Contribution to the special issue on Integration of the journal 'Studies in the History and the Philosophy of the Biological and Biomedical Sciences'. Available online from April 2013, in print from early 2014.

Integrating Data to Acquire New Knowledge:

Three Modes of Integration in Plant Science

Sabina Leonelli

Department of Sociology, Philosophy and Anthropology & Egenis, University of Exeter, UK

Address: Byrne House, St Germans Road, EX4 4PJ Exeter, UK; tel. 0044 1392 725137;

email: s.leonelli@exeter.ac.uk

Abstract

This paper discusses what it means and what it takes to integrate data in order to acquire new knowledge about biological entities and processes. Maureen O'Malley and Orkun Soyer have pointed to the scientific work involved in data integration as important and distinct from the work required by other forms of integration, such as methodological and explanatory integration, which have been more successful in captivating the attention of philosophers of science. Here I explore what data integration involves in more detail and with a focus on the role of data-sharing tools, like online databases, in facilitating this process; and I point to the philosophical implications of focusing on data as a unit of analysis. I then analyse three cases of data integration in the field of plant science, each of which highlights a different

mode of integration: (1) ***inter-level integration****, which involves data documenting different features of the same species, aims to acquire an interdisciplinary understanding of organisms as complex wholes and is exemplified by research on* Arabidopsis thaliana*; (2)* ***cross-species integration****, which involves data acquired on different species, aims to understand plant biology in all its different manifestations and is exemplified by research on* Miscanthus giganteus*; and (3)* ***translational integration****, which involves data acquired from sources within as well as outside academia, aims at the provision of interventions to improve human health (e.g. by sustaining the environment in which humans thrive) and is exemplified by research on* Phytophtora ramorum. *Recognising the differences between these efforts sheds light on the dynamics and diverse outcomes of data dissemination and integrative research; and the relations between the social and institutional roles of science, the development of data-sharing infrastructures and the production of scientific knowledge.*

Keywords: data, integration, plant biology, translational research, model organisms, databases, scientific knowledge, standards.

Highlights:

- Data integration, particularly though online databases and other digital infrastructure, plays a central role in contemporary biological research.
- Plant science constitutes a particularly interesting are to analyse data integration, as it strongly features collaborative efforts to integrate results acquired at multiple levels of organization (molecular, cellular, ecological) and across species.

- I discuss three research traditions in plant science, which exemplify three different modes of integration: **inter-level**, **cross-species** and **translational**.
- This analysis illuminates the challenges of making data usable to the scientific community, the scaffolding needed to transform data available online into new knowledge and the different forms of scientific knowledge that may result.
- I also stress the importance of considering the whole spectrum of scientific activities, including so-called 'applied' research, in order to understand current scientific epistemology.

1. Introduction

The so-called 'data deluge', caused by the overwhelming quantity of information available to scientists through new technologies for the production, storage and dissemination of data, keeps making headlines.[1] Perhaps unsurprisingly, Microsoft researchers have taken the lead in dubbing data-intensive approaches as a brand new approach to scientific research (Hey et al 2009). Equally unsurprising is the position of scholars in the history, philosophy and social studies of science, who are taking a more cautious stand on both the novelty and the revolutionary potential of these developments (see for instance Lenoir 1999, Bowker 2000, and Callebaut 2012 and other papers in the same special issue). Enthusiasts and skeptics tend to agree, however, that digital technologies are fuelling remarkable developments within most scientific fields, including new ways to mine large datasets to extract meaningful patterns; and that philosophers of science ought to take notice and carefully examine their

[1] See the recent special themed issues of Nature (4 September 2008), Science (11 February 2011), the Economist (27 February 2011), and recurring coverage of the data deluge in these magazines as well as the New York Times, the Wired Magazine and most national newspapers worldwide. The latest instantiation of these discussions is the report released by the Royal Society in June 2012 on 'open data', which takes a very strong stance the importance of intelligent open access – which many funding bodies, especially in Europe and the United States, are taking seriously (Royal Society 2012).

epistemic characteristics and possible implications. This paper contributes to this effort by focusing on a central aspect of the data-intensive approach: the integration of large and diverse datasets in order to acquire new knowledge.

Data vary greatly in their format and availability; in the ways in which they have been produced and the materials from which they have been extracted; in the geographical sites, temporal scales and epistemic goals of the scientists generating them; and, most trivially perhaps, in the objects and processes that they can be taken to document. Integrative research efforts need to bridge across these multiple dimensions, by bringing together data obtained in a variety of different settings so that they can be analyzed together and brought to bear on common questions. Data integration thus requires extensive scientific labor, including the development of apposite infrastructures, analytic tools, standards, methods and models. In their recent analysis of integration in molecular systems biology, Maureen O'Malley and Orkun Soyer point to the scientific work involved in data integration as important and distinct from the work required by other forms of integration, such as methodological and explanatory integration, which have been more successful in captivating the attention of philosophers of science (O'Malley and Soyer 2012; see also O'Malley in this issue). This paper extends their argument by looking in more detail at what data integration involves and pointing to the implications of focusing on data as a unit of analysis (sections 2 and 3). I then focus on specific practices of data sharing and re-use in plant science, and argue that focusing on the dynamics of data integration makes it possible to identify at least three distinct modes of integration at work in contemporary biology, which often co-exist within the same laboratory, but whose competing demands and goals cannot usually be accommodated to an equal extent within any one research project (section 4). These three modes are: (1) *inter-level integration*, involving the assembling and interrelation of results applying to different levels of organization within the same species, with the primary aim to improve on existing

knowledge of its biology; (2) *cross-species integration*, involving the comparison and co-construction of research on different species, again with the primary aim to widen existing biological knowledge; and (3) *translational integration*, involving the use of data from a wide variety of different sources in order to devise new forms of intervention on organisms which will improve human health (for instance through agricultural interventions, which are arguably as relevant as medical interventions in fostering human health, though plant biology typically receives less attention than medical research).

Paying attention to the differences and interplay between these modes of integration illuminates the mechanics and challenges of making data not only accessible online but also usable to the scientific community; the large amount of conceptual and material scaffolding needed to transform data available online into new scientific knowledge; and the different forms of scientific knowledge that may result from processes of data integration, depending on which communities, infrastructures and institutions are involved in scientific research. My analysis also underscores the importance of considering the whole spectrum of scientific activities, including so-called 'applied' research carried out by industry or governmental agencies, in order to develop and improve current philosophical understandings of scientific epistemology. As also stressed by Brigandt (this issue) and Bechtel (this issue) with reference to mechanistic explanations, any one component of scientific research (whether data, models or explanations) can potentially contribute to enhancing opportunities for biomedical as well as environmental and agricultural interventions, all of which are of potential value to the preservation and improvement of human health. I will argue that taking the potential social impact of scientific research into account has implications for how the philosophy of science makes sense of the different strategies that scientists may develop when pursuing a research programme.

I shall base my philosophical analysis on historical and ethnographic research that I carried out in the areas of model organism biology, bioinformatics and plant science over the last eight years (documented in detail in Leonelli 2010a, Leonelli and Ankeny 2012 and Leonelli 2012a). In particular, I will use three case studies in contemporary plant science as exemplary for the forms of integration that I wish to discuss: (1) the research activities centered around data gathered on model organism *Arabidopsis thaliana*; (2) the efforts to integrate *Arabidopsis* data with data gathered on the perennial crop family *Miscanthus*; and (3) current investigations of the biology of *Phytophtora ramorum*, a plant parasite that is wreaking havoc in the forests of the South-West of the UK, where I live.

Focusing on plant biology is an important choice here for several reasons. First, despite its enormous scientific and social importance, this is a relatively small area of research - especially in comparison to biomedicine - and has received very little attention within the philosophy of biology.[2] It is also heavily funded by governmental agencies, particularly when it comes to research relating to molecular and genomic aspects of plants – a fact that enhanced opportunities for plant scientists to work on foundational questions, as well as pushing them to join forces and collaborate in order to attract the attention of sponsors and make the best of limited resources (Leonelli 2007). Plant science has indeed been open to generalist and interdisciplinary thinking throughout its history, as even the most reductionist plant scientists tend to be greatly interested in the relations between molecular biology, cellular mechanisms, developmental biology, ecology and evolution (Browne 2001; Botanical Society of America 1994). Further, pre-publication exchanges have been strongly encouraged within this community, particularly again at the time of the introduction of molecular analysis on plants, which was spearheaded by a group of charismatic individuals

[2] Historical work by the likes of Jonathan Harwood (2012), Staffan Müller-Wille (2007), Betty Smocovitis (2008) and Noel Kingsbury (2009) clearly shows the key role played by plant scientists in the development of several branches of biology, including evolutionary theory and genetics.

who explicitly aimed to advance knowledge through open collaboration, thus effectively predating today's 'open science' movement (Rhee 2004, Koorneef and Meinke 2010). These factors have made the community of plant scientists into a relatively more cohesive and collaborative one than more powerful, socially visible and well-funded fields focusing on animal models, such as cancer research or immunology.[3] Given this background, it is not surprising that plant science has produced some of the best available resources for scientific data management and integration to date. Plant scientists' interest in working together, and thus in finding efficient ways to assemble and disseminate their resources and results, long precedes the advent of digital technologies for data sharing, and many of these scientists were quick to seize the potential provided by those technologies to help them in their integrative efforts. As a result, plant science disposes of some of the most sophisticated databases and modeling tools in biology (The International Arabidopsis Informatics Consortium 2012). Its contributions to systems biology are also very advanced, particularly in fostering the development of digital organisms (i.e. the use of mathematical models and simulations to integrate qualitative data in order to predict organismal behavior and traits in relation to the environment).

Yet another characteristic of plant science makes it a particularly fruitful terrain on which to explore ideas on data integration as a key process in producing new knowledge. Plant science produces results of direct interest to several sections of society, including farmers, forestry management, landowners, florists and gardeners, the food industry and the energy industry (through the production of first and second generation biofuels), social movements concerned about genetically modified foods, sustainable farming and population growth, breeders of

[3] Of course this does not mean that plant science is a homogeneous field, or that there are no tensions and non-overlapping programmes at play within it. A major problem is the historical separation between agricultural and molecular approaches to plant science, which came into effect in the second half of the 20th century (Leonelli et al 2012). I also do not mean to assert that there are no comparable examples of cooperation in other fields (for an analysis of such cooperation surrounding model organisms, see Kohler 1994 and Leonelli and Ankeny 2012).

new plant varieties for agricultural or decorative use, and of course national and international government agencies. As I will show, representatives of these groups can be and sometimes are called to participate in the development and planning of scientific projects, thus contributing to choice of goals, opportunities and constraints associated to ongoing research. In significant ways, direct contributions to scientific research by non-scientists make a difference not only to the goals ultimately served by science, but also to its practice, methods and results, including what strategies are used to share and integrate data, and what comes to count as new scientific knowledge arising from such integration. As I hope to illustrate, recognizing the differences in the degrees to which scientific inquiry is brought in contact with other sections of society involves challenging the internalistic view of scientific knowledge that is still favored by many philosophers of science, thus bringing my arguments to bear on recent debates around the social relevance of philosophy of science (e.g., Kourany 2010). Going beyond the view of science as aiming solely to acquire true knowledge of the world may seem a long shot when starting from an analysis of different forms of integration in contemporary plant science; and yet, as I show in this paper, looking at processes of integration 'in action' immediately points to important differences in the types and sources of the data that are being integrated, the integrative processes themselves and the forms of knowledge obtained as a result of integration.[4]

2. Using data to produce knowledge

Before delving into an analysis of different forms of integration in plant science, let me elaborate on what I mean by data and knowledge and what I take to be the relation between

[4] My work is thus aligned with other attempts at broadening the notions of scientific research, collaboration and knowledge traditionally supported within the philosophy of science, including for instance Longino (2002), Douglas (2009), Mitchell (2009), Elliott (2011), Nordmann, Radder and Schiemann (2011), and Chang (2012).

these two key notions. I view data as mobile pieces of information, which are collected, stored and disseminated so as to be used as evidence for claims about specific processes or entities. Thus any material product of research activities, ranging from artefacts such as photographs to symbols, can be considered as a piece of data 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 (through any means ranging from archives to databases, journal publications and stock centres or biobanks). This means first of all that the *evidential value* assigned to an observation or a measurement – the specific claims that it is taken to constitute evidence for – does not need to be specified in advance for that observation or measurement to count as data: what matters is that someone collects and stores that observation or measurement *with the expectation* that it may be used as evidence for one or more claims about the world at some point in the future. What matters in assessing the evidential value of data is their reliability and quality as potential evidence, which is judged by scientists through an evaluation of the ways in which data have been collected and disseminated, including the instruments and materials employed to that effect.

This definition of data is inspired by and compatible with Ian Hacking's (1992) work on marks, even if he limits his analysis to data produced through experiment, while my view includes field observations and the results of simulations and mathematical modelling. It is also largely compatible with James Griesemer's (2006) view on tracking as a key form of scientific inquiry and with Hans-Jörg Rheinberger's (2011) work on the 'medial world of knowledge-making', where he elaborates on the idea of data as things that can be stored and retrieved, and are thus made durable. My approach emphasises the importance of the *use* that is made of data over their intrinsic properties: the evidential value of data comes from their interpretation in relation to specific contexts and goals, rather than as a context-independent quality (Leonelli 2009a). It also separates the analysis of how scientists corroborate their

claims about reality, which is the main issue I am concerned with here, with the analysis of how they develop representations of natural phenomena that are conducive to new understanding ,which is the problem typically discussed by philosophers concerned with modelling.[5] An important implication of this way of viewing data is that, depending on how and where they are used, they can be interpreted as constituting evidence for several different claims and can provide evidence on which to model several different phenomena; and that, given the context-dependence of their use, it is not possible to predict in advance exactly which claims data might be used to corroborate (Leonelli 2009a).[6]

This broad characterization of data is crucial to understanding the link between data and scientific knowledge. Data are not, by themselves, a form of knowledge. Rather, data need to be interpreted in order to yield knowledge; and interpretation, in whichever form and through whichever process it is achieved, involves using data as evidence for a specific claim. Here I agree with Rheinberger's view that the scientific value of data lies in the extent to which they are taken to document aspects of the world: the important question for scientists is what kinds of patterns they are able to extract from data, in order to use them as evidence for specific claims about phenomena. It is that claim, rather than the data, which expresses knowledge about reality, and is therefore often referred to as 'knowledge claim' or *propositional knowledge*. This form of knowledge is also what scientists typically refer to as 'expressing the scientific significance of data'. Another form of scientific knowledge, which is complementary to the propositional form, is the *embodied knowledge* required to handle data and use them as evidence within a specific research context. Embodied knowledge captures

[5] It might be objected here that my account of data is very close to Morgan and Morrison's (1999) account of models as 'tool', and thus that if one adopts the 'mediators' approach to models, there is no difference between a model and data. On my account, the difference between data and models is not ontological, but rather lies in the use made of given artefacts. When something is used as evidence for a claim, it is fulfilling the role of data; when the same thing is used to represent a phenomenon, it is fulfilling the role of a model.

[6] This is not because it is impossible to predict at least some future use of data as evidence: this is clearly the case, as one of the referees of this paper usefully pointed out. Rather, it is because one cannot predict *all* the possible claims that data might be used as evidence for in the future; and also because one cannot predict whether data will *actually* be used as evidence for specific claims until it happens.

what Gilbert Ryle refers to as 'knowing how', thus singling out the knowledge needed to carry out scientific research and distinguishing it from the propositional knowledge used to devise experiments and interpret results ('knowing that'; Ryle 1949). Embodied knowledge includes, for instance, the skills involved in manipulating a mathematical model so that it can be fitted to a given dataset, as routinely done in metabolic control analysis; and the skills involved in understanding that the symbols used to capture sequence data refer to the order and quality of nucleic acid molecules within a genome (for a detailed discussion, see Leonelli 2009b).

Noting the tight relation between propositional and embodied knowledge involved in handling data is relevant to understanding current concerns surrounding the 'data deluge' and the difficulties in transforming data available in public databases into scientific knowledge. Data are made available through public repositories and online databases on the strength of one crucial assumption: that different groups of scientists, including but not limited to the group that originally generates the data, will be able to pick them up, interpret them and use them as evidence for new claims. In this sense the Royal Society stresses the importance of 'intelligent' data access (Royal Society 2012): data circulated online needs to be formatted and visualised in ways that facilitate their use by scientists across the globe, independently of whether those scientists have been involved in the production and collection of those data. Scientists working within different labs across the globe possess their own distinct mix of propositional and embodied knowledge, which makes the re-use of data in new settings both very challenging and potentially fruitful, as different skills and expertise will be employed to interpret the scientific significance of the same dataset. Once a specific dataset is made available to several scientific communities at the same time, each of those communities may have a different perspective on how to interpret it. Hence, one dataset may end up being used as evidence for a much greater number of knowledge claims than if it was kept within the

same research context in which it was originally produced. Scientific pluralism in both the practices (embodied knowledge) and the results (propositional knowledge) of research is thus essential to data-intensive processes of discovery: the generative power of data-intensive science comes from the wide variations in the propositional and embodied knowledge, as well as the goals and interests, possessed by different scientific communities.[7]

3. Standardising knowledge to integrate data

The above vision of data-intensive methods as efficiently channelling diverse research activities into new discoveries, which fuels much of the expectations and hype underlying data-intensive science, is immediately challenged when considering its practical realisation. Here is where the issue of data integration, and all the complications involved in achieving it, comes into the picture. Data cannot be stored and circulated without any organising principles; this basic requirement is ever more pressing when posting data online, given the potential for large storage capability, immediate dissemination and wide reach provided by the web. Data stored in digital databases need to be standardised, ordered and visualised, so that scientists can retrieve them from databases in ways that help them in their own research. These processes constitute an important form of data integration, which involves significant amounts of labour and expertise in biology as well as computer science, including the ability to conceptually order data, format them to fit specific programmes, and developing adequate software and models. Indeed, making data available through databases often requires a sophisticated understanding of what data might be used for, as well as extensive work on the classification and modelling of datasets so that they become compatible with each other,

[7] My interpretation of pluralism is more encompassing than Sandra Mitchell's integrative pluralism, which focuses chiefly on the integration of multilevel explanations and is grounded in the complexity specific to biological systems (Mitchell 2003); my position is closest to Hasok Chang's recent account of interactive pluralism, which involves 'pluri-axial regimes of practice' encompassing all elements of research, including data, models and a variety of possible aims (Chang 2012, 272).

retrievable and re-usable by the wider scientific community – a form of scientific labour that I shall refer to as 'curatorial', as it is currently database curators who are largely responsible for taking care of data in this way.[8] Data curation constitutes an integral part of processes of discovery, where conceptual and practical decisions about how to integrate and visualise data affect the form and quality of knowledge obtained as a result.

It should be noted here that the integrative efforts of curators are not well coordinated with the data integration carried out by database users, such as experimental biologists. This is, on the one hand, because scientists involved in generating knowledge through data re-use typically do not want to spend time participating in curatorial efforts, as it robs them of precious research time (the fact that curatorial activities are not recognised within current systems of credit attribution in science only adds to this problem; McCain 1995, Ankeny and Leonelli 2011, Leonelli and Ankeny 2012). On the other hand, and perhaps more problematically from the epistemic viewpoint, many biologists do not fully appreciate that data dissemination via databases involves substantial integrative efforts, rather than being simply 'conducive' to integration. This is partly due to the pernicious idea that data found online are 'raw', i.e., that they are shown online in exactly the same format as when they were first produced. Biologists who are committed to this idea are reluctant to consider how the integration of data effected through databases is likely to affect the original format of data and the ways in which they are visualised. So databases are viewed by many scientists as a neutral territory, through which data travel without changing in any way; while, in order to work efficiently in disseminating data, databases need to function as a transformative platform, within which data are carefully selected, formatted, classified and integrated in order to be retrieved and used by the scientists who may need them.

[8] For a detailed study of the professional role of curators in making data travel in contemporary biology, see Leonelli (2010b) and (2012b). For historical analyses of how the role of data curators has evolved throughout 20th century biology, see Strasser (2011) and Leonelli and Ankeny (2012).

In their analysis of data integration, O'Malley and Soyer (2012) discuss one key aspect of this situation: 'the activity of making comparable different datasets from a huge variety of potentially inconsistent sources' (p.61). As they illustrate, data dissemination through databases is often expected to remove inconsistencies among sources – such as differences in the instruments used to produce data – and thus make it possible to interpret data as a unified whole, without worrying about their provenance and about how they were merged (e.g., Ideker et al 2007). Other potential inconsistencies to be removed include differences in data formats; standards for what count as valid or reliable data; and data types, ranging from genome sequences to metabolic and even morphological data (as in the recent case of phenomics). Integration within databases can thus be described as the gathering of several different types of data, obtained by a variety of occasionally inconsistent sources, so that they can be searched and analysed as a single body of information.

This form of data integration is largely achieved through the development of standardised descriptions of both the propositional and the embodied knowledge employed in producing the data in the first place. Database curators do not achieve data integration by avoiding worries about data provenance or manipulations of the data that they collect into their resource. Rather, curators are well aware that information about the provenance of data – how they were produced, why and by whom – is crucial to interpreting the data and assessing their quality and reliability; they are also aware that how data are formatted, annotated and visualised in a database determined the extent to which data are re-widely usable. Placing value on the diversity of life histories of data, and of the settings in which they are produced, is thus crucial to the practice of data curation. Thus, overcoming the problem of inconsistent sources, as well as the broader question of how to integrate and visualise data that come in different formats and types, involves putting information about the provenance of data at the centre of the database itself. To this aim, curators are developing standardised descriptions of

techniques, assumptions, methods and conditions under which data are produced, so that information about how data have been generated can be intelligible as widely as possible beyond the boundaries of the producers' lab.[9]

One example of such standards is data formats and confidence rankings. Curators determine which formats (and thus which instruments and outputs) data producers should privilege, and study ways to translate between these formats and others used by data producers. Further, curators are increasingly pushed to provide at least some preliminary assessment of the quality and reliability of data (a process often referred to as 'data control'), such as confidence rankings where datasets are classified depending on the trust placed by the scientific community on the methods and instruments used for data generation (see for instance the controversies over the status of microarray data; Keating and Cambrosio 2012). Another example is the selection of salient features of data production. Curators select and standardise information, usually referred to as 'meta-data', about where specific datasets originally come from, how they were collected and how they were formatted, annotated and visualised upon entry into the database. This involves making decisions about which knowledge, both embodied and propositional, is most relevant to situating data, integrating them with other data, and interpreting their significance. Several ongoing initiatives in bioinformatics, usually initiated by database curators, center upon the criteria for choosing which instruments, experimental conditions, techniques and procedures need to be reported when disseminating data online. One relatively successful effort is Minimal Information on Biological and Biomedical Investigations (MIBBI), which provides standardized descriptions of the embodied knowledge involved in biological experiments (Taylor et al 2008). Another

[9] Curators are able to do this by combining several research skills that typically includes both propositional and embodied knowledge as a bench scientist and some training in information technology and programming (Leonelli 2010b, 2012a). In many cases, curatorial teams are composed of both biologists and computer scientists, in order to cover all relevant expertises; yet, it is often the vision of biologists, and their awareness of what prospective database users need, that guides the development of databases.

is the standardized description of features of seeds to be stored in seedbanks provided by the Plant Ontology (Avrhaman et al 2008). Notably, written descriptions are increasingly seen as insufficient ways to capture embodied knowledge, and are complemented by other media such as videos of data production processes (for instance, the *Journal of Visualised Experiments*). Even a brief overview of these standardising practices indicates that curators are contributing to the formalisation and expression of both the propositional and the embodied knowledge required to make sense of data and transform them into new knowledge, thus playing a key role in facilitating data integration both within and beyond databases (for a study of this process with regard to the formalisation of propositional knowledge through bio-ontologies, see Leonelli 2012b).

4. Data Integration in Plant Science

I will now consider three actual cases of data integration, each of which leads to the production of new knowledge. All three cases come from the field of plant science, which, as I stressed in my introduction, finds itself at the forefront of data integration and data-intensive discovery. One instance of this is The Arabidopsis Information Resource (TAIR), an online database devoted to the storage and dissemination of data collected on *Arabidopsis thaliana*, a key model organism for plant science. Data integration has been a central concern for TAIR curators since its inception, since the database contains several types of data about *Arabidopsis*, including sequence data, transcriptomics, proteomics and even phenotypic data, which all need to be searchable and retrievable by users; and aims to serve comparative research with other organisms, which involves extensive efforts to integrate data across species. TAIR curators, who include experts in various aspects of plant biology as well as computer programmers, have thus long been aware that their database is expected to both

integrate data as part of its efforts to disseminate them, and to foster further integration (of data as well as other components of scientific research) by its users. For instance, a major achievement has been the integration of available data about genes and their functions, enzymes, compounds and reactions to visualize and study specific metabolic pathways (through two tools called MetaCyc and AraCyc; Zhang et al 2005). More recent updates include a complete revision of the Arabidopsis genome, in which 'pseudogenes and transposon genes were re-annotated, and new data from proteomics and next generation transcriptome sequencing were incorporated into gene models and splice variants' (Lamesch et al 2011, 1).

Sue Rhee, who has directed the resource in its first five years, described the type of biology that TAIR is supposed to foster in the following way:

> 'If the next twenty years of biology could be summed up into one word, it would be 'integration'. We will see *integration of basic research with applied research* in which plant biotechnology will play an essential role in solving urgent problems in our society such as developing renewable energy, reducing world hunger and poverty, and preserving the environment. We will see *integration of disparate, specialised areas of plant research into more comparative, connected, holistic views and approaches in plant biology*. We will also see *more integration of plant research and other biological research, from microbes to humans, from a large-scale comparative genomics perspective.* Bioinformatics will provide the glue with which all of these types of integration will occur.' (Rhee et al, 2006, 352)

Rhee implicitly singles out different types of integration here, which I highlighted in italics. I find her preliminary classification illuminating, as it implicitly points to a difference in modes of integration that has not hitherto been explicitly discussed within philosophy of science

literature, and which places data integration in the context of the variety of interests, epistemic and institutional structures and goals at play in contemporary plant science research. I thus want to expand on the three approaches to integration that Rhee distinguishes here, and which I shall call 'inter-level', 'cross-species' and 'translational'. In the next three subsections, I look in more detail at these types of integrative effort by picking case studies that exemplify each of them.

4.1 Inter-level integration and model organism research

The idea of centering research efforts on a handful of species has been a hallmark of 20th century experimental biology, culminating in the sequencing of the genomes of the most popular of these organisms: *E. coli*, *C. elegans*, yeast, *Drosophila*, zebrafish and, for plants, *Arabidopsis*. One of the aims of this research strategy was to integrate knowledge produced on different aspects of the biology of the same organism, so as to understand the biology of the organism as a holistic whole rather than as an ensemble of disconnected parts (Ankeny and Leonelli 2011).[10] Despite the many criticisms leveled at the bias created by this strategy in funding allocation and research directions (Bolker 1995, Davis 2004), model organism research has been immensely successful, particularly in plant science where enormous advances were made in understanding complex processes such as photosynthesis, flowering and root development by focusing research efforts on *Arabidopsis* (Bevan and Walsh 2004). Understanding these processes requires an interdisciplinary approach comprising several

[10] I should note that this reading of model organism biology does not run counter the observation that many research projects under that banner deal with isolated components and pathways. What Rachel Ankeny and I have defended is that the idea of focusing research on one species is historically grounded on the expectation of being able to integrate those disparate bits of knowledge at some point in the future. The attempt to achieve such holistic understanding functions as a key motivating factor for researchers in these communities, and shapes the methods used to pursue research.

levels of organization, from the molecular to the developmental and morphological.[11] *Arabidopsis* provided a relatively simple organism on which integration across these levels could be tried out under the controlled conditions of a laboratory setting. The use of *Arabidopsis* has been highly controversial within plant science at large, with scientists specializing on the study of other plants and/or plant ecology complaining that focusing on *Arabidopsis* took resources away from the study of plant biodiversity, evolution and the relation between plants and their environment (Leonelli et al 2012). At the same time, considering *Arabidopsis* in relative isolation from its natural environments and other plants has been very successful in generating important insights about its inner mechanisms, particularly at the molecular and cellular levels. For instance, the detailed mechanistic explanations of photosynthesis achieved to date, and the resulting ability of scientists to manipulate starches and light conditions to favor plant growth, are largely due to successful attempts to bring the study of enzymes and other proteins involved in photosynthesis (which involves the analysis of molecular interactions within the cell nucleus) in relation with the study of metabolism (which involves the cellular level of analysis, since it focuses on post-translational processes outside the nucleus). The integration of these two levels of analysis is fraught with difficulties, since the evaluation of data about DNA molecules (as provided by genome analysis) needs to take account of their actual behavior within and interactions with the complex and dynamic environment of the cell, which makes it extremely difficult to model metabolic pathways (e.g. Stitt et al 2010; Bechtel and Brigandt, in this issue, also discuss the dynamic modeling of pathways).

This is an excellent illustration of inter-level integration as aiming to understand organisms as complex entities, by combining data coming from different branches of biology in order to

[11] I here adopt a broad definition of 'level of organisation', which is meant to reflect a way to organise and subdivide research topics that is still very popular within biology: that is, the focus on specific components and 'scales' of biological organisation, each of which requires a specific set of methods and tools of investigation that suit the dimensions and nature of the objects at hand.

obtain holistic, interdisciplinary knowledge that cuts across levels of organization of the same organism.[12] This case also instantiates the key role played by databases and curatorial activities in achieving inter-level integration, as the development of centralized depositories for data (first material archives, and then digital databases) has been central to the success of model organism research (Bult 2006, Leonelli and Ankeny 2012). Since its inception, TAIR has been heavily engaged in facilitating inter-level data integration, particularly through the development of software and modeling tools that enable users to combine and visualize several datasets acquired on two or more levels or organisation. Tools such as the above-mentioned AraCyc and MetaCyc, for instance, have been fundamental in enabling researchers to combine and visualize genomic, transcriptomic and metabolic data as a single body of information. This has made it possible to integrate data generated from the molecular and the cellular levels of organization, thus enabling researchers to visualise and study specific metabolic pathways. TAIR curators have also devoted much attention to developing 'evidence codes' to capture meta-data about experimental settings and techniques used to generate data in the first place. Evidence codes, include labels such as 'inferred from mutant phenotype' and 'inferred from reviewed computational analysis', are used by TAIR users to gather information about the provenance of the data found online and thus assess their reliability and quality. These meta-data facilitate inter-level integration by enabling researchers working at a specific level (e.g. cellular) to assess and interpret data gathered at another level (e.g. molecular). Last but not least, TAIR curators have endeavored to collaborate with researchers from all corners of plant science in order to generate keywords to describe the biological objects and processes currently under investigation (keywords that could, in turn, be used to retrieve data relevant to those objects and processes from the database). Crucial requirement for these keywords was that they are intelligible to researchers

[12] What I here call inter-level integration has also been discussed by proponents of mechanistic explanation (Craver 2005, Darden 2005, Bechtel this issue), though Craver (2005) has also pointed to the importance of intra-level integration ignored by proponents of reduction.

from all branches of biology, ranging from genetics to immunology, ecology and, increasingly, evolutionary-developmental biology. To this aim, TAIR curators were involved in several 'content meetings', where biologists specializing on research at a variety of levels of organization met to discuss how to identify and define keywords to enable inter-level communication (Leonelli 2010a). Two prominent results of these efforts are the Plant Ontology and the Gene Ontology, which have been implemented within TAIR as classification systems for the retrieval of data about, respectively, gene products and plant features (Avraham et al 2008; for a philosophical analysis see Leonelli 2010b and 2012b).

It is important for my purposes here to stress that inter-level research on *Arabidopsis*, as in the case of many other popular model organisms, was driven strongly by the scientific community, with the support of funding bodies such as the National Science Foundation, but with little influence from other parts of society which have stakes in plant science – such as, for instance, agricultural research, farmers and industrial breeders (Leonelli et al 2012). Attempting to integrate data resulting from different strands of plant research was seen as requiring expert consultations within the plant science community, aimed to resolve technical and conceptual problems in an effort to acquire an improved understanding of *Arabidopsis* biology. The very idea of using data from the same model organism to achieve inter-level integration exemplifies the image of science often celebrated within philosophy of science circles, as well as many popular accounts of discovery: these are scientists who wish to bring together their results within expert circles that are largely separate from other sections of society; and that this is done in order to acquire a more accurate and truthful understanding of biological processes, resulting in the articulation of reliable explanations of those processes (and related forms of intervention).[13] Indeed, inter-level integration is heavily concerned with the methodological and conceptual challenges deriving from the attempt to collaborate across

[13] This view is compatible with an emphasis on experimental intervention ('learning by doing') and exploratory research as means to achieve biological discoveries (e.g., Burian 1997, O'Malley 2011).

disciplines, such as the challenge of communicating ideas across different research communities, which involves some standardization both of the propositional knowledge produced by each community (so that others can understand it, as in the case of keywords such as in the Gene and Plant Ontologies) and of the embodied knowledge developed (so that experimental results obtained by one community can be reproduced by others if needed, as in the case of evidence codes in TAIR). In order to agree on the keywords and metadata used to classify and retrieve data on metabolic pathways, TAIR curators have engaged in extensive consultations with scientists working on all the relevant levels of biological organisation (including molecular and cellular), so as to make sure that ensure that data integration tools within TAIR were set up in ways agreeable to and compatible with research at several levels of analysis.

The efforts of TAIR curators are thus reminiscent of the challenge of communicating propositional knowledge across different scientific groups, which many philosophers of science have focused upon when reflecting on scientific integration. This scholarship includes Lindley Darden and Nancy Maull's classic paper on inter-field theories (1997) and, most recently, Bechtel and Richardson (2010), Mitchell (2003, 2009) and Brigandt (2010) on integrative explanations. My focus on data integration, rather than on the integration of explanations, models and theories that these authors address, helps to highlight the importance of communicating embodied knowledge, for example in the form of meta-data, in order to achieve integration (a factor that tends to be overlooked in literature focused on explanations and conceptual structures). The focus on data also helps to stress the diversity the epistemic goals and priorities driving integration, as well as the distinct forms of knowledge that may be achieved through pursuing these goals.[14] To this end, I will now

[14] My approach is thus in line with the analysis of epistemic goals developed by Ingo Brigandt (2013) and Maureen O'Malley's emphasis on forms of integration beyond the explanatory level (O'Malley and Soyer 2012, O'Malley this issue).

discuss two forms of integration that operate differently from inter-level integration, and whose results are distinct from the knowledge of plant biology acquired in this case.

4.2 Cross-species integration and biofuels research

In cross-species integration, scientists place more emphasis on comparing data available on different species, and using such comparisons as a springboard for new discoveries, rather than on integrating data across levels of organization of the same species.[15] Consider current research on grass species *Miscanthus giganteus*. *Miscanthu*s is a perennial crop, which means that it can be cultivated in all seasons, without interruptions to the production chain. It grows fast and tall, thus guaranteeing a high yield; and it grows easily on marginal land. These characteristics have made it a good candidate as a source of bioethanol, particularly because it poses less of a threat to food production than other popular sources of biofuels such as corn (whose cultivation for the purposes of biofuel production has taken big chunks of land in the United States away from agriculture, which is deemed to have affected the availability and price of agricultural produce worldwide; Babcock 2011). The potential of *Miscanthus* as a source for biofuels is one of the factors that first spurred scientific research on this plant. And indeed, such research is ultimately aimed to engineer *Miscanthus* so that its growing season is extended (by manipulating early season vigor and senescence) and its light intake is optimized (modify architecture, e.g. several sprawling stems, or increase stem height and number). In this broader sense, research on *Miscanthus* is a good example of research aimed at developing techniques for intervening in the world, and ultimately for improving human life. However, there is at least another reason why *Miscanthus* has become an important

[15] In plant biology, comparisons between strains and varieties regarded as belonging to the same species can often be as interesting as comparisons among species. My analysis here is thus not meant to endorse strict taxonomic distinctions, but rather to capture the importance of comparing differences between groups of organisms. Understood in this broad sense, 'cross-species integration' can also include cross-variety and cross-strain analysis.

organism in contemporary plant science, which has little to do with its energy output. This is the opportunity to efficiently cross-reference the study of *Miscanthus* with research on *Arabidopsis*.

On the one hand, *Arabidopsis* provides the perfect platform on which to conduct exploratory experiments, given how much scientists already know about that system (thanks to inter-level integrative efforts) and the extensive infrastructure, standards for collecting data and metadata, and modeling tools already available on it (e.g. as incorporated in TAIR). On the other hand, *Miscanthus* provides a good test case for ideas first developed with reference to *Arabidopsis*, whose value for other species researchers have yet to explore. Many researchers trained on *Arabidopsis* biology have thus switched to comparative research on these two systems, which has hitherto proved very productive: many experiments needed to acquire knowledge about molecular pathways relating to abiotic stress can be more easily carried out on Arabidopsis than on *Miscanthus*; data collection and integration on *Miscanthus* itself is facilitated by the standards, repositories and curatorial techniques already developed for *Arabidopsis*; and new data types, such as data about how *Miscanthus* behaves in the field (e.g. its water intake), can be usefully integrated with data about *Arabidopsis* metabolism, resulting in new knowledge about how plants produce energy in both species.

Note that this research requires more than simply the transfer of knowledge from one plant to the other: plant scientists need to iteratively move between the two species, compare results and integrate data at every step of the way, in order for new knowledge to be obtained. In other words, this research requires genuine integration between results obtained on *Miscantus* and *Arabidopsis*. For instance, the consultation of TAIR data on *Arabidopsis* genes that regulate floral transition has been a crucial impetus for research on flowering time in *Miscanthus*, since those data provided *Miscanthus* researchers with a starting point for investigating the regulatory mechanisms for this process (Jensen 2007); and the subsequent

findings on the susceptibility of *Miscanthus* flowering to temperature and geographical sites are feeding back into the study of flowering time in *Arabidopsis* (Jensen et al 2011).

Perhaps the most important distinctive feature of cross-species integration is that it fosters studies of organismal variation and biodiversity in relation to the environment, with the aim to understand organisms as relational entities, rather than as complex – yet self-contained - wholes (as in the case of inter-level integration). This is because as soon as similarities and differences between species become the focus of research, plant researchers need to identify at least some of the reasons for those similarities and differences, which unavoidably involves considering their evolutionary origins and/or the environmental conditions in which they develop. Hence, like inter-level integration, cross-species integration may be construed as aiming to develop new scientific knowledge of biological entities. However, the way in which it proposes to expand the realm of existing knowledge is not necessarily by extending the range of inter-level explanations available, but rather by extending the range of organisms to which these explanations may apply. Indeed, while researchers can and often do pursue both inter-level and cross-species integration at the same time, it is also possible to achieve cross-species integration without necessarily fostering inter-level integration. This is the case, for instance, when using comparisons of data about flowering time between *Arabidopsis* and *Miscanthus* to explore the respective responses of the two plants to temperature; such cross-species comparison can eventually be used to foster inter-level understanding of flowering that integrates molecular, cellular and physiological insights, but this is not necessary in order for cross-species integration to be regarded as an important achievement in its own right.

Further, cross-species integration poses a different set of challenges from inter-level integration, whose resolution can easily constitute the sole research focus of a research project. It requires accumulating data that are specifically relevant for the purposes of comparison (for instance, by making sure that data obtained on *Miscanthus* are generated

with tools and on materials similar to the ones available on *Arabidopsis*, so as to make comparison tenable) as well as developing infrastructure, algorithms and models that enable researchers to usefully visualize and compare such data. Thus in our example, TAIR provides a key reference point, but it is not sufficient as a data infrastructure for such a project, for the simple reason that it focuses on Arabidopsis data alone. Indeed, the difficulties of using TAIR for cross-species integration have become so pronounced and visible within the plant science community, that TAIR itself is now undergoing heavy restructuring to secure its future compatibility with *both* inter-level and cross-species analysis (The International Arabidopsis Informatics Consortium 2012). This task is made even harder by the terminological, conceptual and methodological differences between communities working on different organisms, as well as differences in perceptions of what counts as good evidence and the degree to which specific traits are conserved across species through their evolutionary history. These differences need to be clearly signaled when constructing databases that include and integrate data acquired on different organisms. The Gene Ontology project, which as I mentioned earlier was started as a means to disseminate data within a handful of model organism databases such as TAIR, is now used extensively as a platform for the integration of gene products data across species (The Gene Ontology Consortium 2004) and exemplifies the difficulties and controversies involved in rising to this challenge (Leonelli et al 2011). Even the comparison of different genome sequences, which should be among the easiest to accomplish given the highly automated and standardized production of this type of data, is fraught with problems (e.g., Quirin et al 2012). In this context, norms such as the principle of genetic conservation, by which scientists see regions of the genome that are highly conserved across species as potentially linked to important functions (since less relevant regions are assumed to have been selected away through the evolutionary process),

matter over and above the norms of validity and accuracy used to achieve inter-level integration.

In conclusion, the increasing emphasis on cross-species integration can be seen as complementary to, and yet separate from, inter-level integration. The two forms of integration are clearly interconnected. I have illustrated how the inter-level integration achieved for Arabidopsis through model organism biology provides an important reference for cross-species integration in several areas of plant science. Indeed, inter-level integration of data within one species are often the starting point for cross-species investigation and for the integration of data about the same process as it manifests itself in different species. However, this does not necessarily mean that cross-species integration presupposes inter-level integration as a matter of principle, or in all cases. Further, these two forms of integration raise different epistemic challenges, which do not have to be addressed within the same research project; and, perhaps most importantly, they require different sets of data and infrastructures – as illustrated by the practical difficulties in using TAIR, whose primary focus is inter-level integration in *Arabidopsis*, to study other plant species such as *Miscanthus*.

4.3 Translational integration and plant-pathogen interaction

Research on *Miscanthus* could be seen as exemplifying research that has been targeted and structured to serve societal goals – in this case, the sustainable production of biofuels. However, while the goals set by funding agencies and industry have been crucial to the choice and funding of *Miscanthus* as an experimental organism, plant scientists engaged in *Miscanthus* research have not, at least until recently, worried much about how the plants that they are engineering could actually be transformed in biofuel, whether that process would be

particularly sustainable and economical, and how those 'downstream' considerations might affect 'upstream' research. This set of consideration has largely been left to politicians and industry analysis, while plant scientists focus on the task of achieving new knowledge of *Miscanthus* biology. In other words, scientists and curators focusing on cross-species integration are primarily focused on producing knowledge about plant biology that is more accurate and all-encompassing than that already available to them. Their expectation is that knowledge produced in this way will eventually inform the mass engineering of *Miscanthus* plants, thus creating biomass from which bioethanol could be efficiently extracted. This is a reasonable expectation, and the knowledge obtained from *Miscanthus* research will undoubtedly inform biofuel production in the future. However, other research programs in plant science are explicitly planned and shaped so as to serve societal needs *even before* they improve on existing scientific knowledge of the organisms involved, thus de-emphasizing the production of new biological knowledge in favor of producing strategies for managing and manipulating organisms and environments so that they support human survival and well-being in the long term.

Consider for instance research on plant pathogens, which are becoming a serious threat to ecosystems and agriculture worldwide because of global trade and travel that facilitates the dispersion of parasites well beyond their natural reach (Potter et al 2011). Dealing with plant pathogens that are new to a given territory is a matter of urgency, since targeted interventions need to be devised before the pathogen creates much damage; scientific research is a key contributor, as these pathogens are often relatively unknown within the scientific literature and are anyhow interacting with a whole new ecosystem, often with unprecedented results. Recently I participated in a workshop organized at my university to discuss how plant science can help to suppress an infestation of *Phytophtora ramorum*, a plant parasite that landed in the South-West of the UK in the early 2000s and has been ravaging the forests of Devon ever

since. The infestation had gotten particularly worrisome in 2009, when it started to affect large chunks of the local population of larches. The workshop brought together representatives of relevant plant biology and data curation conducted at several research institutes in the UK; the UK Forestry Commission; the Food and Environment Research Agency; private landowners; social scientists; and representatives of other governmental agencies, such as the Biotechnology and Biological Sciences Research Council.

At the start of the workshop, it was made clear that there are several alternative ways to tackle the *Phytophtora* infestation, including burning the affected areas, using pesticides, cutting down the trees, letting trees live and introduce predators, making affected areas inaccessible to humans, or simply letting the infection run its course. A focus of debate was then to determine which scientific approach would provide empirical grounds to choose an effective course of action among all those possible interventions. Acquiring novel understandings of the biology of *Phytophtora* was obviously important in this respect; but it was not the primary goal of the meeting, and it was made clear that choices concerning which research approach would be privileged in the short term should not be based on the long-term usefulness of that approach in providing new biological insights. This was particularly relevant in selecting strategies for data collection and types of data to be privileged in further analysis. For instance, whole genome sequencing was agreed to be an excellent starting point for a traditional research program seeking to understand the biology of *Phytophtora* through inter-level and cross-species integration, especially since data could be compared (through online databases) to data generated on other strains of *Phytophtora* by European and North-American labs. However, many participants questioned the efficiency of this strategy in providing quickly genetic markers for *Phytophtora ramorum*, which could be of immediate use to combat the infestation. It was argued that focusing genomic research on more specific parts of the genome, such as loci already known to be linked to pathogenic traits, would

provide a way for the forestry commission to test trees in areas not yet affected and determine immediately whether the infestation is spreading (the merits and drawbacks of using PCR-based diagnostic were debated at length). Further, much debate surrounded the possible ecological, economic and societal implications of each mode of intervention under consideration, and the science related to it. Biological research was thus not the sole empirical ground to assess the quality and effectiveness of an intervention; other factors included local ecology, touristic value of the areas and the economic value of the wood being cut down (that is, factors that include the environmental considerations that I signaled as central to the cross-species approach, as well as economic and social elements that cross-species integration would not regard as relevant). Only through such an overall assessment could participants and scientists determine the overall sustainability of the research program that was being planned.

Notably, each participant contributed not only a specific perspective on what the priorities are in dealing with *Phytophtora*, but also their own datasets for integration with the molecular and phenotypic data to be gathered by plant scientists. These included data of great relevance to scientific research, which however were collected for purposes other than the scientific study of *Phytophtora*: for instance, geographical data about the spread of the infestation, which was gathered by Forest Research (the research arm of the Forestry Commission) in the course of aerial surveillance and was picked up by mathematical modellers to help predict future spread patterns; and photographs of affected trees in several areas, collected by the Forestry and local landowners, and seized upon by plant pathologists at the James Hutton Institute as evidence for plant responses to biotic stress. Acquiring access to those data constitutes an achievement in itself for plant scientists, since some of the stakeholders involved are more willing to disseminate their data than others. For instance, the Forestry Commission is more reluctant to share data than plant scientists working at the University of

Exeter, for whom contributing to online sequence repositories such as the Sequence Read Archive or GenBank is a part of research routine. Further, scientists at the meeting were not sure about which existing online database, or combinations of databases, would best serve the desired integrative efforts. One obvious candidate would be PathoPlant, a database explicitly devoted to data on plant-pathogen interaction, but its use was not explicitly discussed at the meeting I attended, perhaps because it was not clear to participants that such a database would serve their immediate research goals.[16]

Indeed, plant scientists involved in the meeting at Exeter University found themselves negotiating with stakeholders, some of whom are also arguably involved in scientific research (such as biologists working for the Forestry), whose main aim was not the production of new insights on *Phytophtora* biology, but rather the achievement of a reliable body of evidence that would help deciding how to tackle *Phytophtora* infestations. This negotiation, which is the key feature of this type of integrative effort, is not easy especially given the tendency of plant scientists to reach for inter-level and/or cross-species integration too. Thus, plant scientists at the meeting strongly advocated the expansion of research on *Phytophtora ramorum* into a long-term program that would investigate the relative virulence of the pathogen on different hosts (which would involve detailed studies of the hosts – tree species – as well), assemble whole genome data on all available and emerging strains, and investigate the mechanisms that trigger virulence. All of these research programs, which clearly involve inter-level and cross-species integration, would provide knowledge about the biology of *Phytophtora* that scientists view as crucial to develop better interventions; however, these programs would require substantial funding and considerable time in order to yield results, and scientists were pushed by other parties to articulate more fully how the systematic whole

[16] Interestingly, the development of PlantPatho itself has involved considerable integrative efforts, both inter-level (by integrating data from different features of both host plants and pathogens) and cross-species (by integrating data coming from a variety of different species, though it must be noted that Arabidopsis research has provided much of the data used to start this initiative; Buelow et al 2007).

genome sequencing of *Phytophtora* strains would eventually lead to effective interventions on the infection. In particular, it was argued that although the cross-species and inter-level integration acquired through these approaches was desirable, it was not necessary in order to facilitate decisions on how to eradicate *Phytophtora*. For example, PCR-based diagnostics, though arguably useless to the pursuit of a better understanding of the biology of *Phytophtora* and its hosts, might work perfectly well for the purposes of diagnosing infection.

Negotiations among molecular biologists, scientists working at Forest Research, and other stakeholders are still ongoing at the time of writing, and engagement in these discussions is generating a shared research programme, part of which will involve the development of a database that fosters the integration of data relevant to the study the virulence and potential environmental impact of *Phytophtora*. This case nicely exemplifies the characteristics of translational integration, which privileges the achievement of improvements to human health, for instance through targeted interventions on the environment and the use of existing resources, over the production of new scientific knowledge for its own sake.[17] I take the term 'translational' from current policy discussions of the importance of making scientific research useful to wider society, as instigated for instance by the NIH in the early 2000s. However, I do not subscribe to the linear trajectory of research from 'basic' to 'applied' that is often used within such policy discussions. Rather, I wish to use the category of 'translation' to focus on specific ways in which scientists frame their research so as to respond to a social challenge. In the case I considered, scientists aim to produce new forms of intervention that are targeted to the situation at hand. This is not in itself sufficient to differentiate translational integration from inter-level and cross-species integration. Many philosophers have rightly argued that

[17] It is important to stress here that human health can hardly be construed separately from the health of the environments that support human survival and well-being, and particularly plants as key sources of air, fuel and nurture. In this paper, I thus endorse the idea of green biotechnology, and plant science, as playing a key role in protecting and improving human health. This may be perceived as counter-intuitive by scholars who view red biotechnology, and particularly the medical sciences, as the only form of knowledge that is concerned with human health; I hope that this analysis helps to correct this common misconception.

learning to intervene in the world, and particularly to manipulate organisms in the case of biology, is part and parcel of scientific research and is inseparable from the process of acquiring new knowledge about the world; there is thus no clear epistemic distinction between 'making' and 'understanding', since many scientists develop new types of experimental interventions as a way to acquire new knowledge, and vice versa.

What I think makes translational integration distinct from the other two modes is the strong commitment to producing results that affect (and hopefully improve) human health, which involves the development of research strategies and methods that are distinct from the ones employed to achieve inter-level and cross-species integration.[18] My definition of translation is therefore narrower than the definition provided by Maureen O'Malley and Karola Stotz, according to which translational research consists of 'the capacity to transfer interventions from context to context during the pluralistic investigation of a system' (2011). I agree with them that translational research involves such movement of knowledge, but I also think that these transfers can be geared to satisfying a variety of specific agendas, several of which are not primarily concerned with how scientific knowledge affects society. Many parts of biological research inherit and refashion techniques for intervening on organisms, without necessarily aiming to produce socially valuable results in the short term. Of course, all scientific research has the potential to ultimately improve human health, and yet some parts of science are not explicitly conducted to foster this goal in the short term (which is, incidentally, a very good thing, both because the potential social benefits of science are unpredictable, and because the social agenda for what counts as beneficial to humanity changes with time and across domains). I see the extent to which a group of scientists explicitly subscribes to the agenda of social change – and shapes its research accordingly – as

[18] For arguments concerning the use of biological research, including the forms of mathematical modelling characteristic of systems biology, for medical purposes, see the papers by Bechtel, Brigandt, and Plutynski in this issue.

marking the difference between more 'foundational' scientific research and translational endeavours. I therefore would not agree with O'Malley and Stotz when they conclude that translation is involved whenever techniques for intervention are transferred from one scientific context to another. In their definition, translation is involved, potentially to the same degree, in all three modes of integration which I consider here; while in my analysis, translation becomes a primary concern, with important consequences for how research is conducted and with which outcomes, when scientists commit to fulfilling specific social roles in the short term.

A key implications of the commitment to improving human health is that scientists engaged in translational integration need to pay attention to the *sustainability* of their research program - not only in the narrow sense of worrying about its financial viability, but also in the broader sense of considering the potential environmental and social impact of its outcomes. In practice, this typically involves engaging directly with contexts of production/use, so as to be able to assess the 'downstream' applicability of specific research strategies and prospective results. Crucially, biologists do not possess the right expertise to determine, by themselves, what counts as 'sustainable' research outcomes. This is why they need to collaborate with scientists in industry, state agencies and social scientists, among others, as it is through such engagements that scientists determine what constitutes 'human health' and how to improve it in the case at hand. Indeed, the 'social agenda' for translational research cannot be fixed, for the simple reason that it depends heavily on the ever-changing viewpoints and needs of the many stakeholders involved. Scientists who choose to take time to discuss the goal of their research with relevant parties outside the scientific world, and tailor their own research, tools and methods to fit those discussions, are investing a significant amount of their resources on producing results that might not be revolutionary in

terms of their conceptual contribution to existing biological knowledge (though they may prove to be such!), but rather are primarily meant to serve a wider social agenda.

Researchers involved in these exchanges are often also forced to compromise on their own views of what would constitute a productive research strategy and attractive research findings, in order to accommodate requirements and suggestions by other parties interested in achieving social, rather than scientific, goals. In particular, prioritizing the achievement of sustainable and efficient intervention (where what counts as sustainable and efficient is agreed upon among several different parties) over the acquisition of biological knowledge has important consequences for research practices, and particularly processes of integration, such as the choice of relevant data, and the speed with which data need to be collected and interpreted. This choice of priorities often comes at the expense of time dedicated to exploratory research, and yet this does not necessarily compromise the quality of research and its outcomes. On the contrary, the development of research strategies to pursue socially relevant goals, especially when it is coupled with the awareness that achieving such goals can sometimes be a long-term and complex endeavor, is a form of inquiry that philosophers of science should value and support.[19]

5. Conclusions

I have here identified and discussed three modes of integration at work in contemporary plant biology: inter-level, aimed at understanding organisms as complex wholes across a range of disciplinary approaches; cross-species, aimed at understanding organisms comparatively and

[19] In this sense, my approach is closely aligned with the socially relevant (philosophy of) science proposed by Helen Longino (2002), Janet Kourany (2010) and other leading philosophers interested in feminist philosophy and social studies of science as key sources of insight for the development of the philosophy of science. See for instance the *Synthese* special issue edited by Carla Fehr and Katherine Plaisance (2010); and the review symposium on Kourany (2010) in *Perspectives on Science* 2012, vol. 20, no. 3.

in relation to their environment; and translational, aimed at intervening effectively over organisms in order to improve human health. In all three cases that I examined, data integration is required to generate new knowledge, and curators of online databases play an important role in facilitating it. Collaborative research across groups of scientists is key to this process, as is the commitment to data sharing and re-use (even if such commitment is compatible with a large spectrum of research practices, and some of the parties involved would rather access data produced by others than share their own); and the dissemination of data through databases, as well as their integration within research, entails the standardization of embodied and propositional knowledge, as well as the development of new forms of intervention in and conceptualization of the biological world. Beyond these commonalities, however, each of these cases exemplifies a different way to integrate data, which prioritizes specific methods and prospective results over others, takes different types of data as its starting point, and poses distinct challenges to the curators of online databases. This means that achieving one of these types of integration does not necessarily provide the means to pursue another type: inter-level, cross-species and translational integration are the result of significantly different research strategies and related clusters of instruments, concepts, methods, materials and data.

Inter-level integration has been particularly prominent within 20th century plant molecular biology. This form of integration involves data produced by different subdisciplines on different levels of organization of the same plant species. Thus, most of the research efforts focus on finding ways to overcome disciplinary barriers, such as differences in methods and terminology between molecular and cellular biology, in order to collect and visualize those data within a single framework; and biologists involved in those efforts have tended to prioritize mechanistic understandings of organisms over the study of biodiversity. By contrast, the main challenge in cross-species integration is to find ways of making methods,

terminologies, and material differences involved in the study of two different species (rather than disciplines) compatible with each other; and in exploring the reasons for the similarities and differences between species, which involves taking account of the evolutionary and environmental context of the organisms at hand. Finally, the bulk of research efforts within translational integration goes into the assessment, in dialogue with a wide range of expertises beyond plant science itself, of the sustainability and social impact of the methods and potential outcomes of research. This form of research might not necessarily yield new insights on the biology of organisms, or their evolutionary or developmental history; however, it does yield an understanding of how scientific research needs to be set up and developed so as to yield socially desirable outcomes. Given the context-dependent and ever-changing nature of what counts as 'socially desirable', this involves the constant re-assessment of which data are most relevant to achieving such goal, which expertise's are involved in producing such data, and which methods and data infrastructures are used to disseminate, visualize and interpret them.

It is important to note that the typology proposed here is not meant to be exhaustive of all the ways in which data integration can happen in plant research. These forms of integration are also not meant to be mutually exclusive, and very often they happen alongside each other, and in dialogue with each other, within the same scientific laboratory (as my examples have shown). Most research groups that I have come across are interested and involved in all three types of integration. Indeed, the interplay between those three modes is often crucial to the scientific success of a lab, especially at a time when both scientific excellence and the social impact of research are highly valued by funding bodies across the globe, and heated discussions surround the choice of metrics to assess how scientists fare on these two counts. The development of standards enabling the integration of data across species has been a key step in the development of model organism databases, thus signaling the willingness of

researchers to move beyond inter-level integration and towards cross-species comparisons. Similarly, both inter-level and cross-species integration routinely generate results on which new forms of translational integration can be built. The very case of *Miscanthus* might work in this way if plant scientists actively engage in the search for efficient and sustainable ways to downstream the production of bioethanol, and use these interactions to inform their own bioengineering practices (which seems to be exactly what plant scientists are starting to do, for instance through initiatives such as the UK Plant Science Federation in the UK). There are even cases where all three types of integration are attempted simultaneously, such as the current effort to combine transcriptomic data obtained from large groups of plants (an ideal case of cross-species integration) with metabolic profiling, functional genomics, and systems biology approaches (inter-level integration) so as to reveal 'entire pathways for medicinal products', in ways that promise to revolutionize drug discovery and thus provide a perfect instance of translational integration (de Luca et al 2012, 1660).

Given the multiple and complex interrelation between the three forms of integration that I have identified, one might wonder why it matters to distinguish them at all. The reason has to do with improving existing philosophical understandings of scientific practices and of the temporality and constraints within which research is carried out. Even if these three forms of integration are intertwined in the overall vision of what science is supposed to achieve for humanity, and in the overall trajectory of any one specific research group, their distinctive aims, methods, strategies and norms require that they are taken up to different degrees at any one point in time. All the plant scientists whose work I have discussed here are interested in understanding the biology of specific plants as an integrated whole; comparing different species so as to reach as encompassing an understanding of the plant kingdom as possible (including its evolution, inner diversity and environmental role); and eventually using their research results to address key challenges to human life in the 21st century, such as climate

change, urbanization and population increase (the 'water-food-energy' nexus, as defined by the World Economic Forum; 2011, p.8). Yet, *they are typically unable to pursue all of these goals in equal measure at the same time*;[20] and the choices that they make when considering which data to view as relevant to their research, how to integrate those data and which expertises to involve in that process will be crucial factors in determining which form of knowledge they prioritize as the primary outcome of their efforts. In other words: the pursuit of different forms of integration gives rise to different forms of scientific knowledge, whose value and content shifts in relation to the goals, expertises and methods involved in each research project.

Through this analysis, I hope to have shown the relevance of the infrastructure and standards used to integrate data to achieving different epistemic goals and thus different forms of knowledge; and, at the same time, that prioritizing specific epistemic goals over others might lead to structuring data integration, and the infrastructures and standards used to that effect, in different ways. Considering the procedures and standards developed to facilitate data integration provides important clues about the norms, practices and implications of integrative processes, and the epistemic role of the social and institutional contexts in which such efforts take place. Data integration and the production of scientific knowledge, both propositional and embodied, are strictly intertwined: a crucial question for both scientists and philosophers is exactly in which ways do the worlds of data infrastructures and knowledge production inform each other, and how institutional contexts and epistemic goals affect the development of data integration strategies in contemporary biology.

Acknowledgments

[20] This is something that many funding bodies, particularly those subscribing to the 'Impact Agenda' in the UK and elsewhere, are reluctant to acknowledge.

Warm thanks go to Maureen O'Malley, Paul Griffiths and Karola Stotz, who invited me to the excellent conference on integration (Sydney, April 2012) for which this paper was first drafted, and particularly to Maureen for countless helpful and inspiring discussions. I am also greatly indebted to Ingo Brigandt, three anonymous referees and David Studholme for very thoughtful and useful comments on a previous draft; Kaushik Sunder Rajan, who provided insightful comments at a crucial moment of the writing; participants to the Sydney conference and the Knowledge/Value conference hosted by the University of Chicago Centre in Beijing, where I presented a revised version in September 2012; and to my Exeter colleagues Staffan Müller-Wille and John Dupré, whose provocative ideas continue to be a great source of inspiration. Last but not least, thanks to the plant scientists who welcomed me into their labs and societies, and took time to discuss their fascinating work with me.

Bibliography

Avraham, S., Tung CW, Ilic K, Jaiswal P, Kellogg EA, McCouch S, Pujar A,Reiser L, Rhee SY, Sachs MM, Schaeffer M, Stein L, Stevens P, VincentL, Zapata F, Ware D. (2008). The Plant Ontology Database: a community resource for plant structure and developmental stages controlled vocabulary and annotations. *Nucleic Acids Research,* 36, D449-54.

Babcock, B. A. (2011) *The Impact of US Biofuel Policies on Agricultural Price Levels and Volatility*. ICTSD Issue Paper No. 35.

Bechtel, W. (this issue). From molecules to behavior and the clinic: integration in chronobiology. *Studies in History and Philosophy of Biological and Biomedical Sciences*. doi:10.1016/j.shpsc.2012.10.001

Bechtel, W. & Richardson, R. C. (2010). *Discovering complexity: Decomposition and localization as strategies in scientific research*. Second Edition. Cambridge, MA: MIT Press/Bradford Books

Bevan, M. & Walsh, S. (2004). Positioning Arabidopsis in plant biology. A key step toward unification of plant research. *Perspectives on Translational Biology,* 135, 602-606.

Bolker, J.A. (1995). Model systems in developmental biology. *BioEssays,* 17, 451–5.

Botanical Society of America (1994). *Botany for the Next Millennium: I. The intellectual: evolution, development, ecosystems*. http://www.botany.org/bsa/millen/mil-chp1.html#Evolution Accessed 09/08/2012.

Bowker, G. (2000) Biodiversity datadiversity. *Social Studies of Science*, 3(5), 643-683.

Brigandt, I. (2010). Beyond reduction and pluralism: toward an epistemology of explanatory integration in biology. *Erkenntnis,* 73, 295–311.

Brigandt, I. (2013). Explanation in biology: reduction, pluralism, and explanatory aims. *Science & Education* 22, 69–91.

Brigandt, I. (this issue). Systems biology and the integration of mechanistic explanation and mathematical explanation. *Studies in History and Philosophy of Biological and Biomedical Sciences*.

Browne, J. (2001). History of Plant Sciences. In: *Encyclopedia of Life Science*. John Wylie and Sons. Accessed on 9/08/2012 at http://web.fc.uaem.mx:8080/tareas/55_HistoryPlSc.pdf

Bülow L, Schindler M, Hehl R. (2007). PathoPlant®: a platform for microarray expression data to analyze co-regulated genes involved in plant defense responses. *Nucleic Acids Research*, 35, D841-845.

Bult, C. J. (2006). From information to understanding: the role of model organism databases in comparative and functional genomics. *Animal genetics* 37 Suppl 1, 28–40.

Burian, R. (1997.) Exploratory experimentation and the role of histochemical techniques in the work of Jean Brachet, 1938-1952. *History and Philosophy of the Life Sciences,* 19: 27-45.

Callebaut, W. (2012) Scientific perspectivism: A philosopher of science's response to the challenge of big data biology. *Studies in the History and the Philosophy of the Biological and Biomedical Sciences*, 43(1), 69-80.

Chang, H. (2012). *Is Water H20? Evidence, Realism and Pluralism*. Springer.

Craver, C. F. (2005). Beyond reduction: Mechanisms, multifield integration and the unity of neuroscience. *Studies in the History and Philosophy of Biological and Biomedical Sciences, 36*, 373-395.

Darden, L. and Maull, N. (1997). Interfield theories. *Philosophy of Science,* 44(1), 43-64.

Darden, L. (2005). Relations among fields: Mendelian, cytological and molecular mechanisms. *Studies in History and Philosophy of Biological and Biomedical Sciences,* 36, 349-371.

Davis, R.H. (2004). The age of model organisms. *Nature Reviews Genetics,* 5, 69–76.

De Luca et al. (2012). Mining the biodiversity of plants. A revolution in the making. *Science,* 336, 1658-1661.

Douglas, H. (2009). *Science, Policy and the Value-Free Ideal*. Pittsburgh: Pittsburgh University Press.

Elliott, K. (2011). *Is a Little Pollution Good for You? Incorporating Societal Values in Environmental Research*. New York: Oxford University Press.

Fehr, C. and Plaisance, K. (2010) (Eds.). *Making Philosophy of Science More Socially Relevant*. Special issue of *Synthese* (Volume 177, Issue 3).

Grantham, T. (2004). Conceptualising the (dis)unity of science. *Philosophy of Science,* 71(2), 133-155.

Griesemer, J.R. (2006) Theoretical integration, cooperation, and theories as tracking devices. *Biological Theory,* 1, 4-7.

Hacking, I. (1992) The self-vindication of the laboratory sciences. In Pickering, A. (Ed.) *Science as Practice and Culture* (pp.29-64). The University of Chicago Press.

Harwood, J. (2012). *Europe's Green Revolution and Others Since.* London: Routledge.

Hey, T., Tansley, S. & Tolle, K. (Eds.) (2009). *The Fourth Paradigm. Data-Intensive Scientific Discovery*. Redmond, Washington: Microsoft Research. http://research.microsoft.com/en-us/collaboration/fourthparadigm

Ideker, T., Bafna, V. & T. Lemberger (2007). Integrating scientific cultures. *Molecular Systems Biology* 3, 105.

Jensen, E.F. et al (2011) Characterisation of flowering time diversity in Miscanthus species. *GCB Bioenergy*, 3: 387-400.

Jensen, E.F. (2007) Flowering time diversity in Miscanthus: A tool for the optimization of biomass. Abstracts of the Annual Main Meeting of the Society for Experimental Biology. Consulted on 7 December 2012 on http://hdl.handle.net/2160/4532

Keating. P. & Cambrosio, A. (2012). Too many numbers: Microarrays in clinical cancer research. *Studies in the History and the Philosophy of the Biological and Biomedical Sciences,* 43(1), 37-51.

Keating, P. & Cambrosio, A. (2012) Cancer on Trial: Oncology as a New Style of Practice. Chicago University Press.

Kelling, S., Wesley M. Hochachka, Daniel Fink, Mirek Riedewald, Rich Caruana, Grant Ballard & Giles Hooker (2009). Data-Intensive Science: A New Paradigm for Biodiversity Studies. *BioScience*, 59(7), 613-620

Kingsbury, N. (2009). *Hybrid – The History and Science of Plant Breeding.* Chicago: University of Chicago Press.

Kohler, R.E. (1994). *Lords of the Fly*: *"Drosophila" Genetics and the Experimental Life.* Chicago: University of Chicago Press.

Koornneef, M. & Meinke, D. (2010). The development of Arabidopsis as a model plant. *The Plant Journal,* 61, 909–921.

Kourany, J. (2010) *Philosophy of Science after Feminism*. Oxford: Oxford University Press.

Lamesch, P., Tanya Z. Berardini, Donghui Li, David Swarbreck, Christopher Wilks, Rajkumar Sasidharan, Robert Muller, Kate Dreher, Debbie L. Alexander, Margarita Garcia-Hernandez et al (2011). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research*, 1-9.

Leonelli S. (2007). Growing weed, producing knowledge. An epistemic history of Arabidopsis thaliana. *History and Philosophy of the Life Sciences,* 29(2), 55–87.

Leonelli, S. (2009a). On the locality of data and claims about phenomena. *Philosophy of Science,* 76(5), 737-749.

Leonelli, S. (2009b). The impure nature of biological knowledge. In de Regt, H, Leonelli, S. & Eigner, K. (Eds.) *Scientific Understanding: Philosophical Perspectives*. Pittsburgh: Pittsburgh University Press, pp. 189-209.

Leonelli, S. (2010a). Documenting the emergence of bio-ontologies: or, why researching bioinformatics requires HPSSB. *History and Philosophy of the Life Sciences,* 32(1), 105-126.

Leonelli, S. (2010b). Packaging data for re-use: Databases in model organism biology. In Howlett, P. & Morgan, M. S. (Eds.). *How Well Do Facts Travel? The Dissemination of Reliable Knowledge*. Cambridge, MA: Cambridge University Press, pp.325-348.

Leonelli, S. (2012a). 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, S. (2012b). Classificatory theory in data-intensive science: The case of open biomedical ontologies. *International Studies in the Philosophy of Science,* 26(1), 47-65.

Leonelli, S. & Ankeny, R.A. (2012) Re-thinking organisms: The impact of databases on model organism biology. *Studies in History and Philosophy of Biological and Biomedical Sciences,* 43, 29–36.

Leonelli, S., Charnley, B, Webb, A & Bastow, R. (2012) Under one leaf. A historical perspective on the UK Plant Science Federation. *New Phytologist,* 195(1), 10-13.

Lenoir, T (1999) Shaping Biomedicine as an Information Science. In Mary Ellen Bowden, Trudi Bellardo Hahn, and Robert V. Williams (Eds.) *Proceedings of the 1998 Conference on the History and Heritage of Science Information Systems* (pp.27-45). ASIS Monograph Series. Medford, NJ: Information Today, Inc.

Longino, H. (2002). *The Fate of Knowledge*. Princeton University Press.

McCain, K. W. (1995). Mandating sharing: Journal policies in the natural sciences. *Science Communication,* 16(4), 403-431.

Mitchell, S. (2003). *Biological Complexity and Integrative Pluralism*. Cambridge University Press.

Mitchell, S. (2009). *Unsimple Truths*. Chicago University Press.

Morgan, M.S. and Morrison, M. (1999) *Models as Mediators*. Cambridge, UK: Cambridge University Press.

Müller-Wille, S.W. (2007). Collection and collation: theory and practice of Linnaean botany. *Studies in the History and Philosophy of the Biological and Biomedical Sciences,* 38, 541-562.

Nordmann, A., Radder, H. & Schiemann, G. (2011). *Science Transformed? Debating Claims of an Epochal Break*. Pittsburgh: Pittsburgh University Press.

O'Malley, M.A. (2011). Exploration, iterativity and kludging in synthetic biology. *Comptes Rendus Chimie*, 14(4), 406–412.

O'Malley, M. A. (this issue). When integration fails: Prokaryote phylogeny and the tree of life. *Studies in History and Philosophy of Biological and Biomedical Sciences*. doi:10.1016/j.shpsc.2012.10.003

O'Malley, M.A. & Soyer, O. (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.

O'Malley, M.A. & Stotz, K. (2011). Intervention, integration and translation in obesity research: Genetic, developmental and metaorganismal approaches. *Philosophy, Ethics, and Humanities in Medicine*, 6(2).

Plutynski, A. (this issue). Cancer and the goals of integration. *Studies in History and Philosophy of Biological and Biomedical Sciences*.

Potter, C., Harwood, T., Knight, J. & Tomlinson, I. (2011). Learning from history, predicting the future: the UK Dutch elm disease outbreak in relation to contemporary tree disease. *Phil. Trans. R. Soc. B,* 366, 1966-1974.

Quirin, E.A. et al (2012). Evolutionary meta-analysis of Solanaceous Resistance Gene and Solanum Resistance Gene analog sequences and a practical framework for cross-species comparisons. *Molecular Plant-Microbe Interactions,* 25(5), 603-612.

Royal Society (2012). *Science as an Open Enterprise*. http://royalsociety.org/policy/projects/science-public-enterprise/report/ Accessed 10 August 2012.

Zhang, P., Foerster, H., Tissier, C., Mueller, L. Paley, S. Karp, P and Rhee, S.Y. (2005). MetCyc and AraCyc: Metabolic pathway databases for plant research. *Plant Physiology*, 138:27-37.

Rhee, S.Y., Beavis, W., Berardini, T.Z., et al. (2003). The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Research,* 31(1), 224–228.

Rhee, S.Y. (2004). Carpe diem: Retooling the "publish or perish" model into the "share and survive" model. *Plant Physiology,* 134, 543–547.

Rheinberger, H. (1997). *Towards a History of Epistemic Things*. Stanford, CA: Stanford University Press.

Rheinberger, H. (2010). *An Epistemology of the Concrete*. Durham & London: Duke University Press.

Rheinberger, H. (2011). Infra-experimentality: from traces to data, from data to patterning facts. *History of Science,* 49, 337-348.

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

Stitt, M., Lunn, J. and Usadel, B. (2010). Arabidopsis and primary photosynthetic metabolism – more than the icing on the cake. *Plant Journal*, 61(6): 1067-91.

Strasser, B.J. (2011). The experimenter's museum: GenBank, natural history, and the moral economies of biomedicine, 1979-1982. *Isis,* 102, 60—96.

Taylor, C. F., Field, D., Sansone, S., Aerts, J., Apweiler, R., Ashburner, M., Ball, C. A., et al. (2008). Promoting coherent minimum reporting guidelines for biological and biomedical investigations: The MIBBI project. *Nature biotechnology*, 26(8), 889–896.

The International Arabidopsis Informatics Consortium (2012). Taking the next step: Building an Arabidopsis information portal. *Plant Cell*, 24(6), 2248-2256.

Smocovitis, V. (2006). Keeping up with Dobzhansky: G. Ledyard Stebbins, Jr., Plant Evolution, and the Evolutionary Synthesis. *History and Philosophy of the Life Sciences*, 28(1): 11-50.

World Economic Forum (2011). *Global Risks 2011 Sixth Edition*. Available at http://reports.weforum.org/global-risks-2011/ . Accessed on 8/08/2012.