Paper published in Biology and Philosophy, 2008

Performing Abstraction: Two Ways of Modelling *Arabidopsis thaliana*

Sabina Leonelli

Research Officer
How Well Do 'Facts' Travel?
(Leverhulme/ESRC)

Faculty of Economic History
London School of Economics
Houghton Street
WC2A 2AE
United Kingdom
Tel. 0044 20 7955 7858
Fax. 0044 20 7955 7730

s.leonelli@lse.ac.uk

Abstract

What is the best way to analyse abstraction in scientific modelling? I propose to focus on abstracting as an *epistemic activity*, which is achieved in different ways and for different purposes depending on the actual circumstances of modelling and the features of the models in question. This is in contrast to a more conventional use of the term 'abstract' as an *attribute* of models, which I characterise as black-boxing the ways in which abstraction is performed and to which epistemological advantage. I exemplify my claims through a detailed reconstruction of the practices involved in creating two types of models of the flowering plant *Arabidopsis thaliana*, currently the best-known model organism in plant biology. This leads me to distinguish between two types of abstraction processes: the 'material abstracting' required in the production of Arabidopsis specimens and the 'intellectual abstracting' characterising the elaboration of visual models of Arabidopsis genomics. Reflecting on the differences between these types of abstracting helps to pin down the epistemic skills and research commitments used by researchers to produce each model, thus clarifying how models are handled by researchers and with which epistemological implications.

Keywords

abstraction, *Arabidopsis thaliana,* commitments, modelling, model organisms, skills, understanding

> *"The recognition of a person in the performance of a skill or in the conduct of a game of chess is intrinsic to the understanding of these matters. We must surmise that we are faced with some coordinated performance, before we can even try to understand it, and must go on trying to pick out the features that are essential to the performance, with a view to the action felt to be at work with it"*
>
> *Polanyi 1962: 30*

Much has been written on the way in which models are produced and used to acquire theoretical knowledge about biological phenomena. Model-based reasoning, broadly defined as the use of representational tools to gain epistemic access to natural processes, is now widely recognised as playing a crucial role in the production of scientific knowledge. Further, it is acknowledged that models come in an endless variety of forms, a combination of which is always required by their use in scientific practice.[i] Given this dramatic diversity, and the resulting 'promiscuity'[ii] of the notion of model itself, much attention has been paid to the actual features of models employed in scientific practice, so as to clarify the epistemological status of each type of model as both a product of and a tool towards scientific theorising. Relatively less attention has been devoted to the variety of activities that need to be performed in order to yield adequate models. Scientists do not only refer to models in their explanations: they use them, manipulate them and modify them all the time in order to achieve and justify those very explanations. The adequacy of such use is determined both by the features of the phenomena under scrutiny and by the material, social and institutional setting and commitments of the researchers involved. Focusing primarily on modelling practices, rather than on models thus produced, might prove a useful way to gain insight on some long-standing debates within the philosophy of scientific modelling and representation. I here develop this approach in relation to one such issue. This is the interpretation of the notion of *abstraction*, particularly in the context of the putative distinction between 'abstract' and 'material' models. I address the debate on what it means for a model to be abstract by focusing on the processes required to produce an abstract model. This implies a shift from thinking of abstraction as a property of models ('abstract' as an attribute) to considering abstraction as an activity required to produce models (the verb 'to abstract' as a way of acting) – a shift which, as I shall argue, throws light on the epistemic skills and theoretical commitments required to gain understanding of the natural world.

My analysis begins with the detailed reconstruction of two types of modelling currently employed in the study the flowering plant *Arabidopsis thaliana*. Arabidopsis is the most successful model organism within contemporary plant biology, and possibly the best-known organism within biology as a whole, with the exclusion of *Homo sapiens*. Around 16.000 scientists, based at more than 6.000 laboratories around the globe, are using this plant to pursue very different types of research programmes, ranging across levels of organisation (e.g. from the ecological to the molecular) and the different biological processes that it embodies (e.g. its flowering system, cell-to-cell transmission processes and receptivity to light).[iii] Remarkably, in the face of institutional and disciplinary disunity, the research efforts of the Arabidopsis community are centrally co-ordinated by a steering committee, whose long-term goal is to use the results of Arabidopsis research to study the biology of this organism as an 'intact whole'.[iv] In practical terms, this implies the centralised

production of tools that enable researchers to explore the connections between the knowledge acquired by different teams about each part of the plant. My analysis centres on two such tools, that is, on two types of modelling practices developed to facilitate the integration of knowledge about Arabidopsis. My reconstruction of the steps required to create models used at *The Arabidopsis Information Resource* (section 2) and the *Nottingham Arabidopsis Stock Centre* (section 3) is based on my own interpretation of the models, as shaped by data and experiences I gathered while visiting these two sites (such as interviews with the directors and personnel of the two centres; archival material pertaining to the relation of research conducted within the centres to the various research practices characterising the rest of the Arabidopsis community; and, most importantly, participation in the actual processes of creating and modifying models that I describe in what follows).[v]

1. *TAIR models*

The first type of modelling consists in the activities performed by a research team based at the Carnegie Institute for Plant Biology (Stanford, CA) in order to yield digital representations of the structure and possible functions of Arabidopsis' genes. These two-dimensional models are made freely available to any interested scientist through The Arabidopsis Information Resource [TAIR], one of the major online depositories of data and resources of relevance to Arabidopsis research (URL: http://www.arabidopsis.org).[vi]

TAIR is especially interesting for my purposes, because it has the explicit aim of eventually absorbing and giving access to *all* of the available data about the plant, thus making them accessible to biologists with very different specialisations. Indeed, TAIR is constructed to serve as many as possible of the (often conflicting) needs expressed by its users. Researchers evaluate TAIR databases in terms of the legibility of the images that they produce; their accessibility to researchers without a background in informatics; their flexibility in response to the specific queries of the users and, most important of all, the degree of efficiency with which they collect, store and categorise the unwieldy amount of data hitherto accumulated on Arabidopsis. All these demands weigh heavily on the daily activities of the research team maintaining TAIR: they effectively transform TAIR into a virtual laboratory for the creation of widely usable digital models of the plant.

Up to now, TAIR has been able to construct visual representations of data concerning the lowest levels of organisation in Arabidopsis.[vii] These include tools for the visualisation of its genetic map, its gene expression data and its metabolic cycle.[viii] Creating the models requires working through the following steps:

(1) *Definition of organisational categories:* TAIR databases and visualisation tools are constructed with the help of Java software. The basic set-up of Java programming requires the initial selection of a set of objects and family relations holding among them. These objects and relations are arranged into so-called DAG structures ('directed acyclic graphs', as illustrated in figure 1). Programmers refer to this approach as 'object-oriented'. Biologists have adapted this way of organising data to the idea that each biological system can be investigated via partitioning – that is, by locating its components and the

possible relations among them, depending on the specific mechanism or process requiring explanation. Thus, TAIR starts its modelling activity by selecting *data types* to be represented by the programme as 'objects' of the DAG.

[FIGURE 1]

(2) *Definition of Gene Ontology [GO]:* Scientists constructing databases centred on model organisms have long realised the parallel between biological concepts and the object-to-object relations imposed by database software. To exploit this factor more explicitly, as well as to coordinate the terminology and conceptual framework used across different databases, these scientists have teamed up on a project called 'Open Biomedical Ontologies'.[ix] This project is producing a series of conceptual maps for database construction, generally referred to as bio-ontologies (Bard and Rhee 2004). Those maps are constituted by appropriately defined[x], basic biological concepts which relate to each other as 'part_of' one another, structural and/or functional equivalents ('is_a' relation) or 'dependent_on' one another (figure 2). Each bio-ontology covers a different set of biological objects. The GO, that is the framework adopted by TAIR to structure and visualise Arabidopsis data, starts from gene-related entities (such as gene products and expression) and proceeds to link them to higher-level objects (e.g. metabolic and developmental pathways).

[FIGURE 2]

(3) *Collection of data*: on the basis of the requirements and conceptual order determined by DAG structures and GO networks of concepts, TAIR researchers seek out publicly available data with which to fill each category (figure 3). This involves delving into Arabidopsis literature and selecting publications containing the most up-to-date and complete information about Arabidopsis biology. TAIR programmers have devised a programme called 'PubSearch' to track published material relevant to any specific gene product and select results that fit the available categories. To guarantee that users will be able to trace the experimental context in which data have been obtained, the results are always accompanied by a reference to the sources of the information used.

[FIGURE 3]

(4) *Fitting out*: once the relevant data are collected and matched to the appropriate organisational categories, TAIR programmers devise a 'schema' (figure 4): that is, a way of transforming those data and the relations among them into standard formats visualised through Java.[xi] This also involves fitting data contents to the terminology and relations determined by GO.

[FIGURE 4]

(5) *Design*: trial-and-error phase in which schemas are implemented to create technically and visually satisfying visualisations of the data. The factors involved in this phase are both aesthetic and pragmatic: visualisations have to

be legible, simple and, at the same time, as accurate as possible with respect to the data as originally collected; At the same time, they have to be digitally feasible, that is, realisable via the software at hand.

(6) *Quality Assessment*: As stated above, the main goal of TAIR is to provide scientists with digital models of Arabidopsis molecular biology that are as accessible and useful as possible. To achieve this, TAIR personnel, in consultation with a number of experimentalists, elaborate what they call a 'vision' of how to 'serve the user'. This involves thinking of what a data set should look like for a biologist to get the best information available with the smallest effort. The models should provide answers to questions that biologists actually ask, such as 'what is the relation between this gene cluster and this specific pathway'. An example of such 'vision' is the idea that one should be able to 'fly into' a chromosome, thus accessing a huge amount of information just by clicking on an image of the chromosome. This idea was implemented by constructing a 3D model of a chromosome that could be viewed from all directions by moving the mouse so as to fly around and into it (much in the same way as exploring a virtual space in a videogame).

(7) *Implementation*: the new images and data are added to the material available online (which might be modified to accommodate the most recent models). Figure 5 provides an example of a finished TAIR model representing the metabolic pathway of Arabidopsis.

[FIGURE 5]

Note that most activities required by TAIR modelling involve manual curation - that is, a step-by-step, case-based choice of the appropriate parameters, programmes and visualisation strategies for each group of data. Despite the efforts put into making curation as automated as possible, the choice and insertion of Arabidopsis data into TAIR databases requires the exercise of tacit skills in selecting relevant data and the ability to interpret the concepts used in GO.

2. NASC models

The second type of modeling is performed in the two stock centres responsible for the storage, maintenance, production and distribution of Arabidopsis plants (ecotypes and mutants) on a global scale. I focus on the work carried out by the Nottingham Arabidopsis Stock Centre [NASC], which started to build its collection of Arabidopsis specimens in the early 1990s from stocks originally maintained by the pioneers of research on the plant in the 1920s and 30s. By the late 1980s (a decade after the launch of Arabidopsis as a favourite model in plant molecular biology), the Arabidopsis steering committee had to acknowledge that the standards by which those collections were maintained needed drastic improvement (Meinke and Scholl, 2003). This was mainly due to two factors. First, a standardised and controlled classification of Arabidopsis ecotypes was crucial to the co-ordination of research projects required by the highly centralised Arabidopsis research. One of the main advantages of working on Arabidopsis is its high degree of transformability in the lab. Given the number of available mutants of Arabidopsis, it is important to ensure that laboratories

sharing, replicating and building upon each other's data across the world are not mixing up data from different specimens. The second reason for improving stocking facilities was that different strands of research are associated with specific ecotypes. For instance, the Arabidopsis Genome Initiative [AGI] was entirely based on the *Columbia* ecotype [Col]. As a result, much research on functional genomics keeps using Col, in order to keep as close a match as possible between its specific morphological traits and its genetic make-up. Another ecotype, the *Wassilenska* [Ws], rose to fame because of its putative high transformability – a reputation enhanced by its central role in the revolutionary discovery of how to transform Arabidopsis by exposing it to appropriately modified *Agrobacterium*. Despite current knowledge that Ws is actually no more transformable than other Arabidopsis ecotypes, it is still a favourite variant for mutagenesis experiments.

The increasing importance of distinguishing among ecotypes, as well as of acquiring precisely the ecotype or mutant relevant to one's experiments, determined the establishment, in 1991, of the Arabidopsis stock centres. NASC has been successfully fulfilling the role of producer and distributor of specimens, by elaborating increasingly sophisticated ways of categorising, growing and handling them depending on the needs and experience of laboratories around the world. The production of specimens requires the following, complex sequence of activities:

(1) *Acquisition of seeds*: the NASC receives a sample of seed, together with a free-text (i.e. non-standardised) description of their genetic and morphological characteristics, from a donor (Arabidopsis laboratory or individual researcher);

(2) *Preliminary Classification*: NASC re-phrases the donor's description of specimens into a standardised description. This is relatively easy to do with the plant's genetic make-up (thanks also to the use of GO). The plants' morphology is more problematic, given the great diversity and inaccuracy characterising descriptions of plants at that level. NASC developed its own system of classification, the Phenotype, Attribute and Trait Ontology [PATO] (figure 6). Like Go, this system uses a child-parent network structure; unlike GO, it categorises observations made through direct physical interaction with the specimens ('attributes' describing morphological features - 'traits' - of the organism, or 'phenotype').

[FIGURE 6]

(3) *Cultivation of specimens*, including the processes of ge*rmination*, *growth* of plants (around 25 per type, each producing around 10.000 seeds), *drying* plants as required for the extraction of seeds and *harvesting* of the seeds through pulverisation of the dried plant. As noticeable from the photograph depicting the glass-house (figure 7), these processes are carried out in controlled environments equipped with temperature and humidity controls as well as isolation systems preventing cross- contamination: the growth of plants has been automated as much as possible. Still, cultivation requires a great amount of interventions by NASC personnel. NASC researchers need to evaluate and intervene on the harvest multiple times per day, to secure that the specimens grown in their facilities conform to the expectations of the researchers experimenting on them.

[FIGURE 7]
[FIGURE 8]
[FIGURE 9]

(4) *Final Classification*: both the genetic make-up and the morphological features of the plants are double-checked to ensure that the information used to label the seed samples are accurate.

(5) *Storage*: after seeds have been cleaned and sterilised, some of them are frozen into a 'seed archive' (so that any classified type of seed is always retrievable); the rest is stored in appropriately labelled packages, ready for distribution.

(6) *Distribution*: seeds are put into sterile containers and shipped to users upon request, together with guidelines on how to grow and handle them.

This extensive list of interventions highlights the fact that the transfer of Arabidopsis specimens from the wild to the laboratory involves more than a change of context (though the latter certainly has great impact on the specimens). Specimens need to conform to the expectations of researchers that intend to experiment upon them. They have to display features that are adequate to the research procedures and instruments in use. Obtaining specimens that conform to these expectations requires apposite structures, standardised tools and guidelines and extensive experience in handling the plants: it is a matter of skilful production, rather than mere displacement. By the time that a seed sample is labelled and sent off to a user, researchers have achieved a high degree of control over the plants that will grow from those seeds. These plants are modified to acquire and preserve specific phenotypic and/or genotypic features that are seen by researchers as representative of Arabidopsis wild types. The resulting plants are hybrids of 'wild' traits and traits that are well controlled by experimenters and thus stabilised and reproducible. They can usefully be characterised as *domesticated samples*: on one hand, they are artefacts that are purposefully manipulated so as to become representative of a whole class of organisms (including other variants of Arabidopsis, as well as many other plants, depending on the research context); on the other hand, they are only partially a product of human intervention, as they remain samples of the very phenomenon that they are taken to represent. This ambivalence is a most useful characteristic of NASC specimens, which makes them into three-dimensional *models of Arabidopsis plants*.

3. 'Abstract Models'? Three Interpretations

Let us now pose the analytic question, whether either of these two types of models could be classified as abstract simply by looking at its features. I shall argue that the term abstract can be used to capture at least three different intuitions, and that the content and significance of our classification of models as abstract change depending on which of these intuitions we focus on.[xii]

One intuition indicates whether models are embodied in objects that can be perceived via our senses. In this first sense, the adjective 'abstract' figures as the opposite of

'concrete' and 'material': something is abstract when it is not tangible and/or visible, as in the case of a mental construct or an imaginary object for instance. This is an absolute notion: there are no 'degrees' of abstraction, insofar as an entity is either tangible or not. According to this interpretation, neither TAIR images nor NASC specimens are abstract, since they are both embodied into an image on a computer screen (the former) and an actual organism (the latter).

A second interpretation concerns the amount of information conveyed by a model with regard to its physical meaning or applicability, that is, the way in which it relates to the phenomena that it is taken to represent. Thus, a model is the more abstract, the more it is devoid of physical meaning.[xiii] Notably, this notion of abstraction admits of degrees, which can be measured by the amount of additional information needed for a model to be applicable to the analysis of a particular phenomenon. This second intuition conflicts with the previous one both in its motivation and in its implications. There is a clear difference in the amount of additional information needed to relate TAIR models to the appropriate component of Arabidopsis plants, as compared to NASC models (which are as highly endowed with physical meaning as possible, being actual samples of the plants themselves). To understand the empirical content of TAIR images, one needs to refer heavily both to GO categories and to the object-oriented approach to the organisation of data (as well as, in some cases, the circumstances and experiments yielding the evidence incorporated within each image). In this respect, TAIR images can rightly be claimed to be abstract, and certainly to be more abstract than NASC specimens.

The third sense in which a model can be abstract refers to the number of phenomena that it is taken to represent.[xiv] Even more strongly than in the previous case, 'abstract' is here a relative notion whose application depends on the context: a model is the more abstract, the more specific situations it can be taken to stand for. Within this third perspective, both NASC and TAIR models can be thought of as concrete or abstract, depending on the research goals of the scientists handling them. Prima facie, NASC models qualify as very concrete models, since they are only representative of the specific class of Arabidopsis ecotypes defined by their particular morphological and genetic make-up. However, many molecular biologists use a specific ecotype as a representative of any Arabidopsis variant - or even, in some cases, of any plant or living organism: in this case, NASC models are highly abstract. Similarly, the TAIR image of a metabolic pathway can be taken to represent any similar metabolic pathway found on any organism – thus 'fitting' a large amount of particulars. The same image can also be taken to represent a pathway typical of Arabidopsis plants, or of a specific ecotype of Arabidopsis. In this latter case, the degree of abstraction characterising the TAIR model is lower than in the former.

4. Different Ways of Abstracting

I intend to argue for an alternative way to approach the debate on abstract modeling. This is to shift the very terms in which the debate is run by focusing on the *processes* required to produce a model. This approach is not intended as an alternative to current understandings of what it means for a model to be abstract. Rather, it complements and enriches existing literature by examining the meaning of abstracting in relation to the production of models as representations of phenomena.

I propose to view abstraction as *the activity of selecting some features of a phenomenon P, as performed by an individual scientist within a specific context, in order to produce a model of (an aspect of) P.* This implies defining the process of abstracting as involving the transformation of some features of a phenomenon into parameters used to model it, *as relevant to the specific aspect of biology that the model is deployed to study*. This is not a case of mere parameterisation or standardisation – that is, following Bowker and Star's definition (1999: 13), of constructing consensus around rules governing the production of objects. As I have shown, producing objects with similar, reproducible features is essential to the realisation of the process. Yet, what justifies the characterisation of this process as one of abstracting is the goal towards which objects are produced: that is, to function as models of the phenomenon in question within a given research context. The choice of the features of P to be abstracted is based on the epistemic function expected of the model thus obtained.

Defined in this way, abstraction is one of the processes required in creating a model, rather than an attribute of the model itself (the model thus being 'abstracted' in various ways depending on the specific circumstances and research goals, rather than 'abstract' in an absolute sense). Further, it is an essential process in the context of modelling practice, as it is the process by which all types of models acquire a representational value with respect to some aspects of a phenomenon. Theory can play different roles in this process: the selection of features of P can be entirely based on theoretical assumptions (and thus make no direct reference to actual observations and interactions with P) or it can be largely independent of any theory, as it is based purely on a researcher's proto-explanatory exploration of P.

In the case of TAIR modelling, abstracting consists in picking data that can be used to produce a digital image of a biological phenomenon (such as a metabolic pathway). TAIR curators abstract through reference to the GO framework, which provides the criteria for selecting, organising and displaying data on various aspects of Arabidopsis biology. The main concerns underlying the production of TAIR images are their explanatory power, internal consistency (determined by reliance on GO) and aesthetic value (simplicity and legibility), rather than the degree of accuracy to which their parameters capture the relevant features of the plants. In other words, TAIR modellers prioritise an accurate rendition of the *relations* among elements of the system over an accurate rendition of the empirical meaning of these elements as represented. In visualising a metabolic pathway (figure 5), the emphasis is on the type of relation linking elements such as amino acids and carbohydrates to enable metabolic processes; it is of secondary importance, within these models, whether the little triangles representing amino acids and the little squares representing carbohydrates tell us something about the actual composition and structure of these substances as found in real plants. In this sense, abstracting towards TAIR images is largely an intellectual activity: it is theory-guided, geared towards explanation and requiring no physical interaction with the phenomenal properties to be abstracted. I call this a case of 'intellectual abstracting'.

The goal of intellectual abstracting is to uncover ways in which a model can be *representative for* a given theory. The choice of the parameters used within the model is informed by a well-defined hypothesis about the theoretical outcome that the model

is supposed to illustrate, test, predict and/or elaborate. This is because we start from a theoretically informed 'prepared description' of the phenomena under scrutiny (Cartwright 1983: 133). A description of Arabidopsis microbiology constructed with the help of GO categories constitutes a good example for this. Cartwright notes that "the check on correctness at this stage is not how well the facts known outside the theory are represented in the theory, but only how successful the ultimate mathematical treatment will be" (ibid.). Substitute 'mathematical' with 'conceptual' and this statement becomes applicable to the construction of TAIR images, in which internal coherence and conceptual clarity have priority over the relations between models and the plants that they are meant to portray.[xv]

In the case of NASC, abstracting has very different connotations. It is aimed at the material replication of features of plants. These features are selected as desirable insofar as they allow to explore unknown aspects of Arabidopsis biology (e.g. the regulatory mechanisms responsible for the phenotypic differences between Lan and Col ecotypes). The stability of these features across different laboratory settings is a necessary condition for the resulting model to have representational value. Only through research on the same model can biologists compare their data and come to an agreement on how to explain various aspects of the phenomenon. Without agreeing on the traits that should characterise Lan and Col specimens as models of Arabidopsis plants, there can be no study of their regulatory systems, nor can results achieved through such study be applied to other species. Specimens that do not possess the traits selected by NASC researchers do not constitute trustworthy representatives of Arabidopsis wild types.

The epistemic priority of NASC modellers is to maintain control over the development of traits characterising different ecotypes, thus ensuring the replicability of specimens as well as their non-locality (that is, the stability of their features regardless of the time and location of their use). This is realised mostly by modifying the growth environment of the plants: that is, by ensuring that they are sown and germinated in the best possible conditions (e.g. with enough space, water and humidity) and by growing them in isolated containers under artificially regulated light. Direct interventions on the plants themselves are also involved, in case they manifest unexpected traits and when preparing seeds for storing and distribution. Here, abstracting does not imply establishing relations among given data by reference to specific theoretical frameworks (or, less generally, to actual explanations, as in the case of the explanatory powerful TAIR images). Abstracting involves selecting a limited set of material features of Arabidopsis wild types as potentially interesting for research purposes; devising ways in which these properties can be incorporated into a unique specimen; making sure that specimens with those characteristics can actually be grown; and constructing a toolkit of guidelines, materials and instruments allowing researchers worldwide to grow specimens in the same way.

This type of abstracting is performed by physical interaction between the researchers and the phenomenon to be modelled and is thus based largely on perceptually acquired knowledge about the phenomenon. Background theoretical knowledge is certainly involved in the researchers' choice of which traits to abstract and reproduce in the models: abstracting is theory-*informed*. Theoretical knowledge does not, however, determine the activities and results of modelling: abstracting is not theory-*guided*. The manipulation of models and the selection of traits to be modelled may

require no more than a general interest in exploring one or more aspects of the phenomena that they are taken to represent. NASC specimens are taken to be *representative of* a set of phenomena (amounting, depending on the research context, to 'all plants', 'all flowering plants', 'all weeds' or 'all other Arabidopsis ecotypes'). Epistemic access to phenomena is granted first and foremost by material manipulation, since the amount of intellectual manipulation necessary to handle these models is minimal. To emphasise the contrast with what I called intellectual abstracting, I refer to this type of activity as 'material abstracting'.

Some of the procedures that I list as part of the abstracting process involve the standardisation of materials and terminology. However, considering standardisation alone does not help to confront questions about the epistemological value of the objects thus produced. NASC specimens and TAIR images are no mere products for consumption by interested scientists: they are constructed and used as models, and it is this specific function that I am addressing here. Rather than in the representation of production processes, I am interested in the production of representations. Thinking about standardisation does not, by itself, invite a reflection on how objects become representative of other things; nor does it help to analyse the epistemological implications of using different types of models to study the same phenomenon. In contrast to this, thinking about the process of abstracting involves thinking about how models come to be *representative of* (or *for*) specific phenomena or issues, and with which consequences.

5. ***Skilful Abstracting***

The analytic distinction between intellectual and material abstracting only makes sense when abstracting is taken to denote different ways of handling models, rather than differences among the results of these processes (i.e. among types of models). But what exactly do we learn from it? In which ways is it a fruitful approach to the study of modelling practices?

One insight gained through this approach concerns the different kinds of epistemic skills required to produce, use and interpret models thus abstracted. The very adequacy of models as representations of natural phenomena is determined by the skill with which they are produced and handled as much as it depends on the features of models themselves. It is usually the epistemic community in which scientists work that determines, on the basis of convention, theoretical commitments and past experience, which skills should be exercised, and how, for a modelling activity to be judged as adequate. As evident from my reconstruction of the processes involved in producing NASC and TAIR models, abstracting itself requires the exercise of a number of skills. In producing NASC models, one needs to know how to handle plants, isolate them from each other, make sure they germinate and grow properly, regulate thermostats and ventilators and harvest seeds. TAIR modelling requires the ability to use a computer, type on a keyboard, write programmes in Java-script, search for relevant data through internet searches or contact with the researchers who obtained them, and so forth. These are what I call *performative* skills, involving the ability to interact with the environment (including laboratory equipment and various types of models) in ways relevant to the study of a specific phenomenon. Then we have *theoretical* skills, that is the ability to use various expressions of theoretical

knowledge (such as concepts and theories) towards acquiring understanding of a specific phenomenon.

The very distinction between material and intellectual abstraction can be fruitfully re-formulated as a difference in the type of skills required to abstract. Theoretical skills play a prominent role within intellectual abstracting. TAIR researchers value theoretical skills more highly than performative ones, and use them more often, as evident from the modelling steps described above. Their educational background is geared towards theoretical skills, as the overwhelming majority of TAIR researchers has been trained in developmental biology (that is, the branch of life sciences that embraces most elements coming from other disciplines – such as physiology, evolutionary and molecular biology – and thus has sophisticated theoretical tools at its disposal for studying complex biological processes). The performative skills necessary to running TAIR models – such as, as mentioned above, IT skills - are a prerogative of TAIR *programmers*. This is a sub-group within the TAIR research team that specialises in bioinformatics rather than actual biology and thus lacks some crucial theoretical skills required to make sense of the models that they develop.

To NASC modellers engaged in material abstracting, performative skills take centre stage – most obviously, in carrying out step 2 of the modeling procedure ('*cultivation of specimens'*). Accordingly, researchers working in the NASC laboratory and glasshouses need an educational background as technicians and/or experimentalists. NASC director Sean May values researchers with good performative skills and practice-oriented approach more highly than he does appreciate theory-directed scientists whose focus is on explaining, rather than acquiring, data (such as the ones working at TAIR). Performative skills are a crucial means to obtaining material abstraction, a reality that reflects on the background and skills of researchers hired to produce NASC models.

The gap in values, training and skills privileged within the two research teams is so great that it occasionally generates intellectual and social tensions between them. This reflects the striking difference between the types of skills accompanying different types of abstraction processes. At the same time, the distinction between material and intellectual abstracting does not in any way preclude recourse to performative skills to perform intellectual abstraction, or to theoretical skills to perform material abstraction.

6. Committing to Abstract

In addition to the emphasis on skills, viewing abstraction as a model-specific activity yields a second important insight. Distinguishing among different types of abstracting enables to pinpoint the *research commitments* that are inevitably tied to different modelling practices. I here take inspiration from the notion of commitment as 'a manner of disposing of oneself' that was originally proposed by Michael Polanyi (1967: 302-3). I view commitments as *biases* that become *entrenched* in the successful accomplishment of specific activities, encompassing items as diverse as the theoretical perspective held by the individual biologist engaging in research; the research goals and interests within his or her work; the research conducted by his or her research group; gestures and ways of moving; and the assumptions about the

representativeness of the research materials on which the researcher works, as well as the applicability of his or her results. Research commitments consist of knowledge that is assumed, rather than hypothesised, to be relevant to the study of the phenomenon in question. These biases are necessary conditions for the acquisition, conceptualisation, interpretation and communication of experiences and data: as such, they are crucial means for scientists to pursue their investigations.

Research commitments are strongly tied to epistemic skills. They are two aspects of the same process: acting skilfully implies committing to the activities and results that those skills bring forth, while making a commitment to a specific technique or concept requires learning skills adequate to follow up on such commitment. Yet, the two notions are clearly distinct. Epistemic skills denote the ability to perform an action in a manner deemed as adequate by the relevant epistemic community; commitments consist of a tendency towards (or preference for) pursuing goals that can be obtained through performing that action. As we saw in the examples of TAIR and NASC, considerable time and effort is invested in learning a skill. This means that the set of skills available to any one researcher is limited. Each researcher will learn and perfect skills helpful in pursuing his or her research interests. It also means that trained researchers, who already master several skills, have a strong tendency to pursue projects where those skills can be exercised, rather than embarking on projects where (a) they will have to invest time in mastering new skills and (b) they will not be able to use their old skills. In other words, possessing a skill often entails a commitment to exploiting that skill (thus taking advantage of an otherwise useless investment). Further, once a commitment is made, it implies learning and exercising skills relevant to fulfilling that commitment. This is what distinguishes commitments from a mere promise or a pledge: they imply, and result from, skilful action.

Note that this definition of commitment does not require that a scientist adopting a commitment should believe in its truth. Adherence to a commitment *might* be correlated to an individual's beliefs about reality and truth, but is primarily a *pragmatic necessity* emerging from the individual's need to act and think in specific ways and towards specific goals. As Lakatos pointed out by referring to the 'hard core' of research programmes, there is a set of beliefs, concepts, explanations and values that scientists must uncritically accept as background knowledge in order to pursue any investigation (Lakatos 1970: 133). In my view, the set of possible commitments envisioned by Lakatos should be enlarged to include not just theoretical beliefs and principles, but also specific procedures, ways of communicating and other kinds of tacit, gesturally or socially acquired knowledge. Scientists need to commit to specific procedures, protocols and standard ways to interact with others (each of which requires the acquisition of related skills) in order to follow courses of action that allow them to reach their scientific goals.

Let us now consider whether abstracting TAIR models involves specific commitments, and whether my analysis in terms of abstracting as an activity helps to analyse and compare them. As illustrated by steps (2) and (3) of their construction, TAIR models are heavily structured by the network of concepts referred to as 'gene ontology'. The abundance of data about Arabidopsis chromosomes and gene products, as well as the appeal that molecular biology exercises on research sponsors, forced TAIR researchers to adopt the term 'gene' as the central organisational term of GO (hence the name) as well as the starting point for all TAIR modelling efforts. This

practical necessity implies that the GO network structures its concepts according to their relation to genes. As a result, TAIR visual models are still far from being able to incorporate information about plant morphology, evolutionary history and higher-level physiology. The steps involved in the actual realisation of the abstracting of Arabidopsis biology into TAIR models generated a commitment to a gene-centric vision of Arabidopsis biology. This involved, in turn, a commitment to using Java programming to visualise gene-centric ordering of data as well as a commitment to spreading this view among Arabidopsis researchers of all trades. This bias is explicitly (and regretfully) acknowledged by the TAIR team and condemned by many ecologists and evolutionary biologists. Despite the criticisms, however, the commitment to gene-centrism is unavoidable as long as (a) genomic data overwhelmingly outrun ecological data and (b) the appropriate tools for the integration of genomic and ecological knowledge are not yet available.

NASC modelling does not commit to gene-centrism as heavily as TAIR does. NASC researchers translate their everyday experience with the morphology and physiology of actual plants into a commitment to highlighting their macroscopic features as well as their microscopic ones. This is evident in the use of descriptive terms like 'leaf' and 'stem' in the preliminary classification that constitutes step (2) of the modelling of NASC models. Likewise, the other steps in this modelling process do not require specific focus on the genetic make-up of Arabidopsis specimens. Thus, in the case of NASC, concerns about abstracting the material features of actual specimens determine a commitment to descriptive accuracy in categorizing ecotypes, where description involves both genomic structure and morphological illustrations.

Another commitment adopted by NASC researchers (but not TAIR) consist in the very goal of isolating and reproducing specific ecotypes as the 'right tools' for a given research goal.[xvi] A researcher working in the NASC glasshouse exercises skills such as sowing, harvesting, cleaning and ordering the seeds, as well as feeding the plants, caring for them and checking on their health and growth. These skills are intertwined with commitments to standards for what constitutes a healthy plant in any specific research context. For instance, experiments involving bacteria-induced genetic manipulations are best performed on the Ws ecotype. As mentioned above, this is not because Ws specimens are 'naturally' more susceptible to mutation, as Arabidopsis researchers used to believe in the 1990s: the Ws ecotype is now known to be just as variable as other Arabidopsis ecotypes. Still, for at least a decade, NASC technicians committed to using Ws specimens for induced mutations. This commitment influenced the way in which NASC technicians produced Ws seeds: that is, with an eye to maintaining the genotypic and morphological characteristics of the plant favouring high mutation rates. The commitment influenced NASC researchers to the point that the model resulting from their efforts actually conformed to the researchers' expectations. NASC procedures transformed the original Ws ecotype into plants that are specifically engineered to be transformed through genetic manipulation.

This last example brings me to an important feature of the relation between the process of abstracting and the formation of commitments. As in the case of Ws specimens, pre-existing commitments influence the way in which researchers abstract phenomena to construct models of those phenomena. Modellers choose among available ways to abstract depending on the commitments that pre-exist the modelling procedure and the skills that they are able to exercise. However, there are cases where

the very process of abstracting generates new commitments (or modifies old ones). This can happen through the exercise of new skills. A researcher may decide to learn a skill, such as programming, in order to abstract features of a phenomenon to be used in a model. As a result of this decision, the researcher commits to using a specific type of computer and software, and thus also graphics and diagrams, to model the phenomenon. This implies that the selection of features to be abstracted from the phenomenon depends on the representational capacity of the software that is used. In the case of TAIR, a researcher wishing to model Arabidopsis metabolism will not be able to select cell transmission patterns as a feature to be abstracted from the phenomenon, because TAIR software does not allow to insert that variable into the model. The modeller will have to select different features of the phenomenon, such as gene expression patterns. The whole process of abstracting is affected by the commitments imposed by the chosen method of representation and the skills involved in it.

7. Conclusion: Abstracting and Model-Based Understanding

The notion of abstraction that I here proposed emphasises the actions, choices, displacements, conceptual and physical transformations involved in the creation and use of biological models. This approach deserves philosophical attention. Thinking of abstraction as a process highlights the role of human agency, in the form of skills and commitments, in the production of scientific models.

In the spirit of pursuing a pluralistic approach to the notion of abstraction, I used a detailed reconstruction of the practices required to produce two types of models of Arabidopsis in order to outline a distinction between two types of abstraction processes: the 'material abstracting' required in the production of NASC specimens and the 'intellectual abstracting' characterising the elaboration of TAIR images. The difference among these types of abstracting is determined by the epistemic goals, material circumstances (tools and experimental setting) and relevant skills of the scientists performing them. Reflecting on those differences proves helpful in pinning down the research commitments that biologists need to accept. Further, acknowledging the skilful practices and distinctive commitments involved in abstraction has implications not only for the developers of models but also for their users. Scientists employing NASC specimens and TAIR images to further their understanding of plant biology require skills, materials and commitments that allow them to handle those models successfully.

Acknowledgments

Discussions with Rachel Ankeny, Hasok Chang, James Griesemer, Henk de Regt and Hans Radder were crucial to the development of my analysis. Thomas Reydon, Rasmus Winther, an anonymous reviewer and the editor closely read and commented upon the last draft, which has considerably improved as a result. I also thank the Arabidopsis researchers who shared their time, facilities and thoughts with me: Sue Rhee and her team at the TAIR and Sean May and his team at the NASC. This

research was supported by the Netherlands Organisation for Scientific Research (NWO), The Leverhulme Trust and the ESRC.

References

Bard, J.B.L. and Rhee, S. (2004) Ontologies in Biology: Design, Applications and Future Challenges, *Nature Reviews: Genetics*, 5, 213-222.

Bowker, G.C. and Star, S.L.: 1999, *Sorting Things Out. Classification and its Consequences*, The MIT Press, Cambridge, MA.

Cartwright, N.: 1999, *The Dappled World*, Cambridge University Press, Cambridge, MA.

Cartwright, N.; 1983, *How the Laws of Physics Lie*, Cambridge University Press, Cambridge, MA.

Cat. J.: 2001, On Understanding: Maxwell on the Methods of Illustration and Scientific Metaphor, *Studies in the History and Philosophy of Modern Physics*, 32, 3, 395-441.

de Chadarevian, S. and Hopwood, N.: 2004, *Models. The Third Dimension of Science*, Stanford University Press, Stanford, California.

Clarke, A.E., and Fujimura, J.H.: 1992, *The Right Tools for the Job. At Work in Twentieth-Century Life Sciences*, Princeton University Press, Princeton, New Jersey.

Griesemer, J.R.: 2004, *Three-Dimensional Models in Philosophical Perspective*, in de Chadarevian and Hopwood (eds.), *Models. The Third Dimension of Science*, Stanford University Press, Stanford, California, pp.433-442.

Lakatos, I.: 1970, Methodology of Scientific Research Programmes, in I. Lakatos, I. and A. Musgrave (eds.), *Criticism and the Growth of Knowledge*, Cambridge University Press, Cambridge, UK, pp. 91-196.

Laubichler, M. and Müller, G.B.: 2006, *Modeling Biology. Structures, Behaviours, Evolution*. MIT Press, Cambridge, MA.

Leonelli, S.: 2007, Arabidopsis, the Botanical Drosophila: from Mouse Cress to Model Organism, *Endeavour*, 31, 1: 34-38.

Meinke, D.W. et al.: 1998, Arabidopsis thaliana: A Model Plant for Genome Analysis, *Science*, 282, 672-682.

Meinke, D. and Scholl, R.: 2003, The Preservation of Plant Genetic Resources. Experiences with Arabidopsis, *Plant Physiology*, 133, 1046-1050.

Meyerowitz, E.M.: 2001, Prehistory and History of Arabidopsis Research, *Plant Physiology*, 125, 15-19.

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

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

Polanyi, M.: 1962, *Personal Knowledge*, Routledge, London.

Radder, H.: 2006, *The World Observed/The World Conceived*, Pittsburgh University Press, Pittsburgh.

Somerville, C. and Koornneef, M.: 2002, A fortunate choice: the history of Arabidopsis as a model plant, *Nature Reviews: Genetics*, 3, 883-889.

Internet resources

The Multinational Coordinated Arabidopsis thaliana Functional Genomics Project. Annual Report (2004) http://arabidopsis.info/info/masc_june_04.pdf

NASC
http://www.arabidopsis.info

Open Biological Ontologies
http://www.obo.souceforge.net

TAIR
www.arabidopsis.org

Legends

Figure One - *Directed acyclic graph (DAG) used by Object-Oriented Java Software. Each 'child' (i.e., object derived from/connected to other objects, here represented as a yellow/blue dot) may have multiple 'parents' (objects from which other objects derive).*

Figure Two - *GO: a hierarchy of terms each with a precise definition and relationship to other terms.*

Figure Three – *Gene expression data are collected and ordered according to a given GO schema (here encompassing data from other model organisms as well).*

Figure Four – *'Fitting' a schema. Programmers design a programme based on the relations established by GO and the relevant data selected through PubSearch. While the details of the chosen data and relations do not matter for my purposes, note how the DAG structure is now operationally fitted to the GO schema.*

Figure Five – *A model of Arabidopsis metabolic pathways as displayed in the TAIR website. It provides information about the components relevant to metabolism at the cellular level and about the expression level of genes controlling each component. This is a dynamic model: one can click on any component (the triangles, representing amino acids, or the thin white line connecting several component and representing the pathway itself) to access a variety of sub-models of the components themselves, together with the data used to construct the models and information about the sources from which data has been acquired.*

Figure Six – *Exemplar of PATO.*

Figure Seven – *NASC glass-house for growing specimens (the visible plants are of the Lansberg variety, the third most popular ecotype for laboratory use together with Col and Ws).*

Figure Eight – *NASC technician sowing a specific ecotype*

Figure Nine – *Harvesting of seed (before sterilisation)*

---

[i] See the collections of essays edited by Morgan and Morrison (1999), de Chadarevian and Hopwood (2004) and Laubichler and Muller (2006) for examples of how different types of models are combined in scientific research. Note that it is not within the scope of this paper to provide an innovative definition of what a 'model' is: the broadly defined notion of models as mediators, put forward by Morgan and Morrison (1999), suffices for my purposes.

[ii] Griesemer (2004: 436).

[iii] This estimate is based on data collected by The Arabidopsis Information Resource in 2006.

[iv] The history of research on Arabidopsis and its progressive institutionalisation (including the establishment of the Multinational Arabidopsis Steering Committee) is documented in Somerville and Koornneef (2002), Meyerowitz (2001), Meinke et al. (1998) and Leonelli (2007).

[v] I visited The Arabidopsis Information Centre in August 2004 and the Nottingham Arabidopsis Stock Centre in May 2005. At both centres, Directors Sue Rhee and Sean May provided me with access to their archives and resources. I also had the possibility to interview them and their staff at length, thus gathering information about how they construct and use the models.

[vi] Another two sites are the Munich Information Centre for Protein Sequences [MIPS] and the 'Arabidopsis.info' based at the Nottingham Arabidopsis Stock Centre [NASC].

[vii] The initial focus on genomics was determined by the abundance of data gathered through the Arabidopsis Genome Initiative (the multinational project that successfully sequenced the plant's genome between 1996 and 2000). TAIR personnel insists that the choice to organise the database on the basis of genomic data is pragmatic rather than conceptual, and that the ultimate TAIR aim is to obtain balance and integration of information pertaining to the genomic, the evolutionary and the ecological levels (pers. comm.). Nevertheless, as I show, the commitment to emphasise gene-level data prevents TAIR from giving equal space to data concerning higher levels of organisation in Arabidopsis.

[viii] A model of metabolic cycle is displayed in figure 5. See TAIR website for further examples.

[ix] More information on this project can be found online: http://obo.sourceforge.net/

[x] The definitions used for the concepts employed in GO are agreed upon during GO Content Meetings, in which developers discuss their choices with experts in relevant biological domains.

[xi] Note that 'schema' in TAIR terminology does not denote the organisation of data into various categories (which occurs in steps 1 to 3), but rather the way in which programmers visualise these categories through available digital technologies.

[xii] My list is not supposed to be exhaustive, but rather to give an idea of the confusion underlying the use of the term 'abstract' in discussions of modelling practices.

[xiii] See Cat (2001) for an exploration of this notion of abstraction in relation to Maxwell's work.

[xiv] A good exemplification of this view can be found in Cartwright (1999): a description is abstract insofar as it can be 'fitted out' to a number of other descriptions. Similarly, the characterisation of a

model as abstract or concrete depends on the context in which the model is used. This approach is captured by Radder's definition of abstraction as 'summarising', which he identifies (and goes on to criticise) as one of three main senses in which abstraction works (2006: 110).

[xv] The use of intellectually abstracted models is increasingly widespread among biologists. Take the pervasive use of simulations and algorithms to visualise empirical data, not to mention the push towards formalisation and away from the laboratory brought about by the increasing use of bioinformatics to store, organise and integrate data. These models are especially useful for elaborating explanations or confirming predictions stemming from given hypotheses (what Cartwright calls *interpretative* models in her 1999: 181). They are also fundamental to the integration of biological knowledge concerning specific phenomena (as, for instance, bringing together insights from physiology, genomics and cell biology to understand root development in plants). However, precisely because of their strict reliance on theoretical assumptions, models constructed through intellectual abstracting are very helpful in cases where the goal of their manipulation is to improve the empirical content of a theory. They give little indication as to which features of the phenomenon under scrutiny should be considered relevant to the development of explanatory knowledge about that phenomenon. Further, such models do not help with testing the empirical (descriptive) accuracy of the relation it stipulates between theoretical terms and aspects of the phenomenon.

[xvi] See Clarke and Fujimura (1992).