Accepted for publication in the journal Philosophy of Science on 9 October 2017

The Time of Data: Time-Scales of Data Use in the Life Sciences

Sabina Leonelli
Exeter Centre for the Study of the Life Sciences (Egenis) & Department of Sociology, Philosophy and Anthropology, University of Exeter
Byrne House, St Germans Road
EX4 4PJ Exeter, UK
s.leonelli@exeter.ac.uk

Abstract

This paper considers the temporal dimension of data processing and use, and the ways in which it affects the production and interpretation of knowledge claims. I start by distinguishing the time at which data collection, dissemination and analysis occur (Data time, or Dt) from the time in which the phenomena for which data serve as evidence operate (Phenomena time, or Pt). Building on the analysis of two examples of data re-use from modelling and experimental practices in biology, I then argue that Dt affects how researchers (1) select and interpret data as evidence and (2) identify and understand phenomena.

1. Introduction: Data Time, Phenomena Time and the Epistemic Role of Data Processing Efforts

Existing analyses of the epistemic status and role of scientific data have focused on synchronous aspects of research, often without considering how the diverse time-scales characterizing the handling of data affect processes of inference and knowledge generation. In this paper, I analyse the temporality of the data practices required to facilitate data-to-phenomena inferences, and its impact on researchers' inferential reasoning and understanding of the phenomena under study. I argue that concerns around the temporality and historicity of data practices affect any research situation in which data are (re)used at a time and place other than those in which they are generated. This paper thus considers the epistemological concerns and challenges involved in processing data to facilitate their preservation and analysis in the long term; and in identifying the conditions under which data can be kept, shared and analyzed through time, thus enabling researchers to build on past efforts and boost future research.

Many philosophical discussions of the temporality of data and its implications for research revolve around the credibility of the evidential strategies employed by the historical sciences – typically defined as sciences that attempt to reconstruct and explain long lost events and objects (such as extinct organisms, ecosystems, human cultures and climatic conditions), and which therefore contend with scarce, sporadic and partial data sources. The differential survival of evidence through time has been argued to provide relatively poor evidential ground for knowledge claims, making the historical sciences hostage to "lucky finds" in terms of what they can and cannot

investigate and explain.[1] In this paper, I argue that concerns around whether and how data maintain evidential value through time are not restricted to the historical sciences, but are common to any field where data acquired in previous periods can play a significant role as evidence for subsequent research, and/or in situations where investigators spend long periods of time investigating and revisiting the same datasets. These situations occur both in experimental and field-based research, and whether or not the data in question are quantitative or qualitative. Indeed, I shall argue that experimental data are particularly time-sensitive due to the ever-changing nature of the know-how and laboratory conditions under which they are produced, which makes these data difficult to preserve as meaningful and re-usable sources of evidence. This issue is often underestimated by philosophers who emphasize the degree of experimental control exercised by researchers at the moment of producing data, yet disregard the ease with which such control can be lost once the original experimental set-up changes or ceases to exist, or the data are retrieved and examined by researchers working in different laboratory conditions.[2]

As a starting point for analysis, I propose to distinguish between two types of temporalities involved in knowledge production and interpretation: the temporal dimension of data practices used to prepare and manage data so that they can be subjected to inferential reasoning (which I shall refer to as Data time, or Dt) and that of the phenomena under investigation, for which data are meant to serve as evidence (Phenomena time, or Pt).[3] Dt is closely associated to the ways in which researchers

[1] E.g. Sober 1988, 1-2; Currie and Turner (2016) and Currie (in press) provide a useful overview of these arguments.
[2] As is often the case in contemporary data-centric biology (Leonelli 2016).
[3] My position is thus sympathetic to the analysis of the relation between experimental and historical sciences provided by Carol Cleland (2002) and Derek Turner (2004),

manage time in their work, and particularly to the constraints and opportunities posed by the time spent in the production, dissemination and analysis of data. Pt refers instead to the assumptions that researchers make about the temporal features of their research targets, and the ways in which such assumptions condition their understanding of the natural world as well as their investigative strategies.

The distinction builds heavily on James Bogen and James Woodward's seminal work on the material conditions under which researchers make data-to-phenomena inferences (1988), which stand in striking contrast to the constraints that apply to the development of a priori, logical inferences (Woodward 2000). At the same time, my views on the categories of "data" and "phenomena" differ from Bogen and Woodward's in two respects. First, I explicitly endorse a relational understanding of the epistemology of data, according to which data are identified and conceptualised in relation to their function within specific situations of inquiry (Leonelli 2015, 2016).[4] Second, I favour an interpretation of the ontological status of phenomena as human constructs rather than actual features of the world – though, contrary to McAllister's anti-realist interpretation (2010), I view such constructs as highly constrained by the characteristics of processes and entities in the world, and thus reliably capturing aspects of reality as researchers experience it (Massimi 2009, Feest 2011).

---

though their discussion of the role of the temporal asymmetry of underdetermination does not explicitly consider the distinction between Dt and Pt, thus underestimating the relevance of practical issues of data preservation and handling to the warrant available to claims about past and present events.

[4] Recent work by Woodward (2010) indicates affinities with this view, yet neither Bogen nor Woodward have so far devoted much attention to defining what they mean by data.

These premises are salient to my proposed distinction between Dt and Pt. On the one hand, they are consistent with Dt and Pt being intertwined in scientific practice, with both dimensions typically affected by practical considerations such as the resources, materials, institutional frameworks and technologies available to researchers. As James Griesemer and Grant Yamashita (2002) argued in relation to research on biological model systems, "phenomena have no intrinsic time-scale": the temporality that researchers ascribe to phenomena depends at least in part on the circumstances of inquiry – which often include issues of data access and data analysis. In a similar way, the temporality of data is defined largely by the research contexts in which they are used – which often include specific conceptualisations of phenomena. On the other hand, the interdependence of Dt and Pt does not make it any less useful to distinguish them analytically. Focusing specifically on Dt means paying attention to the efforts involved in data generation, processing, dissemination and analysis, and the large variability in the stages – and related time-scales - through which any given dataset is handled and interpreted. This temporal dimension can have a significant impact on how Pt is measured, but it is conceptually separate from Pt: Dt pertains to the realm of inquiry and research processes (the so-called "context of discovery"), rather than to the knowledge derived from such processes.

In what follows, I use the distinction between Dt and Pt to examine two cases from contemporary biological practice in which researchers attempt to re-use data previously collected by others as evidence for novel claims about phenomena.[5] The

[5] These examples have been researched through an analysis of scientific literature and online tools such as databases, as well as interviews with the scientists involved which I carried out in 2014 and 2015, and which helped me to reconstruct the activities and reasoning involved in data processing and analysis. Full transcripts of those

first case involves the construction of models to track and predict the spread of plant pathogens, which is grounded on the retrieval and integration of data from a variety of sources, and is highly dependent on the accuracy with which Dt is preserved and managed. The second case is typical of experimental work on regulatory mechanisms within molecular and cell biology, and concerns the retrieval and comparison of data collected on two species of yeast to study the role of the cell cycle in regulating transcription in humans, potentially resulting in breakthroughs in the understanding of cancer onset and development.

These examples illustrate how the distinction between Dt and Pt helps to highlight two important features of scientific knowledge production. First, the ways in which researchers acknowledge and document Dt affect the extent to which they can successfully preserve data, integrate them with other data, and (re-)use them as evidence for new claims. In other words, knowledge about Dt affects researchers' ability to identify relevant data and assess their reliability and significance as evidence for a given hypothesis. Second, knowledge of Dt affects researchers' understanding of the phenomena for which data are taken to serve as evidence, and thus the content of the knowledge claims derived from data analysis.

2. Data Re-Use Case 1: Modelling the Global Spread of Plant Pathogens

My first example concerns contemporary attempts to track the global distribution and movements of plant pathogens (such as the fungus *Hemileia vastatrix*, responsible for the infamous coffee rust disease, or the various blights severely affecting the cultivation of major crops) over the last century, with the goal of identifying trends

---

conversations which interviewees consented to make available online are available as Leonelli (2017).

that may help to predict crop pathogen spread across the globe and its potential impact on agriculture. A key source of data underpinning such efforts are observational reports on pathogens. These reports are typically collected by field stations and plant clinics located on various sites around the world and later assembled into a unique body of evidence by initiatives such as the Plantwise database of the Centre for Agriculture and Bioscience International (CABI). CABI uses observational reports to produce maps tracking the geographical spread of different pathogens through several decades (e.g. figure 1).

[Figure 1 here]

These efforts are hampered by the lack of consistent observational data documenting pathogen movements across different parts of the world. Records for low-income regions such as sub-Saharan Africa and South America, for instance, are patchy at best, and significant time intervals are missing even for well-monitored countries. Furthermore, the ways in which different field stations assemble, store and disseminate pathogen observation data are highly variable, and not easily amenable to integration into global maps. To remedy this situation, researchers have devised modelling tools to infer plausible pathogen movements from environmental factors. These models build on available knowledge of the conditions under which fungi are likely to produce spores and infect their hosts, such as temperature and the availability of water on the leaf surface of the plants in question (i.e. when it is too dry, too hot or too cold, the spores die). This knowledge enables researchers to infer infection rates from the triangulation of observational data with measurements of air temperature,

which are often available thanks to the ubiquity of meteorological stations, and estimates of the amount of water in the crop canopy.

Such work can then be used to develop models to predict future trends and target measures to stop the spread of harmful pathogens. It is at this point that this example becomes relevant for an investigation of Dt. This attempt to put old observational data to a new use prompted some of the researchers involved to take a closer look at how the observational data from pathogen reports had been compiled and assembled in the first place, and the extent to which they could be reliably aligned with meteorological data. This brought to light two challenges that had not been apparent at the start of the modelling effort, and yet I found to underpin most cases of data re-use in biology, with significant implications for data analysis and subsequent interpretations.

The first challenge lies in *reconstructing Dt for the key dataset* underpinning CABI maps, i.e. pathogen reports. This involves several distinct events, which in some cases are separated by one or two decades, including:

- Dt1: data collection, e.g. the date on which a local farmer brought an affected plant specimen to a plant clinic for pathogen identification;
- Dt2: compilation of observational datasets into a consistent report about pathogen spread in the region of interest;
- Dt3: official publication of the compiled data, e.g. in a journal or a report;
- Dt4: use of publications as sources for national maps of pathogen spread;
- Dt5: incorporation of national maps into a global digital repository or online database, such as that run by CABI;
- Dt6: retrieval of the data from the repository for further analysis.

It turns out that published maps can be unreliable in their temporal location of data and subsequent data processing and interventions, and often conflate Dt2-Dt6 with Dt1. Particularly in the case of data older than ten years, extensive efforts are now required to date Dt1 and disentangle it from other Dts. In the absence of an accurate timeline for data processing, it is hard for researchers to construct reliable predictive models. The lack of certainty around Dt also decreases researchers' ability to quantify under-reporting and thus the extent to which data may be missing for specific areas/periods/pathogens (Bebber 2014).

The second challenge consists in *aligning Dt across the diverse data sources at hand,* including pathogen location reports, water estimates and climate data. Given the amount of processing required to prepare data for use in modelling, Dt is much longer and more complicated for pathogen distribution and water estimates than for air temperature – a difference made more dramatic by the recent introduction of direct temperature measurements via satellite. A consequence of this temporal mismatch is that integrating data sources requires considerable and expert judgement labor, with researchers needing to use their knowledge of the territory and the species in question to adjudicate specific cases. The ways in which researchers choose to temporally align these different datasets affect their characterization of the phenomena of interest (such as pathogen spread) and their temporal dimensions (e.g. the rate of spread) – which go on to affect the predictive ability of the models in which they are assumed.

3. Data Re-Use Case 2: Identifying Conserved Regulatory Mechanisms Across Species

Shifting now from modelling to experimental practices, my second example concerns the use of experimental data to study the regulatory mechanisms at work in the cell cycle, and assessing potential links between defects in protein regulation and the proliferation of tumor cells.[6] To identify and study regulatory pathways that may be conserved in humans, researchers often resort to analyzing data coming from much simpler forms of life, which are more tractable and easier to study. In the case of regulatory pathways involved in the cell cycle, a successful investigative strategy involves the comparative use of data collected from two types of fungi: fission yeast (*Schizosaccharomyces pombe*) and baker's yeast (*Saccharomyces cerevisiae*). I shall now focus on the management of Dt in relation to the collection and dissemination of data from these two organisms, and particularly the role played by databases in making them available for comparative analysis.

S. cerevisiae has long been a favourite model in biology, with a vast repertoire of knowledge, databases and tools available to researchers interested in studying the cell cycle. It thus constitutes an obvious starting point to identify new regulatory functions associated to cell replication. However, *cerevisiae* spends a lot of time in the G1 phase of the cell cycle, which is problematic for researchers interested in investigating the S and G2 phases of the cycle (see figure 2). *S. pombe*, a much simpler form of life with 3 chromosomes to *cerevisiae*'s 16, turns out to work as an ideal complement: its S and G2 phases are longer, enabling researchers to scrutinize their potential regulatory functions, and the shared evolutionary history of the two organisms makes it plausible to expect that regulatory mechanisms found in *pombe* may be conserved in *cerevisiae*. The systematic comparison of data produced in *cerevisiae* and *pombe*

[6] This line of research was made famous by the work of Paul Nurse, Tim Hunt and Leland H. Hartwell, earning them a Nobel prize in 2001.

produced findings that turned out to be conserved in humans, making this approach useful to understanding the emergence of cancer (e.g. Caetano et al 2014).

[insert figure 2]

What makes this investigative strategy possible is the opportunity to retrieve, visualize and compare yeast data through PomBase and SGD, the main databases for *pombe* and *cerevisiae*. Robert de Bruin, a leading scientists involved in this work, describes the strategy as follows:

> "we used PomBase to find whether that [mechanism] was conserved in fission yeast. We could really easily establish it in fission yeast, and then we could go back to budding yeast now we knew exactly what we're looking at, and then found that also in budding yeast. [..] Now my work in my lab is all focused on that. Without going into fission yeast and having it accessible that easily, I would have never gone in, and I would have completely missed it. And people in budding yeast completely missed even that regulation, let alone the mechanism, and the same in mammalian cells, people have been studying that for decades and completely missed it." (transcript PI_8_A)

Without the quick and accurate comparative tools provided by PomBase and SGD, it would have been much harder for researchers to compare data across the two species, potentially hindering the discovery of significant connections between their regulatory systems. As it turns out, making it "really easy" to explore data in this way requires labor-intensive practices of data annotation and curation, which largely determine which datasets are found online, what information about the data is captured and

made available within databases, and how the data are presented and retrieved for inferential reasoning. Database curators pay particular attention to the selection and inclusion of information about the provenance of data, such as the time at which data were produced and further processed. This often means consulting directly with the original data producers, who are not always accurate when publishing their data as part of research articles, and may quickly lose memory and/or interest in these details after the end of their project. This information enhances researchers' ability to assess the quality of the data and the extent to which each source is comparable to others.

Database curators also participate in the development and application of standard labels identifying the phenomena for which data may serve as evidence, a task made difficult by the diversity of terms used by different biological communities to denote the same genes, gene functions and phenotypes (most often in the case of groups working on different species; Leonelli and Ankeny 2012). The developers of both PomBase and SGD are deeply engaged in the construction of classification systems, such as the Gene Ontology and the Fission Yeast Phenotype Ontology, that make it possible for researchers to look for data of potential relevance to the phenomena they are interested in. This work requires regular updates of data formats and labels to reflect shifts in knowledge base and in the technologies used to produce and disseminate data. Much effort is devoted to providing an accurate record of Dt, i.e. a timeline for the ways in which data are manipulated to remain accessible and usable. Such record is indispensable to the comparison of data acquired on different species, especially when – as in our example – researchers are not sure about the relation between the phenomena under investigation in the two types of organisms. In such cases, precise notations about the temporality and provenance of data are crucial to

interpretation: researchers need to know when and why a certain dataset has come to be associated to a given phenomenon, and how such inference may be triangulated with findings coming from complementary approaches (such as the functional and evolutionary data required to establish a mechanism as conserved).

Remarkably, the better database curators accomplish these tasks, the more their work remains invisible to database users and research funders. The epistemic significance of these practices becomes visible whenever a lack of information about Dt affects the ability of database users to interpret the data. Due to the strong collaborative ethos and relatively small size characterizing the community of *pombe* researchers, PomBase curators are highly successful in eliciting accurate and updated information from data producers. This does not work as well in larger research communities, where database curators are confronted with much larger datasets and data producers have little incentives to participate in data curation. It also fails within the increasingly nested landscape of data infrastructures characterising contemporary biomedical research, where information about Dt is easily lost when data are passed from one platform to another – leading to a situation of high uncertainty around Dt, similar to what we encountered in the first example. Indeed, the yeast researchers I interviewed reacted strongly against the suggestion that it may be efficient to integrate all data relating to yeast species in one single database, which in their view may entail significant loss of information about data provenance.

4. The Epistemic Significance of Data Time

In the cases described above, Dt may span several different events over an extended period, ranging from the moment in which data are originally collected to the times at which they are modified to make them widely accessible and re-usable. Ideally, given the significance of knowledge about Dt for data analysis and interpretation, the researchers and curators involved in data processing (including the compilation and visualisation of data in publications, databases, maps and models) should ensure that information about Dt accompanies the relevant data points in all stages of their travels. This is what happened in my second example, where attentive data curation and a well-constructed database enable researchers to easily find information about Dt and use it for data analysis. However, and particularly in situations where Dt is recorded under diverse working conditions by individuals with different skills and goals, Dt can be remarkably difficult to track and retrieve. Our first example illustrates the problems that can emerge when researchers have limited access to the history of the data that they are analysing.

Knowledge or ignorance of Dt can thus affect research processes and outcomes in (at least) two ways. First, it can alter researchers' perception of which datasets are most reliable as sources of evidence, thus *affecting the evidential value attributed to data* in any given inquiry. Without access to accurate information about Dt, researchers may need to reject whole datasets or modify the ways in which they analyse them (e.g. by shifting their evidential weight in relation to other sources, or seeking to triangulate them with other types of findings). This is clear in the case of pathogen spread, where researchers found it problematic to deal with datasets where they could not distinguish between Dt1 and other Dts, thus throwing doubt on the reliability of existing data maps and related modelling tools. Though less overtly, this also happens in the case of

cross-species study of transcriptional regulation, where researchers needing to re-use data available online need to be able to trace when the data were collected, with which technology and by whom. These examples show that preserving data is not enough to facilitate their future re-use. Equally relevant is the preservation of information about the temporality of the interventions through which data have been assembled, circulated and visualized.

Second, knowledge or ignorance of Dt can determine the frame and resolution at which phenomena are studied, thus *affecting researchers' understanding of phenomena* – including their perception of what can be known, and what aspects of a given target system are worth focusing on. Consider for instance the epistemic risk posed by the *integration of data* acquired on various aspects of a phenomenon of interest by a variety of sources, such as for instance when bringing together genomic, metabolic, environmental and physiological data collected by different research teams in order to study gene-environment interactions.[7] Such integration is crucial to providing novel insights into complex processes (particularly in the emerging landscape of "big data" analysis), yet both examples show that it carries the risk of loss of information about Dt, since not all existing information about data is preserved when bringing large and diverse datasets together. This carries significant implications for data analysis and the reliability of subsequent inferential processes. Wrong information about Dt or difficulties in aligning different Dt can result in predictive failure of models or misleading cross-species inference.

[7] See also O'Malley and Soyer 2012, Leonelli 2016 (chapter 6).

5. Conclusion: Time-Scales of Data Processing and the Limits of Experimental Control

I proposed to focus on the temporal dimensions involved in practices of data processing, and distinguish between Dt and Pt to shed light on the extent to which the diverse time-scales of data and phenomena affect processes of inference and knowledge production. Though the analysis of two examples from biological practice, I have emphasised the complex set of conditions required to preserve data and related meta-data in the long term, and argued that data processing and related temporalities are crucial to inferential processes and (re)interpretation. I thus hope to have illustrated how tracing the movements of data through processes of inquiry, and particularly the conditions under which data do or do not function as evidence, can help to foster philosophical understanding of how data processing affects the content of knowledge claims. Data are defined by their temporal characteristics as much as by their spatial and morphological ones, and underestimating the challenges and time-scales involved in data processing can disrupt inferential reasoning and invalidate the use of data as evidence.

In closing, I briefly examine one implication of this argument, which is that experimental control over data *production* is not enough to guarantee good, (re)usable data. This is significant when considering philosophical accounts that portray historical and experimental sciences as exemplifying two opposed epistemic situations: one in which researchers have complete control over the range, type and quantity of data that they can obtain on the phenomena of interest, thus providing strong warrant for inferential claims made on the basis of those data; and one where researchers cannot control what data are available and are thus "at the mercy of the

processes by which time covers her tracks" (Currie in press), which seems to threaten the epistemic reliability of their claims.[8]

Both Currie and Alison Wylie have critiqued facile dismissals of the epistemic reliability and scope of claims made within the historical sciences, by pointing to the variety of evidence and methods that researchers in those fields can use (e.g. modelling, analogies, various new technologies to study the composition of material traces) as well as the importance of triangulation and consilience in warranting inferences in other fields (Wylie 2017; Chapman and Wylie 2016; Currie in press). My analysis in this paper corroborates their arguments through the following observation: not only the pessimism about the epistemic status of the historical sciences is based on lack of recognition of their methodological sophistication in processing data, but it is also linked to an exaggerated optimism with respect to the potential and warrants of experimental methods in contemporary science, and the extent to which they can really guarantee control over phenomena.

The cases of data handling that I analysed above, and particularly my second example, point to the fact that experimental results are difficult to control – not only at the point at which they are produced, but most significantly at the point of dissemination, storage, and re-use. Data can disappear or become unusable very quickly if not properly curated: it only takes a destroyed hard disk, a misleading annotation or a postdoc changing jobs. Worries about differential survival of evidence and informational destruction are thus arguably as alive with contemporary data

[8] A similar distinction is frequently made between hermeneutic and quantitative approaches to data re-use (as championed by the social and natural sciences, respectively), and is convincingly challenged by James McAllister in his contribution to the PSA Symposium where this paper was also presented (McAllister 2018).

collection in the life sciences as they are for historical sciences and observational data therein. And because experimentalists today operate in what are often characterised as ideal conditions (particularly in the case of molecular biology, where they can avail themselves of ready-made samples, high-throughput instruments for data production, high computational power and myriads of modelling tools), they themselves tend to underestimate the challenges involved in processing data and related information, which can cause real trouble with data interpretation and inferential reasoning down the line. Both scientists and philosophers can learn from the strategies elaborated within the historical sciences to record and update information about Dt and thus maximise the evidential value of existing data collections.

Acknowledgments

This research was funded by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement n° 335925 (project "The Epistemology of Data-Intensive Science"); and the ARC Discovery Grant "Organisms and Us" (DP160102989). I am very grateful to Dan Bebber, Robert de Bruin, Midori Harris, Val Woods, Steve Oliver and the many others who wish to remain anonymous for taking time from their schedules to discuss their research with me. Many thanks also to the audience and other participants in the symposium "Data in Time: The Epistemology of Historical Data" at the 2016 PSA/HSS meeting in Atlanta, where this paper was presented; participants to the Biological Interest Group in Exeter, and particularly Niccolò Tempini, Brian Rappert, Ann-Sophie Meincke, Dan Nicholson, Giovanna Colombetti, Thomas Bonnin and

Staffan Müller-Wille for their comments; and to Adrian Currie, Alison Wylie, James McAllister, James Griesemer, Rachel Ankeny, William Bechtel, John Dupré and David Sepkoski for useful discussions.

References

Bebber, Dan P., Timothy Holmes Timothy and Sarah J. Gurr. 2014. "The Global Spread of Crop Pests and Pathogens." *Global Ecology and Biogeography* 23(12): 1398-1407.

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

Caetano. C. et al. 2014. "Tolerance of De-regulated G1/S Transcription Depends on Critical G1/S Regulon Genes to prevent Catastrophic Genome Instability." *Cell Reports* 9(6): 2279-2289.

Cleland, Carol. 2002. "Methodological and Epistemic Differences between Historical Science and Experimental Science." Philosophy of Science 69 (3): 474–496.

Currie, Adrian. In press. *Rock, Bone & Ruin: An Optimist's Guide to the Historical Sciences*. Cambridge, MA: MIT Press.

Currie, Adrian and Derek Turner. 2016. "Introduction: Scientific Knowledge of the Deep Past." S*tudies in History and Philosophy of Science: Part A* 55: 43-46.

Feest, Uljana. 2011. "What Exactly is Stabilized When Phenomena are Stabilized?" *Synthese* 182 (1), 57-71.

Griesemer, James R. and Yamashita, Grant. 2002. "Managing Time in Model Systems: Illustrations from Evolutionary Biology." Published in German in 2005. "Zeitmanagement bei Modellsystemen. Drei Beispiele aus der Evolutionsbiologie" in H. Schmidgen (ed.), *Lebendige Zeit*. Berlin: Kulturverlag Kadmos, p. 213-241.

Leonelli, Sabina. 2015. "What Counts as Scientific Data? A Relational Framework." *Philosophy of Science* 82: 810-821.

Leonelli, Sabina. 2016. *Data-Centric Biology: A Philosophical Study*. Chicago, Il: University of Chicago Press.

Leonelli, Sabina. 2017. [DATA_SCIENCE] Interviews PomBase Users, January-February 2016. figshare. https://doi.org/10.6084/m9.figshare.5484010.v1 Accessed 10 October 2017.

Leonelli, Sabina and Rachel A. Ankeny. 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.

Massimi, Michela. 2009. "From Data to Phenomena: A Kantian Stance." *Synthese* 182: 101–116.

McAllister, James W. 2010. "The Ontology of Patterns in Empirical Data." *Philosophy of Science* 77(5): 804–814.

----. 2018 (this issue). “Scientists’ Reuse of Old Empirical Data: Epistemological Aspects.” *Philosophy of Science*.

O’Malley, Maureen A., and Orkun S. Soyer. 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.

Turner, Derek. 2004. “Local Underdetermination in Historical Science.” *Philosophy of Science* 72(1): 209-230.

Woodward, James. “Phenomena, Signal, and Noise”. *Philosophy of Science* 77 (2010): 792–803.

Wylie, Alison. 2017. “How Archeological Evidence ‘Bites Back’: Strategies for Putting Old Data to Work in New Ways.” *Science, Technology and Human Values* 42(2): 203-225.

Chapman, Robert and Alison Wylie. 2016. *Evidential Reasoning in Archeaology*. Bloomsbury.

Figures

Figure 1. CABI Map of the global spread of tomato pathogen *Oidium neolycopersici* in 2007.

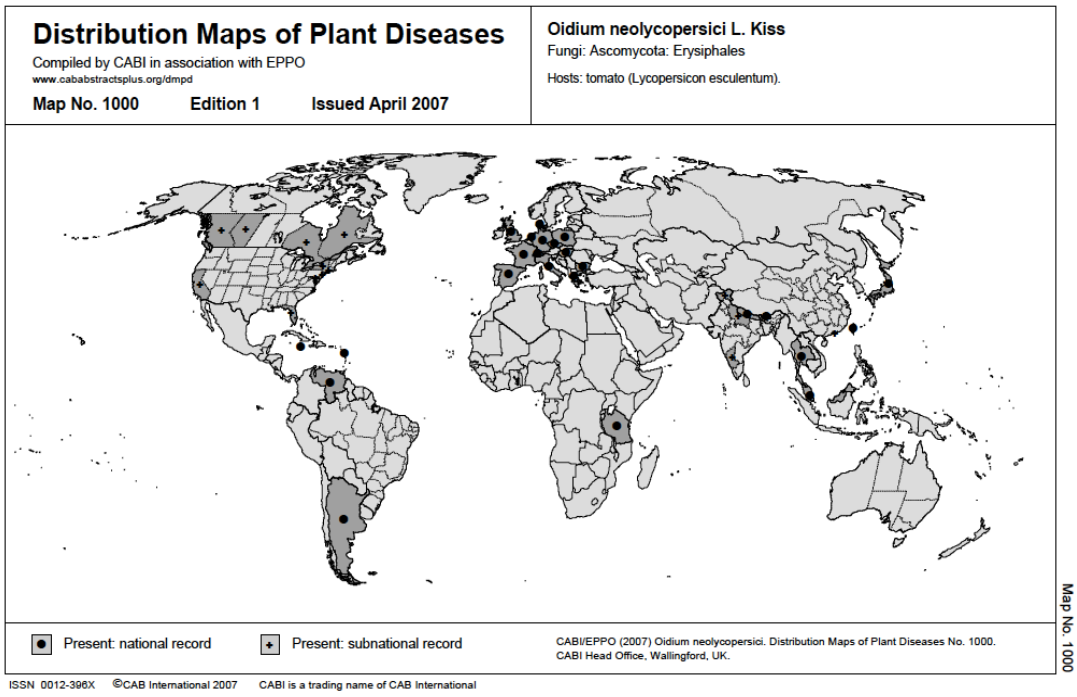

Figure 2 – A simple representation of the different phases of the cell cycle in *S. cerevisiae*. Image produced by Michel Durinx.