Paper forthcoming in *BioSocieties* (submitted January 2012, accepted May 2013)

Global data for local science: Assessing the scale of data infrastructures in biological and biomedical research

Sabina Leonelli, University of Exeter, s.leonelli@exeter.ac.uk

Abstract: The use of online databases to collect and disseminate data is typically portrayed as crucial to the management of 'big science'. At the same time, databases are not deemed successful unless they facilitate the re-use of data towards new scientific discoveries, which often involves engaging with several highly diverse and inherently unstable research communities. This paper examines the tensions encountered by database developers in their efforts to foster both the global circulation and the local adoption of data. I focus on two prominent attempts to build data infrastructures in the fields of plant science and cancer research over the last decade: The Arabidopsis Information Resource and the Cancer Biomedical Informatics Grid. I show how curators' experience of the diverse and dynamic nature of biological research led them to envision databases as catering primarily for local, rather than global, science; and to structure them as platforms where methodological and epistemic diversity can be expressed and explored, rather than denied or overcome. I conclude that one way to define the scale of data infrastructure is to consider the range and scope of the biological and biomedical questions which it helps to address; and that within this perspective, databases have a larger scale than the science that they serve, which tends to remain fragmented into a wide variety of specialised projects.

Keywords: databases, big biology, data infrastructure, large scale, epistemic diversity, data re-use.

Introduction

Much has been written on the impact that data infrastructures such as digital databases are having on scientific research, and particularly the biological and biomedical sciences. Most of this scholarship has emphasised the importance of *quantities* (of data, researchers, investments and research sites, among other factors) as a key motivation underlying the widespread adoption of these infrastructures for scientific communication.[1] Massive research efforts and resources are being devoted to the dissemination of data online by public and private funders across the globe.[2] Many scientists and science funders view databases as crucial tools to handle the vast amount of molecular data produced by technologies such as automated sequencing and microarray experiments (often referred to as 'big data'), and getting them to travel across the world quickly and easily. It is hoped that free and widespread access to large datasets will enhance the use of such data as evidence for new claims, thus generating new paths towards discovery (e.g. Hey et al 2009, Royal Society 2012). The insistence that online databases constitute a solution to the problems posed by large quantities, however, clashes with the *quality* considerations attached to actually using (evaluating and interpreting) data to produce new knowledge. Biology and medicine are notoriously fragmented into countless epistemic cultures, each characterised by different interests, values, forms of reasoning, methods, material objects and standards for what counts

[1] See for instance Stein (2008) and a recent issue of *Science* (2011). STS literature on this topic is also extensive, as exemplified by Hine (2006), Bowker et al. (2010), Edwards (2010) and Leonelli (2012).
[2] The internet, used in conjunction with distributed computing tools such as grids and cloud computing, enables the dissemination and retrieval of information on a geographical and temporal scale surpassing anything seen before. STS scholars who investigated the role played by online databases in supporting large research network include Star and Ruhleder (1996), Bowker (2000), Ribes and Finholt (2009), Baker and Millerand (2010), Leonelli (2010), Parker, Vermeulen and Penders (2010) and Edwards (2010).

as evidence.[3] Further, epistemic cultures are not stable objects: the combinations of individuals, expertises, interests and methods that characterise them are subject to constant change to match the ever-shifting nature of biological knowledge and of living systems themselves.[4] This extensive and dynamic pluralism makes it hard to develop databases that can bridge such diverse expertises, and thus fulfil the specific needs of each community.[5] In confronting both the high quantities and the diverse qualities of research, databases are subject to two seemingly opposite requirements: fostering the global circulation of data and facilitating their local adoption.

What is 'big' about online databases and the sciences that they are meant to support? This paper considers how two groups of database curators have attempted to overcome this challenge, and uses this empirical material to reflect on what scale involves in contemporary biological data infrastructures. I here articulate two complementary answers to this question. *First*, I take a critical stance against the very idea that online databases are catering for 'big biology'. I show that online databases are primarily responsible for providing useful support to the needs and questions emerging from the unique combinations of expertise and interests brought together within any one biological project. In other words, online databases need to cater primarily for 'little science', and if they fail to fulfil this requirement, they eventually cease to be used and funded. As I discuss below, I do not think that Derek De Solla Price's seminal characterization of 'little science' (De Solla Price 1963) fits the type of projects that I am discussing here, which target one scientific question through a specific research approach for a limited period of time (in this sense they are 'little'), but can involve a large number of

[3] For the notion of epistemic culture, see Knorr Cetina (1999); cultural fragmentation and its instantiations within data infrastructures have been documented with regard to environmental, ecological and geological datasets by Bowker (e.g. 2001) and Ribes and Finholt (2009).
[4] Other contributions to this special issue, particularly Emma Frow and Andrew Bartlett, provide excellent examples of the unstable nature of epistemic communities in contemporary biology.
[5] This is especially difficult since scientific needs, questions and technologies keep changing, which forces curators to think hard about temporality and future expectations (the 'long now', as discussed by Ribes and Finholt [2009]).

scientists across different institutions and nations. I thus agree with Niki Vermeulen's argument that biology, in its currently globalised and networked incarnation, embodies a new way of doing science (Vermeulen 2009; see also Leonelli forthcoming); and yet, I disagree with her that the label 'big science' can be usefully applied to most research projects carried out in contemporary biology. *Second*, I show that in order to foster both the global dissemination and the local adoption of data, database curators strive to develop platforms through which methodological and epistemic diversity can be expressed and explored, rather than hidden and/or ignored. I conclude that one way to understand what counts as the scale of data infrastructures is to consider *the range and scope of biological questions that data stored therein can be used to address*; and that if scale is understood in this way, well-functioning databases are necessarily bigger than the science that they serve, which tends to remain fragmented into a wide variety of specialised projects.

A Tale of Two Databases

My arguments are grounded on a historical study of two efforts made to build 'all-encompassing' databases in the biological and biomedical domains over the last decade. Each of these infrastructures was intended to serve a large and cutting-edge research area; and while one is directed to a quintessentially biological field (plant science), the other supports what is arguably the most active area of biomedicine (cancer research). The first case is The Arabidopsis Information Resource (TAIR), which was started in 1999 with funding from the National Science Foundation (NSF) with the objective to gather data relevant to understanding model plant *Arabidopsis thaliana*. In parallel to the success of *Arabidopsis* as a key model organism for plant science as a whole, TAIR became a prominent resource within this area in the late 2000s. Since 2010, its structure and content have been undergoing

significant reworking to guarantee its usefulness to users and long-term sustainability. My analysis is based on the consultation of publications and archives released by TAIR itself and available on its website; and publications and reports on plant bioinformatics issued by the NSF and the Arabidopsis plant community over the past ten years.[6] The second case study is the Cancer Biomedical Informatics Grid (caBIG), created in 2003 to function as a portal linking together datasets gathered by the research institutions and patient care centers under the purview of National Cancer Institute (NCI). The initial goals of caBIG curators included assembling and integrating all existing datasets on all types of cancer research across a wide range of institutions; this ambitious aim was recently scaled down due to widespread critiques of its limited achievements. My analysis is based on a close study of the caBIG website; of archival sources, including publications and minutes of meetings of caBIG curators available online; and on publications and reports on cancer bioinformatics and caBIG's role in it issued by the NCI and several cancer researchers over the past five years. Since their inception, both databases have aimed to serve as wide a research community as possible, while at the same time helping researchers to seamlessly fit data retrieved online into their existing projects. The achievement of these aims was complicated by the huge variety of data to be disseminated; the diversity of loci of data production (and thus the format and methods of data generation); and the ongoing tension between biological standards and protocols, and computational methods used to format, annotate and visualise data so that they are machine-readable. These cases confirm existing findings that high levels of standardisation, accessibility and visibility are key requirements for databases aimed at data re-use on a large scale; and yet, whether these databases are ultimately successful

[6] I have also carried out extensive ethnographic work with TAIR curators and the Arabidopsis community more generally, which has supported previous publications (e.g. Leonelli 2007, Leonelli and Ankeny 2012) and is still ongoing. This intensive engagement has certainly informed my choice of materials for the present analysis, but all empirical information on TAIR used for this paper is available from published literature.

depends on how well they encompass – rather than exclude or deny – the epistemic pluralism that characterises research in the life sciences.

The Arabidopsis Information Resource (TAIR)

TAIR was created in 1999 as a replacement for AtDB (the *Arabidopsis thaliana* Database), a small database containing selected genetic data on the model plant. The NSF funded TAIR to ensure that data coming out of the international sequencing project devoted to *Arabidopsis* would be adequately stored and made freely accessible to the plant science community. The Carnegie Institution for Science's Plant Biology Department, home to prominent plant scientists including *Arabidopsis* veterans Chris and Shauna Somerville, won a national bid to create and host the database; and a former student of Chris Somerville, Seung Yon Rhee, was given the task of directing TAIR, which she undertook with a strong vision for what the database should become in the future. As an experienced experimenter, Rhee thought that TAIR should not be just a repository for sequence data; rather, it should become a repository for all data extracted from Arabidopsis research, a platform facilitating communication and exchange among researchers working on different aspects of plant biology. Further, the database should contain a set of tools for data retrieval and data analysis, which would facilitate the integration of all those data. Finally, the database should enable users to integrate *Arabidopsis* data with data extracted from other plant species, thus paving the way for developments in plant science as a whole (Rhee et al. 2006, 352)

Thanks also to the success of *Arabidopsis* as a model organism (Leonelli 2007, Koorneef and Meinke 2010), TAIR indeed became a reference point for plant researchers across the globe, assembling an impressive array of datasets concerning disparate aspects of Arabidopsis biology, ranging from morphology to metabolic pathways. The database includes various

search and visualisation tools elaborated by the TAIR team to help plant scientists in retrieving and interpreting *Arabidopsis* data. Examples are MapViewer, which allows access to various types of mappings of Arabidopsis chromosomes; and AraCyc, which visualises data about biochemical pathways characterising Arabidopsis cellular processes. TAIR provides abundant information about how these tools have been constructed, how they should be used and which types of data are included; and allows users to order Arabidopsis seeds directly from Arabidopsis Stock Centres.[7]

The importance of TAIR to plant science became a hot topic for scientific debate in 2008, when the National Science Foundation decided to cut funding to the resource. The motivations for this decision were several, and included financial constraints as well as the wish for TAIR to accommodate the changing needs of the plant science community, and particularly the increasing importance of tools for the analysis and comparison of data across different plant species. At the same time, the decision to cut TAIR funds, thus effectively making it impossible for its curators to extensively review and revise its contents, was controversial within the plant science community. Scientists' protests poured in from all corners of the globe, *Nature* published an editorial on the key role of TAIR in plant science, and several working groups were set up to find ways to support TAIR in the long term, thanks also to the effort of the Multinational Arabidopsis Steering Committee [MASC] and the Genomics Arabidopsis Research Network, or GARNet (Ledford 2010; Bastow et al. 2010). These events illustrate how TAIR, in a similar way to other model organism databases such as WormBase and FlyBase, has played an exemplary role in demonstrating the value of data curation to experimental biology (Leonelli and Ankeny 2012). This involves substantial curatorial labour, for reasons that I shall now describe.

[7] For scientific details on TAIR and its components, see e.g. Huala et al. (2001), Garcia-Hernandez et al. (2002), Rhee et al. (2003), Mueller et al. (2003).

First, consider the variety of datasets that TAIR attempts to host under the same digital umbrella. Quantity is not the main problem here. There are of course worries about securing the hardware and memory necessary to store the masses of data collected by TAIR on a weekly basis, but the most urgent problems confronted by curators concern the diversity in quality and provenance of those data. Here we can see one way in which the extensive fragmentation of biological research into different epistemic communities affects the set-up of databases: each groups tends to use different methods, instruments, formats and protocols to produce data, which makes it very hard to integrate those results with data produced by other groups, even when they document the same biological aspect. For instance, the most straightforward type of data curated by TAIR is sequence data documenting *Arabidopsis* genome structure, and yet even in that case complications abound. Data obtained through the first sequencing project are being constantly updated and checked, in order to correct mistakes or inaccuracies arising from the attempt to merge datasets produced by different groups in different locations. These curatorial efforts include adding novel genes, updating exon/intron structures of existing genes, deleting mispredicted genes, merging and splitting genes, changing gene types, and adding splice-variants (Swarbreck et al 2008). These efforts only increase in the case of functional, metabolic and morphological data about plant mutants, whose production is even more dependent on the preferences and local conditions of data producers.

Another crucial difficulty is posed by the unpredictability of the uses to which data hosted by TAIR could be put in the future. TAIR curators devoted years of efforts to making TAIR as accessible and interesting as possible not only to Arabidopsis specialists, but also to other plant scientists and even biologists working on other kingdoms. This is an extremely ambitious goal, which curators have pursued from the early days of TAIR development and which involved frantic consultation of literature dealing with information management, in the

hope of finding suggestions about how to integrate and visualise the most diverse information in the simplest possible way, without losing sight of the diversity of cultures through which data are produced and re-used. One early strategy adopted by curators was to create several different search engines within TAIR, each of which would provide a different perspective on *Arabidopsis* biology. They devised a search engine visualising the location of genes on *Arabidopsis* chromosomes; another displaying data about gene expression; another focused on data about biochemical pathways; and so on. The possibility to gather data about the same phenomena from different perspectives, they reasoned, would maximise the information available to users while minimising losses in the accuracy or the richness of data. Most importantly, users would be allowed to formulate their queries in a variety of different ways, reflecting their own epistemic commitments: they would be able to choose among different parameters and ways to display the results of their searches (for instance, when searching a specific gene locus on a chromosome, TAIR users can view their results in the form of a genetic, physical or sequence map).

Not all of these tools have been found to be equally valuable and accessible by plant researchers, and TAIR curators have reduced their ambitions over time, focusing increasingly on updating sequence and functional data on Arabidopsis rather than including new data types and tools for comparison across plant species (which might be viewed as one reason for their loss of funding). Still, what I want to stress here is that the construction of TAIR involved not only collecting diverse types of *Arabidopsis* data from multiple sources, but also elaborating strategies through which data could be organised, retrieved and visualised by users from various parts of plant science. In Rhee's words,

> "Ultimately, our goal is to provide the common vocabulary, visualisation tools, and information retrieval mechanisms that permit integration of all knowledge about Arabidopsis into a seamless whole that can be queried from any perspective. Of equal

> importance for plant biologists, the ideal TAIR will permit a user to use information about one organism to develop hypotheses about less well-studied organisms." (Rhee website, accessed January 2005)

During its 12-years-long existence, TAIR could be said to have become a virtual laboratory, experimenting, with mixed results, with different ways of visualising both data and the biological phenomena which data are used to interrogate. The main challenge has been to imagine what users want, since TAIR aims to reach several types of scientific audiences, ranging from developmental to molecular and theoretical biologists (not to mention the differences in epistemic cultures to be found within and across those categories). To confront this problem, TAIR curators have drawn insight from their own experience as bench researchers specialised in different areas of *Arabidopsis* biology. For instance, developmental biologist Eva Huala, who became TAIR director in 2005, proposed that the user should be able to "fly into" the chromosome, i.e. to view and explore a three-dimensional representation of *Arabidopsis* chromosomes that is produced and constantly modified on the basis of incoming experimental data. This meant that TAIR should provide complex, three-dimensional visualisation tools that would allow users to click on the image of a specific chromosome and see a representation of the inside of the chromosome. On the one hand, this representation would have to be realistic enough as to convey ideas about the actual structure and physiology of chromosomes; on the other hand, it would have to contain specific references to the data from which the model was generated, so as to allow users to trace the sources and original context of the data. Further, from its inception TAIR has collaborated with other model organism databases, with the ultimate goal to provide an integrated platform with repositories of data on other organisms. In this sense, TAIR curators have indeed invested effort in facilitating comparative research. External collaborations included consultations with experts in each relevant field; participation in the development of the Gene

Ontology as a controlled vocabulary to disseminate data about gene products across species; and meetings with curators working on other model organisms to compare strategies and develop a joint resource (the Generic Model Organism Database).

Initially curators put a lot of emphasis and hope in the idea that users would recognise the benefits of a resource such as TAIR and would, as a consequence, support it both by donating data to it and by helping curators to get it right, for instance by sending feedback and engaging with the technicalities of how data were collected, stored, visualised by the database. The idea that users would enter a strongly collaborative relationship with database curators, however, proved to be misguided. Providing input is difficult and time-consuming, as it requires familiarity with the software and classification systems used by TAIR. Users have no real incentive to do this, especially since no formal credit system is yet in place for data donation within biology. This results in users expecting curators to take full responsibility for how data are presented, so that users can access the data they need and get on with their research. Rather than asking biologists for direct feedback on the vision of the database, curators thus started asking users for queries that would likely be submitted to a database such as TAIR. The crucial issue for the TAIR team became: can we answer this query with the current tools? By asking this question, TAIR curators effectively brought together their concerns about information management and user-friendliness, thus elaborating designs for easily accessible, and yet rich databases.

Curators also made assumptions about how the plant community should organise itself so that a database like TAIR reaches its full potential. They strongly relied on the existence of a collaborative, open access ethos within the community, which Rhee herself aptly characterised through the motto ‘share and survive’ (Rhee 2004). This constitutes a surprising exception in the competitive context of biological research, and of molecular biology in particular. Since the early 1980s, when *Arabidopsis* was re-discovered as a model organism

and research efforts on it acquired momentum, prominent *Arabidopsis* researchers agreed that the data acquired through research on this model organism should be kept freely available (Leonelli 2007, Koorneef and Meinke 2010). Many biologists and research institutions involved in Arabidopsis research continue to foster this ethos. This situation has had a strong impact on how TAIR was developed and on the expectations that TAIR curators placed on their user community.

The Cancer Biomedical Informatics Grid (caBIG)

caBIG is an ambitious bioinformatics initiative within an area that is vastly better funded and more visible than plant science. The curatorial vision behind caBIG parallels both the commitments to open access and data re-use fostered by TAIR curators and their ongoing struggle with serving widely diverse user communities, although caBIG curators have arguably been much less successful than TAIR curators in implementing effective feedback mechanisms for their curatorial choices and thus gaining the support of prospective users. As I shall argue, this is partly due to the difficulties encountered by caBIG curators with understanding and capturing the pluralism characterizing cancer research, which has meant that this resource, though widely and easily accessible online, has not yet been widely adopted as a tool to foster medical research.

caBIG was started in 2003 as a corollary of the 'big science' programme of the National Cancer Institute (NCI) in the United States. Its initial functions were to facilitate data sharing among stakeholders in NCI research, including clinicians, research scientists, pharmaceutical companies, primary healthcare providers, and patients; and showcase the results obtained by NCI in its fight against cancer. Given the scope of the enterprise, as well as the diversity of

ethos, training and results characterizing these groups, the caBIG remit was immediately seen by many bioinformaticians and cancer researchers as hugely ambitious and possibly unrealistic (Ochs et al 2010). Despite this early skepticism and a rather critical review of the caBIG pilot phase (National Cancer Institute 2008), its reach and ambitions expanded towards becoming a portal to all cancer data, no matter where those data were first obtained and which national agencies see themselves as responsible for their dissemination.[8] Andrew von Eschenbach, who was directing the National Cancer Institute (NCI) at the time when CaBIG was started, and Kenneth Buetow, the Associate Director for Bioinformatics and Information Technology at the National Cancer Institute and the founder of CaBIG, summarized their vision as follows:

> "Medical research teams have operated, in effect, as cottage industries, each collecting and interpreting data using a unique language of their own making and in virtual isolation from other teams. Biomedical informatics has the potential to be the powerful critical means to achieve the necessary degree of integration as it provides the mechanisms and tools to support standardized sharing, management and analysis of diverse data across the bench-to-bedside continuum and back." (Eschenbach and Buetow 2006, p. 22)

caBIG was funded to spearhead efforts towards data re-use across diverse research and clinical contexts. In a similar way to TAIR, caBIG embodies the promise of data infrastructure to facilitate data management on a very large scale. The challenge faced by caBIG curators is however even more substantial. caBIG operates in a densely populated research environment, where hundreds of projects around the world are already utilizing digital databases to store and disseminate their results. It therefore made little sense to re-

[8] The attempt to embrace data production contexts beyond the (American) ones within the NCI pursuit brings considerable complications, as exemplified by recent comparisons between European and US clinical trials (Kohli-Laven et al 2011).

invent the wheel and construct a central resource from scratch, as TAIR curators were tasked to do.[9] Rather, caBIG started with bringing together existing repositories of cancer data, by providing a superstructure through which users can access the data they are interested in no matter where it is actually located and who is administering the relevant database. caBIG thus proposed to facilitate data re-use by providing a common point of access to data, while at the same time exploiting existing databases produced and maintained by other organizations. Robert Beck, a participant in the caBIG Strategic Planning Workspace, presents the following user scenario to illustrate the usefulness of such an endeavor:

> "A researcher involved in a phase II clinical trial of a new molecularly targeted therapeutic for brain tumors observes that cancers derived from one specific tissue progenitor appear to be strongly affected. The trial has been generating proteomic and microarray data. The researcher would [now] like to identify potential biochemical and signaling pathways that might be different between this cell type and other potential progenitors in cancer, deduce whether anything similar has been observed in other clinical trials involving agents known to affect these specific pathways, and identify any studies in model organisms involving tissues with similar pathway activity. [..] With caBIG compliant components now under development, the researcher would be able to perform the analysis routinely, with data flowing through systems and analysis being automatic. This analysis will yield biomarkers and potential drug targets gathered from multiple workspaces and make it possible to develop treatment modalities faster, less expensively, and more effective for patients." (Beck 2005, p. 10)

Realizing this vision of seamless data dissemination and retrieval involves building standards aimed at bridging the gaps between existing databases (what Beck calls "caBIG compliant

[9] Though as I noted above, TAIR curators were building on previous, smaller-scale efforts to collect Arabidopsis sequence data.

components"), thus facilitating the retrieval and visualization of very different types of data, ranging from genomic sequences to symptoms of individual patients, coming from hundreds of different sources. The way in which caBIG has attempted to achieve this is by pushing the databases collected under its purview to adopt common formats and follow basic structural rules enabling basic *interoperability* across different databases. The notion of interoperability constitutes a specific way to envisage data dissemination, integration and re-use. As in the case of TAIR, CaBIG curators wish to leave as much room for selecting and interpreting data as possible to their users. Interoperability, defined as "the ability of two or more systems or components to exchange information and to use the information that has been exchanged" (Covitz 2004, p. 8), seems well-suited to making data travel across diverse environments without necessarily affecting the local norms of the contexts in which data are produced. Interoperability requires only a minimal amount of consistency and coherence among existing databases: just enough to enable users to move from a data domain, type and source to another, while leaving them free to read and use the data that they find as they see fit, to answer any scientific query they may have. To make this possible, the structure of caBIG is highly modular. Since its inception, caBIG has been organized into separate pilot groups working on different areas of data management, ranging from curating data acquired through medical imaging technologies to managing tissue biobanks and data coming from clinical trials. These pilot groups were recently formalized into 'domain workspaces', each of which oversees the storage and management of different data types, and aims to create tools to enable user communities to overcome physical, legal and scientific barriers to data sharing (table 1).

[TABLE 1]

Table 1 constitutes only a glimpse of the structural and organizational complexity and scope of caBIG. What becomes clear from even such a limited perspective is that this resource is not supposed to operate on a shared, unified understanding of what it could be used for. The idea underlying caBIG is that its use will have distinct characteristics depending on specific biological and biomedical queries and on the contexts from which the queries come. By capitalizing on interoperable modules, caBIG is trying to walk the line between a centralizing, top-down structure which produces universal standards and provides a focal point of reference for all users; and a decentralized, bottom-up resource informed by curators and researchers around the world. This attention to diversity and decentralisation is not only tied to the curators' awareness of the variety of epistemic cultures and research traditions involved in cancer research broadly construed; it is also tied to their awareness of the regimes of competition and intellectual property involved in high profile biomedicine. The challenge of constructing a data-sharing community around cancer research was highlighted by John Niederhuber, the director of the National Institute of Cancer, as the most significant contribution that caBIG can hope to make:

> "caBIG® is an example of a new approach to organizing medical research in the future that is really both an experiment and yet a transformation at the same time. No single individual or organization can manage the amount of data that we deal with now in biomedical research. Ideally, this information must be available online, in real time, so doctors, patients, and organizations like ours can use it quickly. This is really creating a new community of research. That's what caBIG® is. It's not just a technology; it's a cultural change." (Leaflet caBIG, January 2011)

Remarkably, the cultural change is viewed as emerging from the inclusion of as many stakeholders in cancer research as possible, including patients and national funding agencies across the globe:

> "By having this infrastructure in place, we have the capacity to both support next generation clinical research and also, more importantly, inform next generation practice outcomes and bring molecular medicine into the clinical practice environment. However, if we're going to do this on a broader scale, we need to bring more participants to the table. […] We need more players. These include both our clinical communities, as we currently have, but also a much more meaningful engagement of care providers and consumers. We need to have a full partnership of funders — not just government as has been the case to date with caBIG — by bringing in other players to help underpin this infrastructure." (Buetow 2008, p. 5)

It is remarkable that despite such strong public acknowledgments of the importance of inclusion and plurality in caBIG, curators do not seem to have been particularly effective, or indeed interested, in capturing the interest of potential users. The resource is widely seen as too complex for researchers to understand and use; the modules developed to enable interoperability are also widely viewed as too restrictive in their choice of data format and standards, and thus unfit to encompass the very diversity of research practices that they were devised to capture. A recent review of CaBIG achievements by the NCI concluded that 'the level of impact for most of the tools has not been commensurate with the level of investment' (National Cancer Institute 2011). This stark critique has led to a decrease in funding; and it is particularly striking when contrasted to the NSF argument for cutting TAIR funding, which is not tied to a stark critique of TAIR achievements (which are recognized to have been high, though of course its functionality needs to improve and evolve as research moves forward), but rather stressed that the plant community should take more responsibility for securing its long-term sustainability (Bastow and Leonelli 2010).

A detailed assessment of caBIG's failure to establish itself as the most useful digital platform in cancer research lies beyond the scope of this paper. It might be argued that this is only a question of time, caBIG curators having taken long to think through the structure and computational requirements of their resource, which delayed vital consultations with stakeholders and the management of data themselves. Indeed, the NCI review remarks that "perhaps the greatest impact of the caBIG® program on cancer research has been to gather several communities around a virtual table to help create and manage community-driven standards for data exchange and application interoperability" (National Cancer Institute 2011). 'Community-driven' is a key term here: identifying which communities are involved in cancer data analysis, and which methods, instruments and terminologies they use, is a difficult and yet foundational task for caBIG curators, and one that they arguably will take even more seriously in the future. At the same time, caBIG needs to put that knowledge to work, by implementing data searches that highlight the diverse knowledge and provenance of available data, and yet manage to address the specific scientific queries of prospective users.

The approach to data curation taken by CaBIG can be viewed as resembling that of TAIR insofar as it emphasizes (1) open access to data as to advancements in biomedicine; (2) the belief that centralized data management is not only compatible with the existing diversity in biomedical research traditions, but can actually foster its development; and (3) the belief that effective data re-use can be achieved through the successful management of data access. The similarities in the visions proposed by TAIR and caBIG might be at least partly derived from their common cultural and political embedding (these are both tools funded by US agencies and aiming to serve scientists around the world). Yet, these two resources have independent histories, employ different standards and support vastly different communities, which makes the communalities in their vision ever more striking.

Scale and the scientific usefulness of data infrastructures

The conclusion I wish to draw from my brief analysis of caBIG and TAIR is that digital infrastructures, rather than science itself, are what needs a big scale. I have shown how these databases are responsible for supporting the needs and specialised questions posed by research projects across a variety of scientific areas. Echoing the findings of STS scholarship on standards (Timmermans and Berg 1996, Bowker and Star 2000, Timmermans and Epstein 2010), both resources view standards adopted for data dissemination as obsolete if they are not immediately useful to users in their current practices. If biologists with different expertises cannot use databases to collect and re-use data to address their own specific queries, the databases have failed to fulfil their role, no matter how good the rationale underlying their construction and the amount of resources invested in their development. Since the start of their work in the early 2000s, TAIR and caBIG curators have been acutely aware of this overarching goal and of the difficulties of reconciling it with the requirement of making data travel far and wide across multiple epistemic communities. Their efforts to resolve this tension resulted in the development of databases where methodological and epistemic diversity is explicitly expressed and explored, though with differing degrees of effectiveness. TAIR and caBIG thus do not aim to fully standardise and/or unify the knowledge and data that they contain, nor to homogenise the instruments, methods and terminologies used by their user communities. Rather, they attempt to make such diversity visible, so that prospective users can take that into account when interpreting data retrieved through those resources for their own investigative purposes. This attempt has not yet been entirely successful, as demonstrated by the fact that both TAIR and caBIG are undergoing extensive revisions and caBIG in particular has been widely critiqued. Yet, these very shifts and critiques signal the extent to which responsiveness to users' research interests, and

particularly the capacity to encompass and serve as many prospective queries as possible, is the most important factor in determining the long-term usefulness and sustainability of data infrastructures.

Incorporating a large variety of possible viewpoints and prospective queries has been, and continues to be, the most complex and labour-intensive task involved in the development of TAIR and caBIG. Both databases responded to this challenge by diversifying search tools and developing sophisticated visualisations, archives of data provenance (the methods and instruments originally used to generate data) and links to biological materials (most obviously in the case of TAIR). Setting up and updating these resources occupies much of curators' time and creative efforts. caBIG and TAIR illustrate how existing diversity in data types, disciplines, resources, instruments, methods and goals is valued as breeding ground for innovation; and the 'data friction' generated by the interaction of different communities (Edwards 2010, Edwards et al 2011) is viewed as a fruitful terrain for scientific advance. For instance, ensuring the comparability of *Arabidopsis* data with data extracted from other species has been crucial to the planning and development of TAIR, though its full potential is yet to be realised; and the overly complex structure of caBIG results from the ambitious goal to include data coming from clinical and biological contexts, as well as directly from cancer patients.

So what does the notion of scale mean for data infrastructures? My observations lead towards an interpretation of scale that has less to do with the quantities of resources, data and personnel involved in the development of data infrastructures, and more to do with their scientific role as platforms towards new discoveries. In other words, the scale of data infrastructures can be measured through *the range and scope of biological questions that data stored therein can be used to address* - where range indicates the number of research areas and specific queries potentially served by the database and scope indicates the types of

organisms whose study can thus be fostered. Under this interpretation, the scale of TAIR is not captured by the number of curators working for the resource, the number of users consulting it or the quantity of data stored therein. Rather, it relates to the range and scope of queries that TAIR can be used to address: the variety of biological questions captured by TAIR search tools (ranging from the genes involved in developmental processes to the morphologies of different ecotypes or the history of specific loci) and the number of plant species and strains that TAIR can help to study.

This definition of scale is only one among many possible interpretations, and indeed we discuss the multi-faceted nature of this term in the introduction to this special issue (Davies, Frow and Leonelli, xxxx). I focus on this interpretation here because it has considerable implications for the conceptualisation of databases as forms of 'big science', particularly when compared to the idea of 'big science, little science' articulated in De Solla Price's classic 1963 account and Niki Vermeulen's recent work. Vermeulen endorses De Solla Price's view that science is becoming increasingly more collaborative, international and expensive. She extends that argument to biology and argues that the quantities involved in research projects (including the number of researchers involved and their varied locations across the globe) are affecting the quality of the research under way (Vermeulen 2009). I broadly agree with this view, and yet the focus on scale interpreted as numbers (of locations and resources) may lead to overlooking one key source of continuity between 20th century and 21st century research: the tendency of biologists to focus on very specific questions and to structure networks, expertises and collaboration around those (a tendency reinforced by current funding regimes through the focus on short-term projects). The long-term aims of contemporary biology and biomedicine involve the integrated understanding of organisms and environment as complex, interrelated systems; but this goal is pursued through division of labour, with myriads of research groups looking at different aspects of biological systems.

The curators of large data infrastructure such as TAIR and caBIG are mindful of the fact that their resources need to serve such multiple questions, so as to be of use to as wide-ranging a group of scientists as possible.

Biological research relies on a web of expertises finely tuned to specific research interests, which means that the science that databases attempt to foster is unavoidably a localised and situated affair - not in terms of the quantity and geographical/disciplinary location of researchers involved, but rather in terms of its focus on very specific questions and outputs, which can vary greatly across projects and over time. By contrast, caBIG and TAIR strive to provide overarching infrastructures that serve as many specialised uses of data as possible. This implies that their scale is necessarily bigger than the scale of the science that they support, which tends to remain fragmented into a wide variety of specialised projects. Large-scale infrastructure, as embodied by these two databases, provides a platform where biological queries can be expressed and addressed on a case-by-case basis. This view of databases as platforms for dialogue and collaboration is exemplified by the popularity of the notion of database interoperability, as discussed in the case of caBIG. Whether interoperability works in practice remains an open question: what matters to my analysis is the very existence of this discourse and the ongoing attempts to enforce it.

In closing, it is important to note that emphasising the role of databases in managing and fostering epistemic diversity calls into question the rhetoric of data re-use employed by funders and database curators alike. I illustrated how facilitating data re-use is extremely complex, requiring multiple displacements. Research moves from existing research projects to databases to new projects, with high potential for misunderstandings across different research loci. Acknowledging this means recognising that making data accessible is a different challenge from making them re-usable. This apparently simple point has often been overlooked by key science funders, such as the National Science Foundation, where until

recently expectations about data re-use were not backed up by empirical studies of the needs and expectations of actual users.[10] Many curators strive to identify and serve those wishes as well as possible, and these efforts are crucial to the success of large-scale databases; yet, relatively little systematic and empirically-grounded research grounds these intuitions, which is striking given the level of investment in these resources. Database curators have to reconcile two different sets of expectations by both users and funders: on the one hand, databases are supposed to circulate data as widely as possible, thus making it possible to conceive of biological data as global commodities which should be exploited and re-used by researchers across the world; on the other hand, most biologists continue to view data as highly contingent products of a specific research group interested in addressing specific questions, and struggle with the idea of re-using such data within a different research context, to address new questions. This tension between the current urge to globalise data and the importance of locating data into specific research contexts is crucial to social scientific attempts to understand what is 'big' in contemporary life sciences, and what it means to expand the scale of biological research, whether in a geographical, scientific or social sense.

Acknowledgments

Ranjit Singh has helped me to gather some of the empirical material on caBIG, particularly the material for Table 1. I further benefitted from the insightful comments of four anonymous referees and from discussions with participants to the 'Making It Big' workshop at the University of Exeter, March 2011, especially Gail Davies, Kaushik Sunder Rajan and Emma Frow; Alberto Cambrosio and Karen Baker; and the several database curators who

[10] Some studies have been undertaken to measure the expectations and values of caBIG stakeholders, yet as the authors conclude 'The study population is small and limited to a single institution. To support generalizability of the findings, it is necessary to enlarge the population by increasing the number of participants from the chosen institution and/or to include other academic institutions in future studies' (Novick et al 2009).

generously took time to discuss their work with me over the last nine years, particularly Sue Rhee. This research was funded by the ESRC as part of the ESRC Centre for Genomics in Society.

References

Baker, K.S., & Millerand, F. (2010). Infrastructuring ecology: Challenges in achieving data sharing. In: J.N. Parker, N. Vermeulen, & B. Penders (Eds.), *Collaboration in the New Life Sciences*. Ashgate.

Bastow, R., Beynon J., Estelle, M., Friesner, J., Grotewold, E., et al. (2010). An international bioinformatics infrastructure to underpin the Arabidopsis community. *Plant Cell*, 22, 2530-2536.

Beck, R.J. (2005). The caBIG strategic plan: What is the caBIG community? In: *caBIG Annual Meeting, 2005*. https://cabig.nci.nih.gov/2005_Annual_Meeting/2005_Presentations/caBIG_Strategic_Plan_Robert_Beck.pdf (last accessed March 2012)

Bowker, G.C., & Star, S.L. (2000). *Sorting things out: Classification and its consequences*. The MIT Press.

Bowker, G.C., Baker, K.S., Millerand, F., & Ribes, D. (2010). Towards information infrastructure studies: Ways of knowing in a networked environment. In: J. Husinger, M. Allen & L. Klasrup (Eds.), *International Handbook of Internet Research*. Springer.

Bowker, G.C. (2000). Biodiversity datadiversity. *Social Studies of Science*, 30(5): 643-683.

Buetow, K. (2008). Pathway to a new model for biomedicine: caBIG and beyond. In: *caBIG Annual Meeting, 2008*. Washington, DC. Retrieved October 28, 2010 from

https://cabig.nci.nih.gov/2008AnnualMeeting/Buetow_PathwaytoaNewModelforBiomedicine_caBIGAnnualMeeting2008.pdf.

caBIG Website https://cabig.nci.nih.gov/ (last accessed March 2012).

Covitz, P.A. (2004). Cruising the Cancer Biomedical Informatics Grid caBIG: From village to city. *caBIG Workspace and Working Group Kickoff meeting, 2004* https://cabig.nci.nih.gov/overview/Kickoff_Documents/kickoff_plenary_covitz (last accessed March 2012)

D. J. de Solla Price, D.J. (1963) *Little Science, Big Science*, Columbia University Press.

Edwards, P.N., Mayemik, M.S., Batcheller, A.L., Bowker, G.C., & Borgman, C.L. (2011). Science friction: Data, metadata, and collaboration. *Social Studies of Science*, 41(5): 667-690.

Edwards, P.N. (2010). *A vast machine. Computer models, climate data and the politics of global warming*. The MIT Press.

Eschenbach, A.C., & Buetow, K. (2006). Cancer Informatics Vision: caBIG. *Cancer Informatics*, 2, 22-24. Retrieved October 28, 2010 from https://cabig.nci.nih.gov/in_the_news/CI2vonEschenbach.pdf.

Hine, C. (Ed.), (2006). *New infrastructures for knowledge production: Understanding E-science*. London: Information Science Publishing.

Huala, E., Dickerman, A.W., Garcia-Hernandez, M., Weems, D., Reiser, L., et al. (2001). The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis and visualisation system for a model plant. *Nucleic Acids Research*, 29(1): 102-105.

Kaye, J. (2011). From single biobanks to international networks: developing e-governance. *Human Genetics*, 130, 377-382.

Knorr-Cetina, K.D. (1999) *Epistemic Cultures: How the Sciences Make Knowledge*. Cambridge, MA: Harvard University Press.

Kohli-Laven, N., Bourret, P., Keating, P. & Cambrosio, A. (2011) Cancer clinical trials in the era of genomic signatures: Biomedical innovation, clinical utility, and regulatory-scientific hybrids. *Social Studies of Science* 41(4): 487-513.

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

Ledford, H. (2010) Plant scientists fear for cress project. *Nature* 464: 154.

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. (2009). Centralising labels to distribute data: The regulatory role of genomic consortia. In: P. Atkinson, P. Glasner, & M. Lock (Eds.), *The handbook for genetics and society: Mapping the new genomic era.* London: Routledge, pp. 469-485.

Leonelli, S. (2010). Packaging small facts for re-use: Databases in model organism biology. In: P. Howlett, & M.S. Morgan (Eds.), *How well do facts travel*? *The dissemination of reliable knowledge*. Cambridge University Press.

Leonelli, S. (2012). When humans are the exception: Cross-species databases at the interface of clinical and biological research. *Social Studies of Science* 42(2): 214-236.

Leonelli, S. (forthcoming). Data interpretation in the digital age. *Perspectives on Science*.

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

Mueller, L., Zhang, P., & Rhee, S.Y. (2003). AraCyc: A biochemical pathway database for Arabidopsis. *Plant Physiology*, 132, 453-460.

National Cancer Institute (2011). *An Assessment of the Impact of the NCI Cancer Biomedical Informatics Grid (caBIG®).* Report available on the National Cancer Institute Website http://deainfo.nci.nih.gov/advisory/bsa/bsa0311/caBIGfinalReport.pdf (last accessed March 2012).

National Cancer Institute (2008) *CaBIG Pilot Phase Report 2003-2007.* Available on the National Cancer Institute Website https://cabig.nci.nih.gov/overview/pilotreport.pdf (last accessed March 2012).

Novick, Y., Escalon, L. & Rolnitzky, L. (2009) Academic cancer researchers' perspective on a federally mandated centralized comprehensive database of all cancer clinical trials results. *Journal of Clinical Oncology* 27(15S): e17576

Ochs, M.F., Casagrande, J.T., Davuluri, R.V. (2010) *Biomedical Informatics for Cancer Research*. Springer.

Parker, J.N., Vermeulen, N., & Penders, B. (Eds.), (2010). *Collaboration in the New Life Sciences*. Ashgate.

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

Rhee website http://dpb.carnegiescience.edu/labs/rhee-lab (accessed January 2005).

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

Ribes, D. and Finholt, T.A. (2009) The Long Now of Technology Infrastructure: Articulating Tensions in Development. *Journal of the Association for Information Systems* 10(5): 375-98.

Science (2011). Special issue: dealing with data, 331(6018).

Stein, L.D. (2008). Towards a cyberinfrastructure for the biological sciences: Progress, visions and challenges. *Nature Reviews Genetics*, 9(9): 678-688.

Star, S.L., & Ruhleder, K. (1996). Steps towards an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research* 7(1): 111-134.

Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T.Z., Garcia-Hernandez, M., et al. (2008). The Arabidopsis Information Resource (TAIR): Gene structure and functional annotation. *Nucleic Acids Research*, 36, D1009-D1014.

TAIR Website http://www.arabidopsis.org (last accessed March 2012).

Timmermans, S., & Berg, M. (1997). Standardisation in action: Achieving universalism and localization in medical protocols. *Social Studies of Science*, 27(2): 273-305.

Timmermans, S., & Epstein, S. (2010). A world full of standards but not a standard world: Toward a sociology of standardization. *Annual Review of Sociology*, 36, 69-89.

Vermeulen, N. (2009) SuperSizing Science: On Building Large-Scale Research Projects in Biology. Maastricht University Press.

Table 1. *The following table depicts the organizational structure of caBIG, the different roles of each domain workspace and the number of participating organizations for each of these workspaces at the initiation of the project in 2004 and in 2009. Relevant information was extracted from the caBIG website in October 2010.*

Workspace	Scope of work	In 2004	In 2009
Domain Workspaces:			
Clinical Trials Management Systems Workspace	Development of modular, interoperable and standards-based software tools designed to meet diverse clinical trials management needs. The tools developed are configurable to work with trial sites with little or no clinical data management systems in place as well as those with robust systems, and take into account the diversity of clinical research activities and local practices that exist among trial sites.	15	92
Integrative Cancer Research Workspace	Production of modular and interoperable tools and interfaces that provide for integration between biomedical informatics applications and data to enable translational and integrative research.	24	74
In Vivo Imaging Workspace	Identification of ways in which the wealth of information provided by imaging – from the molecular level to the clinical imaging of patients performed at academic and other research centers – can be shared, optimized, and integrated.	3	59
Tissue Banks & Pathology Tools Workspace	Integration, development, and implementation of tissue and pathology tools to facilitate the integration of, and access to, information from geographically-separate areas.	13	77
Strategic Level Workspaces:			
Data Sharing & Intellectual Capital Workspace	Facilitate data sharing between and among caBIG participants by addressing legal, regulatory, policy, proprietary, and contractual barriers to data exchange. Development of recommendations for policies, procedures, and best practices, preparation of white papers and comment letters on proposed policies and guidelines, development of problem scenarios that illustrate issues confronted by caBIG participants, support reviews of caBIG tools under development, and provision of education and outreach to caBIG participants, their IRBs and their technology transfer offices.	14	27
Documentation & Training Workspace	Facilitate widespread adoption, dissemination, and use of caBIG interoperable tools, standards, and data sets within the larger cancer and biomedical communities and support the creation and dissemination of documentation and training materials for caBIG-related projects and community-wide resources.	10	16

Strategic Planning Workspace	Assistance to caBIG leadership with strategic planning and vision development activities, provision of strategic insights with regard to engaging and interacting effectively with the biomedical cancer research community and creation of white papers and planning documents that help identify and prioritize additional activities for the caBIG program as a whole.	16	25
Cross Cutting Workspaces:			
Architecture Workspace	Ensuring consistent application of the caBIG development principles to the distributed groups doing the actual integration and implementation activities throughout the caBIG project.	9	25
Vocabularies & Common Data Elements Workspace	Evaluation and integration of systems for vocabulary and ontology content development, as well as software systems for content delivery. They are also responsible for developing standards for the representation of ontologies and vocabularies used throughout the caBIG system, as well as assessments of existing systems proposed for use within the caBIG.	7	37