation depend on the goals, instruments, expertise and beliefs characterising the context in which they are produced. According to them, data are bound to remain in that context, as only in that context is it possible to assess their value as evidence for the existence and behaviour of phenomena. I agree with Bogen and Woodward that data are produced in a setting characterised by an arguably unique ensemble of methods, instruments, aims and background knowledge. I also share their intuition that the interpretation of data is necessarily bound to a local context, as it rests on the scientists' expertise in handling specific instruments and materials. However, I wish to stress that data are often made to travel across research contexts: that is, *data can and often do become non-local evidence for local claims*. This observation has deep epistemological implications, which lead my analysis towards different conclusions from the ones drawn by Bogen and Woodward.

Contemporary biological research constitutes an excellent case for examining data travels in a data-rich environment. Biology has yielded immense amounts of data in the last three decades. This is especially due to genome sequencing projects, which resulted in the accumulation of billions of datasets about the structure of the DNA sequence of various organisms. Researchers in all areas of biology are now 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 in a specific tissue or in whole embryos, data about genes' position on the chromosomes and their mobility through time, data about morphological effects correlated to 'knocking-out' specific genes, and so forth.

These results are obtained through experimentation on a small group of organisms whose features are particularly tractable (i.e. apt to being investigated through available laboratory techniques). These organisms, including fruit-flies (*Drosophila melanogaster*), worms (*C. elegans*), mouse cress (*Arabidopsis thaliana*) and mice (*Mus Musculus*), are referred to as 'model organisms', because it is assumed that results obtained from them will be applicable to most other species with similar features. Researchers are aware that this assumption is problematic. Cross-species transfers of knowledge are a shot in the dark, as researchers cannot be sure of how species differ from each other unless they perform accurate, case-by-case comparative studies. Indeed, the assumption of the reliability of cross-species inference is a strategic rather than a dogmatic choice. Despite the uncertain representational value of model organisms, the majority of biologists agree that cooperation towards the study of several aspects of one species is a good strategy to advance knowledge, as results acquired on that one species can be used as a starting point for the study of other species. Focusing research efforts on few species enables researchers to integrate data about many different aspects of their biology, so as to obtain a better understanding of organisms as complex wholes.

In the light of this background, it is not surprising that biologists consider the circulation of data across research contexts as the main priority in model organism research. If data obtained by the labs involved are not made accessible to all other labs working on the same organism, there would be no cooperation towards an integrated understanding of the organism ever happen and thus no justification for the shaky assumption of reliable cross-species inference. Especially over the last two decades, the quest for efficient means to share data across model organism communities has become a lively research area in its own right, which is usually referred to as bioinformatics. One of the main objectives in

bioinformatics is to exploit new technologies such as the Internet to construct digital databases that are freely available for consultation (Rhee et al 2006).

The construction of databases to make data travel is no easy feat. My analysis focuses on the recent efforts to develop databases to gather, organise and distribute the immense, heterogeneous mass of available data about model organisms. The first part of the paper illustrates how databases confront the challenge presented to biologists by the accumulation of data on model organisms. In particular, I analyse the phases through which data are made to travel through databases, and the consequences of such travelling for biological research. The second part of the paper uses this case study to critique some of Bogen and Woodward's claims. On the one hand, there are cases in scientific research where *data are non-local entities* shared and used across a wide range of research contexts. On the other hand, there are cases where *phenomena are local entities* (that is, context-dependent *concepts* meant to refer to objects in the world) insofar as their construction and interpretation depend on the presuppositions and expertise characterising the communities that use them.

References

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

Hacking, I. (1992) The Self-Vendication of the Laboratory Sciences. In: Pickering, A. (ed.) *Science as Practice and Culture*. The University of Chicago Press.

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

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