Inside the sequence universe: The amazing life of data and the people who look after them

Tahani Nadim

Goldsmiths, University of London

Thesis submitted in fulfilment of the requirements for the degree of Ph.D.

July 2012

I, Tahani Nadim, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

This thesis provides an ethnographic exploration of two large nucleotide sequence databases, the European Molecular Biology Laboratory Bank, UK and GenBank, US. It describes and analyses their complex bioinformatic environments as well as their material-discursive environments – the objects, narratives and practices that recursively constitute these databases. In doing so, it unravels a rich bioinformational ecology – the "sequence universe". Here, mosquitoes have mumps, the louse is "huge" and self-styled information plumbers patch-up high-throughput data pipelines while data curators battle the indiscriminate coming-to-life caused by metagenomics.

Given the intensification of data production, the biosciences have reached a point where concerns have squarely turned to fundamental questions about how to know *within* and *between* all that data. This thesis assembles a *database imaginary*, recovering inventive terms of scholarly engagement with bioinformational databases and data, terms that remain critical without necessarily reverting to a *database logic*. Science studies and related disciplines, investigating illustrious projects like the UK Biobank, have developed a sustained critique of the perceived conflation of bodies and data. This thesis argues that these accounts forego an engagement with the database sui generis, as a situated arrangement of people, things, routines and spaces. It shows that databases have histories and continue established practices of collecting and curating. At the same time, it maps entanglements of the databases with experiments and discovery thereby demonstrates the vibrancy of data. Focusing on the question of what happens at these databases, the thesis follows data curators and programmers but also database records and the entities documented by them, such as uncultured bacteria. It contextualises ethnographic findings within the literature on the sociology and philosophy of science and technology while also making references to works of art and literature in order to bring into relief the boundary-defying scope of the issues raised.

Table of contents

List of figures

List of abbreviations

BGI	Beijing Genomics Institute
BLAST	Basic Local Alignment Search Tool
BLink	BLAST Link
BoL	Barcode of Life
BoLD	Barcode of Life Database
CDS	Coding DNA sequence
DDBJ	DNA Databank of Japan
EBI	European Bioinformatics Institute
EMBL	European Molecular Biology Laboratory
EMBO	European Molecular Biology Organisation
GO	Gene ontology
GOS	Global Ocean Sampling
HGP	Human Genome Project
JCVI	J. Craig Venter Institute
IHSD	Icelandic Health Sector Database
INSDC	International Sequence Database Collaboration
MVZ	Museum of Vertebrate Zoology
NAR	Nucleic Acids Research
NBRC	NITE Biological Resource Centre
NCBI	National Center for Biotechnology Information
NIH	National Institutes of Health
NLM	National Library of Medicine
WAL	Women's Art Library
WSG	Whole Genome Shotgun
WTGC	Wellcome Trust Genome Campus

Acknowledgments

This thesis is the product of prolonged and ornate wanderings between and across themes and disciplines. It was supported by a Whitehead Scholarship, awarded in 2007 by the Centre for the Study of Invention and Social Process, Department of Sociology, Goldsmiths. My supervisor Mike Michael stood by me through all distractions and follies, providing unwavering guidance, encouragement and practical support. I would therefore like to express my sincerest appreciation and gratitude for his endurance and mentorship. I think we did make a sociologist out of me in the end.

I would like to thank my colleagues and friends who have provided support, insights and encouragement, in particular, Ann-Christina Lange, Noortje Marres, Jenn Barth, Uli Beisel and Alex Wilkie. The generosity and tolerance extended to me by Dr. Jacqueline Cooke and Dr. Vanda Playford deserve a special thank you.

I would like to express my sincerest gratitude to my magnificent friends Anna Soucek, Gabriela Flores Zavala, Isabel Waidner, Catherine Grant, Melissa Castagnetto and Althea Greenan. Sophie Macpherson, especially, made the final stages of this PhD unexpectedly sunny. None of this would have been possible without the love, support and patience of my wonderful family, Helene, Hassan and Ahmed Nadim; Susi Goldmann; Alfi, Philip, Anna, Eva, Susi and Franzi Zoubek; and Ragnhild Rød.

Lastly, I wish to dedicate this thesis to my grandmother, Emilie Zoubek.

Chapter 1. Upsetting the database logic, towards the database imaginary

This thesis is a qualitative sociological study of two DNA sequence databases, EMBL-Bank and GenBank. Using ethnographic methods, including ethnographic interviews, field notes and non-participant observation, this research is a response to the simple explorative question: What happens at these sequence databases? The central argument brought forward in this thesis is that these databases constitute novel sites for discovery (the "sequence universe"), for both the biosciences and social science. This chapter presents the rationale for this research – the articulation of a *database imaginary* to complement the database logic that is prevalent in most critical engagements with (DNA sequence) databases. It details some of these engagements before portraying EMBL-Bank and GenBank by means of historical and scientific descriptions. Following on from this, I situate the two amidst other bioinformational resources and tools, advancing the suggestion that an appraisal of EMBL-Bank and GenBank needs to take into account this "sequence universe", a heterogeneous assemblage of people, data, objects and environments that defies conventional spatial and ontological boundaries. In doing so, I place this research amongst the literatures of science studies, and, more broadly, concerns around *naturecultures* (Haraway 1997). This chapter concludes with specifying terms and limits and providing a chapter overview.

Introduction

In 2001 I was working my first proper job as a programme manager for the Women's Art Library (WAL), an arts organisation that combined a membership scheme for women artists, a library and archive of women's art and feminist art history, the publication of an art magazine (*Make*) and a public events programme. A few years earlier, while still on my undergraduate degree programme (in art history), I had assisted in cataloguing the

archive of the Artist Placement Group (APG), a formative art group that was active in the UK and Europe from 1966 to 1989.[1] This entailed me sifting through filing cabinets in the house belonging to one of APG's co-founder (together with Barbara Steveni), the late John Latham (1921-2006). Sitting in a tiny room overflowing with *stuff* I logged everything I found – notes, letters, manifestos, film rolls, business cards, diaries, scrap paper – in a Word document. No spreadsheet or table, no classifications, not even a standardised vocabulary: I just recorded what was present at the time in any terms available to me. At WAL I found myself working as part of a small group of committed women, not least the "archivist" herself.[2] I learned about the intricate lifeworld of the archive, its human and nonhuman constituents and the ever-changing and often precarious relations between them. There were systems in place: an MS Access database for managing members' records as well as the library's holdings; shelving systems for grouping publications of the same order (group exhibition catalogues, solo exhibition catalogues, *catalogues raisonnés*, art history, etc.); filing cabinets filled with thousands of 35mm slides arranged by artist name in hanging folders; stacked archival boxes containing ephemera and objects, some labelled with artist or artist group name, others labelled with names of specific occasions or movements. The order of the archive was both systematic and *ad hoc* and the only reliable finding aid was the archivist herself.

Archives, for me, have always been very real, messy, affective and unpredictable *places* – far from the guardians of grand narratives suggested by Derrida's *Archive Fever* (1996) but instead contingent collectives faithfully caring for people, materials and their stories. In the course of carrying out research for an event on cyberfeminist art, I read Lev Manovich's *The language of new media* (2001), a landmark account of the emergence of

[1] APG (1966-1989) was an artist initiative that organised placements of artists in UK and European industries and public institutions. Negotiating unique agreements with central and local government and industries, they formed a number of "artist-with-government" and artist-with-industry associations. See Slater (2000) and Rasmussen (2009) for detailed appraisals of the APG's important contributions to re-thinking the artist/public relationship and the concept of "intervention".

[2] She is reluctant to call herself "archivist" because, as she put it, "if I were an archivist I would've created a system of archiving which would have worked efficiently without me. An archivist creates records. Me, I could find things, and this was because I'd worked as a go-between or conduit for so long or I'd arranged to put it there. An archivist is a slave to the rules and imposes them on the material, but I was a slave to the archive! I'd been absorbed into its ecology and became an able guide."

new media and digital culture. I found myself reacting strongly to his claim that the *database* had replaced the *narrative* and become a new "symbolic form" that "represents the world as a list of items" (2001, p.225). I thought that such a verdict foreclosed a lot of interesting questions one could ask of databases, especially in light of the recently completed draft sequence of the human genome whose release coincided with the publication of Manovich's book. Surely a database cannot be so different from an archive: Each database must come with its own history and lifeworld, with people that work there and materials that live there and a vibrant traffic of knowledges, gestures, commitments and, indeed, *narratives*. And so I embarked on the present study which explores two databases at the heart of bioscientific research, EMBL-Bank and GenBank.[3]

Together with the DNA Database of Japan (DDBJ), EMBL-Bank and GenBank form the International Nucleotide Sequence Database Collaboration (INSDC), established in 1987, which provides the world's most comprehensive collection of nucleotide sequence information. EMBL-Bank is a product of the European Molecular Biology Laboratory's (EMBL) European Bioinformatics Institute (EBI), located within the Wellcome Trust Genome Campus in Hinxton, near Cambridge. GenBank is produced by the National Center for Biotechnology Information, a division of the National Library of Medicine (NLM), located on the campus of the US National Institutes of Health (NIH) in Bethesda, Maryland, USA. This thesis explores the two databases by means of ethnographic observations and interviews with participants at EMBL-Bank and GenBank as well as through the analysis of a range of objects encountered in the course of my research and ethnographic travels.[4] It contextualises ethnographic findings within the literature on the sociology and philosophy of science and technology while also making references to works of art and literature in order to bring into relief the boundary-defying scope of the issues raised. In the following,

[3] I use the term "bioscientific" research to refer to a broad range of sciences that work with nucleotide sequence data. Although there are alternative terms, such as "genomic science", that point to the centrality of DNA and RNA in their organisation and direction, they tend to be primarily associated with biomedical research. Although this plays a very large part in bioscientific research, I do not want to skew my perspective towards the notion of "human" (as in "human benefits") concerns. Hence, bioscientific research encompasses all research dealing with organic matter on a molecular level.

[4] All interviewees have been anonymised. Respondents at EMBL-Bank are coded using the prefix "EB" followed by a number (1-6) while respondents at GenBank are coded using the prefix "GB" also followed by a number (1-24). See chapter 3 for more details on the research methods applied.

I shall introduce some of this thesis' key terms (database logic, sequence universe, bioinformational artefact) while providing some historical and disciplinary context for EMBL-Bank and GenBank.

From database logic to database imaginary

While Manovich was admittedly more concerned with the formal aspects of the database as a new experiential medium, others were focusing their critical attention on the effects of the "computerisation of society" (to borrow the title of the famous report by Simon Nora and Alain Minc).[5] In the winningly titled book *Database Nation* (Garfinkel 2000) the author recounts a proposal put forward by the US Bureau of the Budget to establish a "database that would contain every person's electronic birth certificate, proof of citizenship, school records, draft registration and military service, tax records, Social Security benefits, and ultimately, their death records and estate information" (2000, pp.13–14). This Orwellian picture of total capture by data and of indiscriminate equivalence is frequently evoked in relation to personal health data where it is enrolled in conveying a new kind of state control over citizens – what Rose and Novas call "biological citizenship" (N. Rose & Novas 2008). Others, like Gugerli (2009), offer a more mixed account in arguing for the database as a qualitatively novel signifying practice that can suggest new ways of knowing oneself and the world. The conceptions of the database as new symbolic form, new technology of governance (P. Miller & N. Rose 1990; Aas 2004) or, indeed, new mode of production (Poster 1990; 1995) share a commitment to what I shall call a "database logic". This logic posits that the database is inherently a rational, universal and coherent operation. In doing so, it already offers a range of conclusions: Databases alienate, calculate, process, order and do not answer to anyone. While Mol (2008) has revealed a more nuanced meaning of "logic" in relation to locally cohering *practices*, the database logic purported in the above-mentioned accounts remains remarkably unbothered by practice. By and large, they do not concern themselves with

[5] The *Computerization of Society* was published in 1978 as a report to the French president Valery Giscard d'Estaing. Similar to Lyotard's *The Postmodern Condition: A Report on Knowledge*, also commissioned by government (by Quebec's Conseil des Universités), it outlined the anticipated impact of information and communication technologies on knowledge, society and economics.

what actually happens at specific databases. Neither do they examine the wider organisational ecologies in which databases are embedded.

But as the use of databases proliferates and becomes more and more ingrained in practices, critical attention has begun to channel into situated engagements with specific database projects. Here, science studies have contributed empirical or in any case situated accounts of individual databases, most notably the Arabidopsis Information Resource (Leonelli 2007a; 2008) and a mouse genome mapping resource (Hine 2006).[6] At the same time, scientists themselves have started critically assessing their own database resources, such as the protein database Swiss-Prot (Bairoch et al. 2004), the *Drosophila* database FlyBase (St. Pierre & McQuilton 2009; Tweedie et al. 2009) or the *Saccharomyces* Genome Database (Dwight et al. 2004). Concurrently, an interest in the nature of data has emerged as "data should be conceptualized not as the end-products of research, but as part of an evolving data stream." (Hilgartner & Brandt-Rauf 1994, p.359) This certainly encourages empirical study of the issues and devices surrounding scientific data production, management and distribution. Instead of a mere intermediary, "[transporting] meaning without transformation", the database emerges as a mediator whose "input is never a good predictor of their output, their specificity has to be taken into account every time." (Latour 2005, p.38) Importantly, databases are used by scientists not just to search the literature but as part of their experimental activities: identifying genes or determining gene structures, comparing sequences across species or predicting biological function. The kinds of databases used in these activities include model organism databases, bibliographic databases, sequence databases, taxonomy databases, protein databases as well as databases that cater to very specific matters, such as the *Homophila* Database for human disease genes that have cognates in *Drosophila*. In addition, there are "metadatabases" such as the Gene Ontology database that maintain the vocabulary and authority lists for the already mentioned databases.

Databases provide an informational environment but they also provide material environments through which a whole range of objects is (recursively) constituted. By

[6] Science studies have also begun to examine how scientist are *using* data, looking in particular at issues of re-use and the everyday practices of data management (Bowker 2000; Birnholtz & Bietz 2003; Zimmerman 2008).

empirically exploring two databases this thesis seeks to account for this rich bioinformational ecology and attend to some of the *situated* practices which build and maintain it. Rather than commencing from a database logic then, the present research is concerned with assembling a *database imaginary*, recovering the "[u]nexpected things" and "creativity" that often underwrite any logic (Mol 2008, p.8).

How biology learned to love the database

Biology, like most every other science, has come to encompass the production and processing of vast amounts of data. But no other science encounters quite the amount of anxiety around this development. The spectacle of data generation certainly harbours a need for caution, from concerns about data ownership in the wake of the Human Genome Project (HGP) to uncertainties besetting informed consent procedures and continuous worries about the adequacy and safety of technical infrastructures. Genomic research itself has now come to a point where concerns have squarely turned to fundamental epistemic questions about how to know *within* and *between* all that data. In particular, the development of so-called next-generation sequencing or high-throughput sequencing, capable of generating massive amounts of data (as much as 1 billion bases) in a single experiment that can be completed in a matter of days has led to much self-reflection. A special issue of *Nature* (September 2008) entitled "Big Data" gives voice to some of the most pressing concerns over the data "torrents":

> Ultimately, that could mean the genomes of most of Earth's 1.8 million named species, along with individual variants produced by projects such as the '1000 Genomes' programme for humans. And there's all the rest of the quantifiable information about life on Earth – data on protein structure and function, biomolecular interactions, signalling and metabolic pathways, and much more. The challenge is to make sense of the deluge. (Waldrop 2008, p.22)

Making sense is exactly what EMBL-Bank and GenBank are engaged in as storing and preserving data go hand in hand with making data (more) intelligible. At the same time, they are devising new formats and protocols for incorporating new kinds of data, inventing new tools for processing, visualising and relating data, and continuously training and advising the wider bioscientific community. As Dwight et al. point out, "the collection

and assimilation of data is, in itself, not enough to make a database useful. The data must be incorporated into the database and presented to the user in an intuitive and biologically significant manner." (2004, p.9) Thus, recalling the image of an "evolving data stream", sequencing DNA is really just the beginning and routing and accessioning data into the database are but two small steps in a multi-layered, recursive, distributed effort.

The concerns over data management betray another significant development, the emergence of new fields of enquiry. On one hand, this saw the development and application of computational tools and resources for utilising biological data as well as the development and application of methods, modelling and simulations based on those tools and resources in the study of biological systems. The production and availability of biological data has made it possible to profile the activities of molecules within a cell and cell populations and the interactions between every genetic element, from molecules to supracellular structures. It has given rise to new fields such as epigenetics, phylogenetics, synthetic biology, metagenomics (detailed in chapter 6) and many more "omics".[7] On the other hand, the refined attention to data matters reveals a shift from quantity to quality, from generation to integration, from elemental unit to system (or phenomenon). This is indicative of the post-HGP era where "biological questions can be approached from levels ranging from single genes and proteins to cellular pathways and networks or even whole genomic responses" (Pevzner et al. 2001, p.6) and, we can now add, whole environmental genome responses (Venter et al. 2004). The scope here includes all life on Earth and beyond (astro- and exobiology). This also points to a more qualitative take on "scale": Honing in on molecular micro-levels unleashes all kinds of complexities that demand large-scale infrastructures while enacting one of many versions of simplicity – information gain brings information loss (Strathern 2004a).

Critical responses: data begets life

While the achievements built on the steady increase of DNA sequence data are certainly plentiful, there is good reason for retaining a critical disposition towards unbridled

[7] Apart from genomics, which encompasses the quantitative study of genes and regulatory as well as non-coding sequences, there are transcriptomics (the study of gene and RNA expression), proteomics (the study of protein expression), metabolomics (the study of metabolites and metabolic networks), pharmacogenomics, and toxicogenomics. And the list keeps growing.

acceptance of the paradigms underwriting these developments. Especially given that some of the key terms remain fundamentally uncertain. These include "the gene" (Fox Keller 2000; Dupré 2004), heredity (S. Fuller 2009), "species" (Eglash 2011), "the genome" (Calvert 2007) and the relationship between phenotype and genotype expression as well as the role of the environment (Fox Keller 2010). One of the most concerted critiques raised by science studies is directed at the perceived conflation of information and life. While the present research is very much indebted to science studies and its technoscientific critique, it is also an attempt to recover novel terms of scholarly engagement with bioinformational databases, terms that remain critical without necessarily reverting to the database logic outlined above. In the following, I will detail some of the conventions which underlie technoscientific critique.

The story of biology in the 20th century is often told as a story of convergence of biology and computer and information science by both historians of science (Lenoir 1999; Hagen 2000; 2011; Chow-White & García-Sancho 2012) and science studies (Fox Keller 1995; Kay 2000; Franklin et al. 2000; Franklin 2000; Dillon & Lobo-Guerrero 2009). For the latter, the molecularisation of biology meant an informationalisation of life, an unsettling of the ontological basis of life that brought about a "discursive shift and a reinvention of history, a reconfiguration of epistemic, experimental, and social structures, and a remaking of what had been the space of representations of molecular biology before the 1950s." (Kay 2000, 39) The concerns expressed on the basis of this convergence are perhaps best evoked by Waldby's study of the Visible Human Project (VHP). For Waldby, the VHP exemplifies the "production of readability" – turning biological matter into a readable and therefore processable text (Waldby 2000, p.29). A similar position can be seen in more recent accounts that seek to come to terms with "bioinformation". This is seen as a radically new kind of entity which upsets traditional nature-culture borders while its generation and circulation renders a new information-based resource economy (H. Rose 2001; Harvey & Mcmeekin 2002; Parry 2004; Thacker 2005). In this economy, bioinformational databases and similar resources, particularly biobanks, converge

commercial, governmental and scientific rationalities while creating hybrid shareholders.[8] A fundamental concern relating to the interweaving of biological and informational orders relates to the *deterministic capacities* of information technologies.[9] As Waldby writes: "if organisms are posed as systems of information for example, this knowledge works for the array of biotechnologies (e.g. polymerase chain reaction, viral vectors, 'reading' enzymes, gene databases) which are designed as informatic tools, which encode, decode, record or reprogramme." (Waldby 2000, p.28) Once committed to the language of information and its standards and classification systems, it is computational logic which will set the "material conditions of possibility" (ibid., 37). She suggests that through projects such as the VHP or the HGP "the limits of the human as species is set out as a large yet finite information database, a spatial, graphic ordering which acts as a digital archive" (ibid.). The incursion of information technologies such as databases on biology and the subsequent re-definition of life according to computational logics are seen to radically flatten ontologies as they partake in "a multi-stranded attempt to systematically organise and productively manipulate a totality of knowledge concerning living things" (Mackenzie 2003a, p.318).[10] DNA sequence databases have emerged as a prescriptive model for this data capture, a model that oftentimes occults the uncertainties and contingencies that characterise experimental work in the wet lab (Fujimura & Fortun 1996). Flower and Heath (1993) have expressed particular concern in relation to GenBank's authority over a

[8] Comprehensive collections in this vein are Tutton and Corrigan's collection *Genetic Databases* (2004), Häyry et al.'s collection *The Ethics and Governance of Human Genetic Databases* (2007), and Gottweis and Petersen's collection *Biobanks: Governance in Comparative Perspective* (2008).

[9] On one hand, this anxiety relates to the foreclosure of accountability and contestation in relation to technical objects: Once accepted and ingrained into scientific method in the form of, for example, a unit of measure, an instrument or a reference value, more efforts are required for effective contestation. On the other hand, the flight from the real points to scenarios that elude political due process by quite literally circumventing traditional political fora. This has proven particularly worrisome in developments of national biobanks like the UK Biobank (Tutton et al. 2004; Gibbons 2007), the Icelandic Health Sector Database (Pálsson 2008), the Swedish UmanGenomics (Hoeyer 2003; H. Rose 2006), or the Estonian Genome Project (Korts 2004). Pálsson (2008), for example, showed how decisions on the IHSD were rushed through parliament while Hoeyer and Tutton (2005) demonstrate how the way in which "ethics" are enrolled by UK Biobank prevent the articulation of other kinds of concerns. Rooted in the institutionalisation of ELSI (ethical, legal and social issues) that began with the Human Genome Project, much of the work on genetic databases raises issues such as intellectual property, commercial exploitation, trust, benefit, consent while also introducing questions of governance (Petersen 2005).

[10] Critiques of information formulate a similar concern, arguing that the endless production and availability of information gives rise to novel governmental arrangements that render the citizen into commercially viable data (Poster 1990, 1995) and predicate "empowerment" and "agency" on the gathering and processing of information (Cruikshank 1999).

consensus on standards and classifications for recoding DNA sequence data. To them, databases such as GenBank have become "engine[s] of genetic governmentality" (1993, 30) – a turn of phrase which, like Waldby's assessment, carries a distinct Foucauldian inflection. Indeed, Foucault's work remains instructive for critiques of biotechnologies, from polymerase chain reaction (Rabinow 1996) to more recent work on genetic databases (Tutton & Corrigan 2004).

In summary, much critique is levelled at the rationalising and reductionist processes of databases, implying the notion of a *database logic*. In concert with biopolitical classificatory systems and standards and new economic formations, they are posited at the heart of the molecularisation and concurrent informationalisation of biological sciences. From this position, DNA databases and biobanks emerge as formidable processing plants whose data streams and calculative procedures reinforce the new molecular "governance of life" (Gottweis 1998) concerned with harnessing and harvesting bodies and body parts. The assumption is that if calculation co-produces calculable bodies, then conversely, databases produce bodies-as-data. Focusing on what gets lost in translating phenomena into bioinformatic artefacts is one way of taking a more ethical and political stance towards such systems (Bowker and Star 1999). Though these works forge important trajectories for relating genetic databases to wider socio-political concerns they do so by foregoing an engagement with the database *sui generis*. A too stringent conception of database logic aligns technical affordances offered by databases (data storage, data generation, data exchange, data transmission) with socio-political processes (avalanche of data puts an end to privacy; people rendered into mineable data vessels). For example, it does not permit encounters with and in the database, obscuring the people, machines, artefacts and practices which one can find there.

Yet, many of the works quoted above contain residuals and openings that allow for more imaginative engagements with, and less deterministic assessments of, data and databases. Before outlining these in light of the present research, I will introduce the databases, EMBL-Bank and GenBank, as well as some attendant resources and tools, which *in toto* form what I call the "sequence universe".

The databases: EMBL-Bank and GenBank

The International Nucleotide Sequence Database Collaboration (INSDC) was formally established in 1987 by its three collaborative parties, GenBank, EMBL-Bank and the DNA Database of Japan (DDBJ). Data between the partners are exchanged on a daily basis and are provided free of charge. Thus, data submitters only need to provide sequence data to one of the databases. Submissions are then incorporated in a single namespace, meaning that a search yields the same results regardless of which database is searched (Guy Cochrane et al. 2010). Data is stored in database records which contain the nucleotide sequence and its description, scientific name and taxonomy of the source organism, and a feature table that specifies coding regions and any other sites of biological interest.

The INSDC represents the biggest initiative in public domain data sharing but one instance amongst thousands of bioinformational data resources.[11] Data can be accessed and/or downloaded by anyone through a web browser. INSDC databases are data *custodians*, not owners. This means that original data submitters retain ownership as well as editorial control over sequence data submitted. Most journals require, prior to publication of an article, deposit of any protein and/or nucleotide sequence in public databases. Therefore, in order to publish, scientists are required to submit their sequence data to the databases. The continued success of INSDC databases – and although there remains discontent over data quality, there is little doubt that they are fantastically successful – is very much due to researchers submitting their data, to journal editors endorsing submission requirements, and to legal frameworks mandating researchers to deposit data generated by publicly funded research into public repositories.[12]

[11] In response the proliferation of resources, the journal *Nucleic Acids Research* (NAR) began publishing its annual Database Issue in 1993. This contains descriptions of the most important database resources and helps researchers keep track of new developments. The 2008 issue listed 98 *new* databases. Not coincidentally, NAR was the first subscription-based journal to move to an open access model (in 2005). This followed from a letter published in *Science* magazine entitled "Building a GenBank of published literature" (2001) which was signed by 24 Nobel laureates and asked for open access to published research.

[12] EBI recorded an average of 5.3 million requests per day in 2011, not including the genome portal Ensembl, compared to 4.1 million per day in 2010 (EBI-EMBL Group 2012). EMBL-Bank received and processed more than 8×10^{10} bases in 2011, a four-fold increase from 2010 (ibid.). From 1982, the number of bases in GenBank has doubled every 18 months. There are no usage statistics (publicly) available for GenBank or its search portal Entrez.

EMBL-Bank and GenBank are the principal archives for nucleotide sequence data.[13] Their data "form a core foundation upon which scientific understanding of biological systems has been assembled and our exploitation of these systems will develop" (EBI-EMBL Group 2012, p.14). Importantly, sequences are derived from across the spectrum of living matter: archaea, bacteria, eukaryotes, viruses, viroids as well as synthetic sequences and unclassified ones. Aside from the raw sequence data, that is, strings of A, Ts, Cs and Gs, EMBL-Bank and GenBank contain biological annotation and bibliographic information. Annotation is crucial to the intelligibility and processability of sequences as it provides an additional level of biological information (see chapter 5 and 6). The use of EMBL-Bank, GenBank and related resources (see below) in (post-)genomic sciences is ubiquitous. In addition to sequence data submissions, EMBL-Bank and GenBank facilitate a number of functions: Primarily, they allow researchers to compare and analyse sequence data. Yielding information about genetic orthologues in other species and paralogues within a species can reveal, for example, variations occurring between species as well as between members of a species while also making deductions about the evolutionary history of life on Earth. Sequence comparison is also used for locating the placement of a gene on the genome. Similarly, they offer information about intergenic regions (non-coding sequence) and neighbouring genes as well as details on how a gene or transcript has been assembled. Wider applications include the validation of drug targets in pharmaceutical research and the location of biomarkers and subsequent cellular processes in clinical-oriented research.

Submissions to EMBL-Bank and GenBank come from individual laboratories as well as from large-scale sequencing projects such as genome sequencing centres and environmental sampling projects. Data content in both databases is determined by the submitter, which often makes for redundant or conflicting entries. Submissions are made using web-based submission tools accessible through web browsers. These and the presentation tools available online to view and process data are developed and

[13] A DNA sequence describes the base sequence of a specific DNA molecule, specifically the sequence of its four nucleic bases (adenine, thymine, cytosine and guanine). Sequence data has theoretical (evolutionary genomics, epigenetics, functional genomics, synthetic biology and so on) as well as practical application (medical diagnostic and treatment, forensics, food science, environmental sciences and so on).

maintained by each INSDC partner independently. The fact that the databases develop tools points to a crucial aspect of their workings: EMBL-Bank and GenBank "*capture*, *preserve* and *present* the permanent scientific record" for nucleotide sequence and attendant information (Karsch-Mizrachi et al. 2011, emphasis added). These activities, as this thesis demonstrates, are anything but trivial and demand concerted efforts by humans and nonhumans within and beyond the databases. Aside from making data publicly available, they offer curation support and a range of services, from sequence similarity and text searches to complex integration with other data resources (described below). They are actively engaged in basic research, working with different scientific communities in finding ways to further bioscientific knowledge. This is why the workforce at the databases is largely made up of scientists – molecular biologists, geneticists, bioinformaticians, even (paleo)botanists.

Not surprisingly, the first archive of sequence data was created by a biochemist, Margaret O. Dayhoff (1925-1983), one of the pioneers of bioinformatics (Strasser 2010; Hagen 2011). Her *The Atlas of Protein Sequence and Structure*, published between 1965 and 1978, represented the first exhaustive collection of macromolecular sequences and is regarded as the prototype for GenBank (and all the sequence databases that followed).[14] Dayhoff's *Atlas* was the first attempt at a comprehensive molecular database. It included amino acid sequences of about 70 proteins that had been collected from published literature and manually re-keyed into the database. Finding sequences in the literature was time-consuming so Dayhoff and her group developed and applied computational methods for the comparison of protein sequences. This allowed researchers to discover distantly related sequences and duplications within sequences and to infer evolutionary histories from alignments of protein sequences.[15]

[14] Dayhoff's collection became known as the Protein Information Resource (PIR) and in 1988, mirroring GenBank's development, it entered into an international collaboration with the Munich Center for Protein Sequences (MIPS) and the Japan International Protein Information Database (JIPID). The result of this collaboration was the PIR-International Protein Sequence Database. PIR is now located at Georgetown University Medical Center (GUMC). In 2002 it joined the European Bioinformatics Institute and the Swiss Institute of Bioinformatics to create a single global database of protein sequence and function (through merging PIR-PSD, Swiss-Prot and TrEMBL databases).

[15] It was understood that small differences between homologous protein sequences indicate evolutionary change on the molecular level.

This brief historical context anticipates three key moments which the present research develops for EMBL-Bank and GenBank: databases and database work have histories and therefore continue certain practices of collecting and curating; databases are entangled with experimental research practices; and databases are embedded within heterogeneous ecologies of practices, resources and institutions.

GenBank

The original version of GenBank, the Los Alamos Sequence Data Bank, had been running since about 1979 and is credited to the efforts of Walter Goad, a scientist who had worked on the hydrogen bomb. In the early 1960s his interest shifted to the fledging field of molecular biology (then still called biophysics) and the rapidly increasing accumulation of data it caused. Working as part of the Theoretical Biology and Biophysics Group (T-10), he sought to address issues such as storage, retrieval and analysis of molecular data using computational tools.[16] In the beginning the idea was to apply "all this computer horse-power" (GB3) accumulated at Los Alamos to biological problems, more specifically, to medicine. In 1980 the NIH issued a call for proposals for a national nucleotide sequence database to which both Dayhoff and Goad responded.[17] Two years later the contract was awarded to Goad at Los Alamos where it was housed until October 1992 when it moved to the NIH and became GenBank (Benson et al. 2010).

GenBank is part of the National Library of Medicine, more precisely, of one of its division, the National Center for Biotechnology Information (NCBI) and under the overall auspices of the US Department of Health and Human Services. The NCBI, headed by the biologist David Lipman, was established in 1988 in order to provide access to biological data and analysis tools and carry out research in computational biology. It creates and maintains over 40 databases and employs staff from a range of disciplines including medicine, molecular biology, biochemistry, genetics, systematics, computer and information science and mathematics. NCBI databases include literature, molecular and

[16] The so-called "T Division" at Los Alamos is the Theoretical Division responsible for basic research and applications in mathematics, chemistry, biology, engineering and physics. It was founded in 1974 by George I. Bell to apply physics' expertise to immunology.
[17] For a more detailed account of this history see Smith (1990), Strasser (2010) and Hagen (2011).

genomic databases. Currently, over 380,000 species are represented in GenBank and new taxa are being added at a rate of about 3,800 a month (Benson et al. 2010).

EMBL-Bank

Following a call by leading European biologists to set up an organisation and laboratory for cooperation in molecular biology, the European Molecular Biology Organisation (EMBO) was founded in 1963. Initially funded by the Volkswagen Foundation and smaller grants from the government of Israel and Interpharma, it moved to an intergovernmental funding model in 1969 with the establishment of the European Molecular Biology Conference (EMBC) which, 5 years later, created EMBL (Tooze 1974).[18] It took some time to convince European member states to provide long-term support since "molecular biology did not need a major piece of experimental equipment whose construction and operation costs demanded that governments pool their resources to make it affordable" (Krige 2002, p.548). The objective was to create a central laboratory in Heidelberg, Germany, and to further a network of scientists and their interactions across Europe. This was very much seen as a response to the perceived advances of US biosciences (Strasser 2003). To this day, EMBO remains a central organisation for European life sciences, providing scientists with training, career development and networking activities and influencing European science policy while also undertaking research and developing community tools and resources.

EMBL set up the EMBL Data Library in 1980. The idea for a central resource was first discussed during a workshop convened in Schönau on the application of computers in working with sequence data that saw a centralised European data repository as "an ideal project to consolidate and legitimate EMBL" (García-Sancho 2011, p.77).[19] The Library became the first globally supported central resource that made nucleotide sequence data freely available. While other centralised efforts were being undertaken elsewhere, most notably Dayhoff's *Atlas* and the Los Alamos Sequence Data Bank (see above), the EMBL Data Library was the first to secure "central support for (…) a permanent resource"

[18] EMBC is EMBO's intergovernmental funding organisation which solicits contributions from 27 member states including Turkey, Switzerland and Israel.
[19] Historians of science Krige (2002) and Strasser (2003) have charted the often difficult history of EMBO and the EMBL Data Library.

(Hamm & Cameron 1986, p.5). With a full-time staff of seven, its first release, on magnetic tape, was made in April 1982, leading to an article in *Nature* proclaiming "Europe leads on sequences" (Walgate 1982).[20] Data contents were mostly abstracted from journals though there were some direct submissions by authors. By 1985 a yearly compendium of data was published in collaboration with GenBank.

EMBL-Bank is now produced by the European Bioinformatics Institute (EBI), which was established in 1994. Like the NCBI, the EBI carries out basic research, develops and distributes resources and tools, and organises training and education for researchers from Europe and beyond. Its major funders include the EU, the NIH, the Wellcome Trust and UK Research Councils.[21] The database has now been subsumed under the European Nucleotide Archive (ENA). This consolidates EMBL-Bank data with data from the European Trace Archive and the Sequence Read Archive (SRA). The latter contain the "*raw* raw data" (EB1), the raw reads from capillary electrophoresis platforms, DNA sequence chromatograms or traces (in Trace), and data by sequence runs performed by next-generation sequencing platforms (in SRA). ENA too provides submission services and tools, search services and data presentation and retrieval services. As of October 2010, ENA overall contains approximately 500 billion raw and assembled sequences and completed genome sequences for 1,400 cellular organisms and for 3,000 viruses and phages (Leinonen et al. 2010).

Beyond the databases, the sequence universe

GenBank and EMBL-Bank do not, *on their own*, provide much comprehensive contextual information. They are basic primary data archives and represent the foundational DNA datasets on which all genomic (and almost all protein sequence) is based. It is however difficult to think about, or indeed work with, GenBank or EMBL-Bank "on their own". It is

[20] The tapes were distributed free of charge and ordered by writing to the EMBL Data Library which sent a package containing the tape, a printed version and a user manual defining data formats, contents and indices used. Remarkably, no restrictions were placed on the re-use or re-distribution of data and third party services did indeed distribute the Library through on-line means.

[21] More recently, the UK government allocated capital funding to its most ambitious project to date, ELIXIR, a pan-European research infrastructure for biological information in Europe. This will also fund an "Industry and Innovation Suite to promote the use [*sic*] biological information in applications in medicine, biotechnology and the environment" (EBI-EMBL Group 2012).

not only that the three databases in the INSDC exchange new data on a daily basis. They form the heart of a complex suite of resources and tools, some developed by their respective hosting institutions others, because all data is in the public domain, developed by third parties. In exploring the databases, as is described in chapter 4, it is often difficult to know when one has left one resource and entered the domain of another. These complex meshings also testify to their multi-faceted applications as they serve a variety of different user communities, not just biomedical research but ecologists, taxonomists, botanists, mycologists.

NCBI currently maintains and distributes about 42 databases. Recent additions include the BioSample database that provides annotation for biological samples used in studies submitted to NCBI; CloneDB, which supports descriptions, sources and statistics on available genomic libraries and clones from genome-sequencing projects; PopSet which is a collection of related sequences and alignments derived from population, phylogenetic, mutation and ecosystem studies submitted to GenBank (Sayers et al. 2010). One of the most widely used databases at NCBI is PubMed, a database of bibliographic citation covering over 24,000 life science journals dating back to the early 19th century. PubMed, together with the BLAST tool (described below), provides the key resource for making sense of data as it provides "a crucial bridge between the data of molecular biology and the scientific literature" (Sayers et al. 2010). The Taxonomy database, a curated set of names and classifications for all organisms represented in GenBank, forms another crucial service as it guides the organisation of all NCBI biological databases. Lastly, RefSeq, the reference sequence database, contains, as the name suggests, reference sequences which have been synthesised from information from across different sources. These are non-redundant, highly curated and verified sequences that "provide a foundation for uniting sequence data with genetic and functional information." (Pruitt et al. 2002)

The EBI maintains over 63 databases which on average receive 3.5 million hits per day. Resources at EBI include gene expression datasets like ArrayExpress Archives, which stores functional genomics experiments and the Expression Atlas, a curated subset of ArrayExpress; protein repositories such as UniProt, a comprehensive resource for protein sequences maintained in collaboration with the Swiss Institute of Bioinformatics and

Georgetown University, Washington DC; InterProt, a database classifying protein families, and PRIDE, the Proteomics Identification Database. The EBI also provides databases for macromolecular structures, small molecules, enzymes and reactions, interactions, pathways and networks, ontologies, and literature. In addition, it provides common tools and support for more specialised databases such as TAIR, WormBase and PhytoPath, a database for plant pathogens.

Making sense of the sequence universe

Genomic analysis interweaves different approaches: literature and database searches, sequence comparisons and mapping data for finding gene information by position relative to other landmarks (Dombrowski & Maglott 2003). It requires the integration of heterogeneous data and methods: homology and orthology prediction, expression data, 3D structure, text mining and phylogenetic profiling. These can bring to light, among other things, biological processes, interactions and functions, pathways and metabolic networks as well as different connections between genetic elements. Thus, integration, comparison and mapping of data and objects are vital for genomic analysis. Much of this is made possible by integrated search portals, Entrez (NCBI) and Ensembl (EBI), the basic local alignment search tool (BLAST) and visualisation tools such as map viewers. Entrez is the central retrieval system for NCBI databases and currently integrates 35 databases that together contain over 570 million records.[22] Ensembl, produced collaboratively between the EBI and the Wellcome Trust Sanger Institute, is an integrated platform for vertebrate genomics. Ensembl Genomes provides the same service for plants, fungi, bacteria, protists and non-vertebrate metazoans. The Ensembl projects integrate data from EBI-internal as well as external sources to provide comprehensive, evidenced-based annotations as well as comparative genomics resources. The key tool facilitating most genomic resources at NCBI and EBI is BLAST. BLAST performs sequence comparisons and is the most fundamental and frequent type of analysis performed on EMBL-Bank and GenBank data (Benson et al. 2010). As one respondents puts is: "So if BLAST is down we're all 'Aaah! I can't verify anything!'" (GB16) Used by external researchers as well as by staff, BLAST

[22] Entrez is described in more detail in chapter 4.

finds regions of local similarity between sequences (through calculating the statistical significance of matches in the database). In inferring functional and evolutionary relationships between sequences it provides a statistical measure of significance from which biological relevance is inferred.

Amidst the sequence universe, scientists are faced with massive quantities of noisy, error-ridden connected data (C. E. Jones et al. 2007). There is very little experimental data and support – only about 1% of genes have experimental verification to assigned functions (Sjölander et al. 2011). Aside from data quality, many other challenges have arisen and continue to accompany data generation. These include consolidation and integration of past, present and future data as well as visualisations of data in more user-centred settings. More traditional problems such as standards and consistent terminologies (Ashburner et al. 2000), or data storage and safety remain (Doctorow 2008). Another challenge concerns the business model for resources such as EMBL-Bank and GenBank (Bastow & Leonelli 2010) and how looking after your data and ensuring its reusability might be integrated into the scientific curriculum and, above all, merit system.

These challenges are at once epistemic, technical, social and cultural and suggest that bioinformational resources offer themselves to manifold sociological analysis as well as mutual benefits for all disciplines involved. Thus, this thesis seeks to contribute to an open framework for imaginative and inventive engagements with such resources, committed to the assertion that we need a "deep understanding of the nature of information infrastructures" (Bowker 2006, 127) in order to make ethical and accountable decisions in relation to the collection, storage and distribution of data – not just in relation to its production.

The present research

These brief characterisations of EMBL-Bank and GenBank contain polyvalent beginnings for critical study. For the present research, the initial question was simple: What actually happens at the databases? Having worked *in* archives, I was familiar with the effort that goes into maintaining archival structures such as catalogues, classification systems, social relations (for example, between users and archivists or between contributors of archival

materials and institutions housing the archive). I was therefore interested to see how such practices translate into the environment of large sequence databases: What work was carried out on the items received? How were they maintained? What kind of expertise was required in that work? How does a database look like as a workplace, an organisation, an institution, a scientific resource? And in response to Manovich: Where were the narrative and imaginative moments and how to recover them from the database logic?

As noted above, the convergence of information technologies and biology has given rise to a body of critical literature which espouses what I call a "database logic". Yet, some technoscientific critique also points towards non-reductionist moments. Waldby, for example, advances the suggestion that bioinformational technologies "throw into question the viability of a distinction between a natural inside and technical outside of the organism" (Waldby 2000, p.40). This is of course the premise of *natureculture*s (Haraway 1997), a cosmology which refuses the separation of the natural and the cultural and instead encounters hybrids, monsters and assemblages.[23] A growing number of work is suggesting more imaginative ways to contextualise (and critique) data and databases, taking into account the anti-essentialist insights gained from engagements with naturecultures where objects and technologies are foremost understood as socio-material assemblages. This work includes philosophical engagements with sequence data and algorithms (Mackenzie 2003a; 2006; Parisi 2010), *laboratory studies* of genomic databases (Hine 2006; Leonelli 2007b) as well as accounts of the materiality and life of data (Brouwer et al. 2003; Thrift 2005; M. Fuller 2009) and the recursive traffics between data and bodies (Rosengarten 2009; N. Myers 2008; N. Myers & Joe Dumit 2011).

This thesis is in dialogue with this work as it too describes EMBL-Bank and GenBank as a heterogeneous assemblage of people, objects and environments – the "sequence universe".[24] In order to render the contents and contours of these assemblages, it portrays four different enactments of the sequence databases: as an ethnographic site

[23] This has been taken up as both commitment and programme in science studies (Haraway 2003; Michael 2004; 2006), philosophy (Barad 2007; Stengers 2010) and geography (Whatmore 2002; Davies 2003; Hinchliffe 2007).

[24] Though I derived the term "sequence universe" from the book *Bioinformatics* (Baxevanis & Ouellette 2001), I soon found that it is a common designation for figuring bioinformational data worlds (Grabowski et al. 2007; Levitt 2009; Povolotskaya & Kondrashov 2010).

and discovery environment, as a place of work, a dialogic text and an agonistic forum. These enactments resonate with how I encountered the sequence databases in the course of my research. At the same time, they are congruent with how the databases work (and often don't work) in the sense-making activities of post-genomic science. In response to the initial question – what happens at the databases? – these enactments suggests a number of observational, experiential, interpretative as well as speculative answers.

Note on limitations and terms

Before outlining the coming chapters I wish to stipulate some riders that frame the scope of this thesis. Firstly, the questions do not concern the internal technical specificities of the databases, that is, their software, algorithms, database management systems or data architecture, although, where necessary, some basic operations are described. Neither does it examine the nature and wider implication of DNA sequence data. This brings me to my second qualification, namely the eschewal of the users of the databases. Although chapter 7 contains views from the scientific community about these resources, this thesis is not intended as an account of human-computer interface interactions or of they ways in which the databases find concrete application in bioscientific research.

I refer to the databases and associated resources and tools as "bioinformational" throughout. Unlike "bioinformatic" this term spans a much broader set of concerns and is not encumbered by any disciplinary affiliations and restrictions. Indeed, exploring the scope of the "bioinformational" constitutes one key aspect of this thesis. At its most basic, it indicates the convergence of the biological and the informational without implying any ready-made assumptions about the effects of this convergence, for example, that the biological is thereby simplified or reduced. In a similar vein, this thesis makes reference to "*in silico*". This is understood as a new and distinctly "virtual" research and discovery environment in which research processes such as experiments (also referred to as "*in silico* hypothesis testing") are carried out. More generally, as Moretti (2011) has recently argued, "*in silico*" can indicate any kind of inferences made using (mostly) computational tools. Accordingly, I suggest that the recursive movements between *in silico* and *in vivo*

and *in vitro* not only create models, simulations and hypotheses (B. Palsson 2000) but that these movements are constitutive of the sequence universe.

Chapter overview

The next chapter (chapter 2) situates the sequence database between archive and laboratory. In doing so, it reviews literature on the archive in conjunction with key studies of laboratories. The first part suggests that, like the archive, the database is a "figure seen twice" (Riles 2000), both concrete lived reality and abstract epistemic arrangement. The latter is related to what I call "database logic". In the second part, I begin to unsettle this by drawing on the figure of the laboratory as it has appeared in laboratory studies. Specifically, three laboratory lessons are discussed in detail: the laboratory as a place of mundane actions; the laboratory production of inscriptions; and the laboratory-field borderlands. These lessons are each refracted through the lens of bioinformational developments and the database, which – as a new site for discovery and experimentations – imparts novel contents to these lessons.

Chapter 3 addresses the methodological challenges posed by the sequence universe as a distributed information infrastructure. In particular, it focuses on the ethnographic conventions of co-location and site, neither of which appears feasible in encounters with the databases. Instead, I suggest, following Barad (2007) following Haraway, "diffractive methods", which are introduced through Agnes Varda's film *The Gleaners and I* (2000) before being discussed in more analytical terms in relation to the methods employed in the present research. With reference to Law (2004), I argue for *method assemblage* as an appropriate means to meet the sequence universe because it makes room for co-presence, multi-sited fields and imagination.

Chapter 4 embarks on an exploration of what I call the "sequence universe" by both narrative and analytical means. It takes the figure of the "doubtful guest", an enigmatic creature devised by the illustrator and author Edward Gorey (1925-2000), as a foil for furnishing ethnographic, bioinformational, viral and analytical landscapes. The doubtful guest, of no discernable species or motivation, appears unannounced at the gates of a stately home, which it commences to occupy in both insidious and stoic fashion, very

much like the objects encountered in the sequence universe by biologist Sandra Porter. Her "discovery" of mumps in yellow fever mosquitoes is related alongside my ethnographic journeys to EMBL-Bank and GenBank.

In chapter 5 I move inside the database and describe the work of data curators. Observing their curation practices *in situ*, I suggest that data curation continues an ecology of practices that has been developed in natural history. To illustrate the historical continuities, the chapter introduces the work of Joseph Grinnell (1877-1939), first director of the Museum of Vertebrate Zoology in Berkeley, CA. I argue that data curation is very much centred around the notion of *care* and that the interactions between sequence data and curators, carried out via customised computer programmes, demonstrate a *haptic visuality*.

The succeeding chapter takes a closer look at the database records by means of two sample records: one detailing a coding region on fungal RNA putatively associated with a phototropic response, the other documenting the 16S rRNA sequence of an uncultured bacterium sequenced as part of J. Craig Venter's *Sorcerer II* expedition. It introduces the notion of *non-commensurate reading*. This explores the flat files that make the database records as *dialogic texts* constituted and continuously re-constituted through their relations with other texts and other entities. Reading the records performatively establishes a *cumulative relationality* that points to the indeterminate and inventive ways in which the sequence universe is strung together.

Chapter 7 is centred on an open letter signed by a group of mycologists and published in *Science* (2008) that faulted GenBank's accuracy in relation to fungal sequences. The letter called for a return to practices of community-based annotation established in natural history and caused a number of favourable and unfavourable responses that related to the perceived benefit/threat of such "wikification". This chapter details and interprets the letter and responses while also describing the practice of annotation in relation to the issue of accuracy. I appraise the open letter and accuracy as effective political devices for making representation for organisms, in this case fungi.

In the final chapter, I return to my initial question – what happens at these databases? – and review the key premises established in the previous chapters. I suggest

that the enactments of the databases gathered in this thesis point to different instances of integration where integration is vague and doubtful but nevertheless effective in *making sense*. With reference to Michel Serres' philosophy of prepositions, I argue that this is primarily by means of differential relationalities that allow inventive and productive connections to emerge. In closing, I argue that database imaginaries and their pluripotent prepositions hold important lessons for the role of data and databases in biodiversity debates and actions.

Chapter 2. Figures seen twice: from archive to database and laboratory

This literature review situates the databases in relation to literature about two related sites: the archive and the laboratory. The first part of the chapter introduces the archive and how it remains a pertinent form for thinking about databases. It suggests that like the database, the archive too is a "figure seen twice" (Riles 2000): on one hand, an epistemic arrangement and explanatory tool; on the other, a site to be explored and an object to be analysed. The section concludes with reviewing the importance of "site" in science studies, particularly in the form of the laboratory. The second section is concerned with presenting and critiquing key notions from laboratory studies such as "centre of calculation" and "inscription". This serves to present the key tropes of this thesis – data curation, bioinformational artefact and the sequence universe. The third section looks at how laboratory studies have given rise to a conception of posthumanist politics that have radically decentred human agency, changed the unit of analysis and shifted focus to performative practices. In doing so, the chapter assembles texts drawn from history, cultural studies, postcolonial studies, science studies, the history and philosophy of science and human geography.

Introduction

The archive offers another, less personal, point for departure in thinking about a database imaginary. The first and most obvious is that EMBL-Bank and GenBank are officially *archives*, that is, institutions dedicated to collecting and preserving nucleotide sequence records. Although this throws up some interesting tensions, particularly around the workings of different temporal registers, in the following I attend to some of the commonalities. Like the archive, the database too has come to serve as both concrete artefact and discursive construct. This is not to say that the one is opposed to the other or that these two formations, artefact and discourse, are indeed mutually exclusive. Rather, I

would like to suggest that we can distinguish between works that study specific archives empirically from works which take the archive as a particular epistemic arrangement for modern knowledge, hence, suggesting the operations of an archival logic. "Memorization by inventory", as Le Goff put it, "(…) is not only an activity of organizing knowledge in a new way, but also an aspect of the organization of a new power." (1992, p.62) Here, the archive becomes an "ideological construction for projecting the epistemological extension" (Richards 1993, p.15) of nation states but also for equating rationalisation and bureaucratisation with documentation.[25]

Studies of archives bearing a more empirical hue can be seen as a complimentary response to this archival logic, furnishing it with case studies that exemplify its operations. For some this addressed a perceived empirical oversight, most vociferously voiced by Steedman's (2002) plea to take the archive fever literally.[26] Stoler (2002) and Arondekar (2009) challenged the primacy of the archival logic, criticising the lack of due attention to the form, the materials and situated context of the archive. Others, like Richards (1993), demonstrated the inextricable co-dependence between the ordering and preservation of knowledge and the creation and maintenance of empire. Here, the archive enacts the principal interface between knowledge and the state, forming a vital component of the imperial mission in conjunction with the more diffuse archival practices such as travel diaries or artistic iconography described by Saïd (1978).

Figure seen twice

One pertinent parallel between archive and database that has emerged for the present study concerns what Riles has termed a "phenomenon seen twice" (2000, p.26). Just as the archive, the database too has become both discursive formation *and* lived, concrete situation. It is real, literal and conceptual; it is site and metaphor; it contains histories, multitudes and expectations. And just as the archival logic has trumped over more situated accounts of specific archives so has the database logic come to surpass critical

[25] The intrinsic absurdity of this constellation – the modern bureaucratic apparatus as both outcome of *and* remedy against the chaotic vagaries of modernity – has provided a fertile ground for authors from Kafka and Musil to Sebald.

[26] Steedman's work too draws attention to this by writing about the deadly dust accumulating in the archive, which contained anthrax spores from the ancient sheep blood used in the glue for bookbinding.

engagement with the concrete workings of, and importantly *at*, specific databases. This is what I have argued in my introduction: The database logic is pervasive particularly in technoscientific critique. But I would like to offer another parallel, one that introduces the focus for the rest of this chapter. To do so, I shall briefly return to Le Goff and the *new history*, established by the Annales School, of which he remains an eminent proponent. This new history demanded of historians a more critical encounter with documents and much greater attention to questions regarding the constitution of object and subject of history. During the nineteenth century historians had developed (or rather, translated) scientific methods by establishing common standards of inquiry and verification. As Appleby et al. (1994) noted, archives became veritable *laboratories of history*, sites from which past events were reconstructed on the basis of documents. The archive, Featherstone observed, "always contains potential surprises as the life history trajectories by which material travels backwards and forwards between the known and the unknown, between rubbish, junk and sacred priceless records and icons have a high degree of contingency" (2006, p.593). Here, the archive emerges as an incongruous yet operational assemblage of objects, stories, people and indeed organisms that make the archive into an environment for curiosity, discovery and, importantly, experimentation. The "popular archive" described by Lynch (1999) as well as the many instances of archival practices in contemporary art do indeed suggest a new hybrid site, the *archive-cum-laboratory* which offers a pertinent figure for situating the databases in the present research.[27] The

[27] Lynch describes the creation (by Lynch and his colleagues) of an archive on the O.J. Simpson trial that consisted of reproductions of publicly available documents (interview transcripts, news footage etc.). Rather than riding on the archive's exclusivity (original, unique documents housed in a specific locale) which usually guarantees the archive as a place of privilege, it served as a device "for examining how legal circumstances interact with the contingencies of police investigation and 'scientific' analysis" (1999, p.78). In this instance Lynch and his colleagues subverted the "archival reason" while also turning the archive into a quasi-laboratory in which to observe and critically assess the interactions between its constituents.

There are many examples for what has been called the "archival turn" in visual art: Archival practices can be seen in Gerhard Richter's *Atlas* (1962-ongoing), a process archive of thousands of photographs, notes, diagrams and newspaper cuttings; Mary Kelly's *Post-Partum Document* (1973-79), which assembles dirty diapers and text fragments into an archive of separation; Barbara Steveni's performance *I am the Archive* (2001) in which she delivers a history of the Artist Placement Group through a suitcase filled with documents and objects that bear witness to the elusive practices of APG; Ellen Gallagher's *An experiment of unusual opportunity – Everyone's got a little light under the door* (2009) which, based on records documenting the infamous Tuskegee Experiment, explores the working relationships between science and racism. For an overview and discussion of archival practices in art see Harding (2002), Foster (2004) and Enwezor (2008) .

following therefore details some of the ways in which the issue of space has come to bear on science, particularly in relation to the laboratory, which, just like the archive, also straddles two kinds of appearances, as metaphor and as historical, concrete lifeworld.

Initial situations

For Foucault (1979; 1994) the spatialisation of discourse in figures such the clinic or the prison allowed a microsociological investigation into the technical and administrative dimensions of state power. Similarly, science studies have dedicated much attention to the question of site in the sciences. This work contends that the settings of science, the spatial organisation and order of people and things, have direct bearing on what and how knowledges are produced (Ophir & Shapin 1991; Golinski 1998; Galison & Thompson 1999; Livingstone 2003; Gieryn 2006; Powell 2007; S. Wainwright & C. Williams 2008; Finnegan 2008). On the one hand, in situating scientific practices, this literature counters the universalising tendencies that have characterised the philosophy of science in the early 20th century.[28] These accounts avow the (Lefebvrian) premise that "the social and the spatial are mutually constituted" (Wainwright and Williams 2008: 161) and aim at unravelling the socio-political, economic or cultural conditions of science. On the other hand, localising the doing and making of science has afforded rich linkages to other disciplines like geography, ecology and anthropology through trading concepts and concerns and developing a repertoire of useful terms and tools with which to pry open the black boxes populating the production of scientific knowledge.[29]

With reference to this thesis' disciplinary context, science studies, the archetypal site of science and, importantly, the *study of science* has been the laboratory. While the history of science has furnished detailed accounts of individual laboratories as "cultural institutions" (Hacking 1992, p.33), science studies have taken to the laboratory much like

[28] Though as Wainwright and Williams have pointed out, contenting oneself with the question of "how space shapes science" can at times be "unpromising" in countering universal knowledge claims (2008, p.161).

[29] There is of course a danger in extrapolating metaphors and methods from disciplines which have used them more habitually and therefore developed more nuanced engagements with them. I am thinking most prominently of geography and anthropology but also ecology, disciplines in which space and place and their relation to social practices have been at the heart of polarising debates. But at the same time recent years have seen STS methods and concerns taken up by geographers in studying things such as malaria (Beisel 2011) or fair trade coffee (Barth 2009).

traditional anthropologists have to "exotic" islands.[30] In adopting the tools of the latter, most notably ethnography, science studies, starting with Latour and Woolgar's *Laboratory Life* (1986), have produced rich accounts of laboratory goings-on, shedding light on how science produces its claims and turns them into facts. Visual art and urban studies have become equally enamoured by the laboratory not as a site of/for science but as a privileged space for experimentation and innovation in late capitalism.[31] In fact, "[l]aboratories, factories, and studios disperse and recombine in ways unimaginable half a century ago" (Galison & C. A. Jones 1999, p.534), suggesting a displacement similar to the one experienced by the database and the archive. Laboratories, too, are more (and at times less) than discrete locales.

Large-scale scientific data collections such as EMBL-Bank and GenBank can generate novel questions relating to the situation of science despite, or indeed, because they are pervasive digital resources that readily escape Cartesian conventions. As studies on global finance (Sassen 2002; Knorr Cetina & Bruegger 2002), climate data modelling (Yusoff 2009) or "invisible" infrastructures such as wireless networks (Mackenzie 2005b) have demonstrated, what we generally perceive to be "virtual" is by no means immaterial. On the contrary, information and communication technologies and the digital spaces they give rise to are entangled with material worlds that require our critical attention in order to appreciate the extent to which they re-organise existing orders and geographies.[32] In the present case, there are additional good reasons for trying to situate the sequence databases. For one, in the course of my research I encountered them as distinct *locales*.

[30] Notable historical accounts detailing individual laboratories include Tycho Brahe's laboratory at Uraniborg (Shackleford 1993), Robert Boyle's gentleman house (Shapin 1988), and the Balfour Biological Laboratory for Women at Cambridge University (Richmond 1997).

[31] In relation to urban studies, see the example of New York as a "social laboratory" in Mollenkopf and Castells (1991). The characterisation of the city as a laboratory is discussed by Gieryn (2006) while the characterisation of the factory as laboratory is discussed by P. Miller and O'Leary (1994). In the visual arts, the laboratory has not only superseded the artist's studio as a place of creative invention but has become a mode of production that is experimental, interdisciplinary and concerned with engaging issues and publics. The use of the "laboratory" as a mode of work (the Institute for Contemporary Art in London was initially conceived as a "laboratory" for contemporary art) and device by which to explore certain themes has indeed proliferated. Examples include the Danish art group Laboratory of Insurrectionary Imagination, Carsten Höller's *Laboratory of Doubt* (2006) or Olafur Eliasson's exhibition *A Laboratory of Mediating Space* (2006).

[32] The materialities of electronic waste are described by Gabrys (2011a) while recent Greenpeace reports (2011; 2012) discuss the energy consumption of the "cloud" and the "dirty data" of services such as Google and Amazon and their re-appropriation of the post-industrial landscape left by the abandonment of aluminium plants on the northwest coast of the United States.

Getting to them was not easy and once I arrived, I followed participants through different spaces – office, corridors and conference rooms – while also exploring their wider environmental and institutional settings. EMBL-Bank and GenBank are *places* where people work, where decisions are (creatively) negotiated, where multifaceted material worlds unfold and different epistemic cultures meet. This highlights the second rationale for situating the databases: Their implication in larger (institutional) arrangements such as the European Molecular Biology Organisation (EMBO), the National Institutes of Health or scientific communities (see chapter 7) means that they can occupy very definitive *positions* in relation to other organisations as well as in relation to certain issues (for example, taxonomic nomenclature).

Aside from these methodological considerations, the notion of space enters in an equally literal fashion when considering the scientific, political and institutional developments that saw the creation of the databases. EMBL-Bank is a product of an institutional development that was started in the early 1960s by scientists who sought to establish a momentum for molecular biology that would match – in scope and significance – the one already in place for atomic physics and space research (Krige 2002; Strasser 2003; Tooze 1974). And this ambition was deeply entangled with a distinct aspiration to build a common Europe. The foundation of EMBO was driven by the need to close the perceived gap between the US and Europe in terms of research productivity, the need to address the imminent withdrawal of US funds for European research as well as the migration of European scientists to the US (Gottweis 1998; Strasser 2003). In the US too, the institutional establishment of genomic science through the Human Genome Project created large genome centres in order to "realize economies of scale, and, no less important, to encourage synergistic collaborations" among different disciplines such as molecular biology and crystallography (Hilgartner 2004, p.114). This, as Knorr Cetina (1999) pointed out, ran contrary to the bench-top culture of molecular biology and therefore required not only the development of technological infrastructures for manipulating DNA but also the engineering of certain orders of communication and collaboration between the centres and between the centres and outside entities such as biological resource centres. Nowadays, the big genome centres (for example, the Sanger

Institute, the Genome Institute at Washington University, the Human Genome Sequencing Center at Baylor College of Medicine, the Broad Institute, the J. Craig Venter Institute) are flagship stores amidst a fleet of commercial and positively pedestrian sequence facilities oftentimes housed in nondescript hangars that populate industrial zones and are not surprisingly referred to as "factories" by many of my respondents. Like climate science, the development of genomics, with attention to its material infrastructure, followed a pattern from system to network to web (P. N. Edwards 2010).

We can see that the development of genomics as a discipline entailed from the start a mobilisation of performative spatial arrangements: Where science was done had an impact on how it was done and *vice versa*. If the content and method of scientific research is indeed shaped by the place in which research activities are carried out, then databases like EMBL-Bank and GenBank point not only to novel spatial constellations but equally reflect the direction of science in the wake of their "informational turn" (Beaulieu 2004). The history of molecular biology, or more precisely, of its genesis as a discipline *sui generis* has involved decidedly spatial matters. The (literal and conceptual) trope of the gene furthered a general spatialisation of knowledge concerning hereditary and phenotypic expression constitutive of the shift from classic genetics to molecular biology (Rheinberger & Gaudillière 2004). This brings me to my last rationale for situating the databases which concerns the inner spatialities projected by the genomic enterprise. The biosciences have seen a proliferation of sites and contexts that have reconfigured existing spaces and invented radically new ones: from genome and sequence centres, protein landscapes, transcription sites and coding regions, body parts and stem cells to the spaces configured by biometric security, biopiracy, personalised health care and the tissue economy. Certainly, the widespread application of biotechnologies, the progress of "geneticization", the generation of synthetic life and the indiscriminate exchange of metaphors across disciplines, particularly between biology and information science, demand of scholars a greater critical sensitivity of where and how to look.[33] With the

[33] Geneticization, a term initially coined by the writer and activist Abby Lippmann, points to a conflation of the social and the biological. It renders neo-positivist attitudes that foster the cementation of differences and inequalities in (biological) matters of facts not just acceptable (race, homosexuality, sex) but prudent and efficient (behavioural genetics, pharmacogenetics).

intensification of the molecular gaze (Nelkin & Anker 2003), that is the rapid exploration of genomes, proteins and molecular interactions, the problem of site and situation changed scales and magnitude (Barnes & Dupré 2008). We now inhabit radically new "genomic geographies" (Fujimura & Rajagopalan 2011) and "bioscapes" (Burri & Joseph Dumit 2007). At the same time, this development is often characterised as continuing the move from the field to the laboratory, a move that has taken us even further into abstractions and away from the lived realities of organisms *in situ*. This has impacts on the training and practices of biologists as much as on the paradigms guiding their epistemic pursuits.

Situating databases

EMBL-Bank and GenBank are continuations of the kinds of places that were established to house, archive or showcase scientific materials, such as natural history collections, science libraries and museums, and even scientific indices. They embody a material and semiotic dialogue with both atavism and expectation. Seen as foremost *archives*, they develop and make available sophisticated discovery tools without which bioscientific research would be unthinkable. In fact, they are designed to be discovery environments that let scientists explore and test things, much in the same way as a laboratory might do. Though the apparent schism between natural history and experimental biology is premised on maintaining a distinction between archival practices such as collecting, ordering and describing, these practices continue to play a crucial part in doing science. Not just that but they do not necessarily preclude an "experimental disposition" (Ronell 2007), which is to say that collecting, categorising and classifying share with the experiment certain creative, inventive and contingent facets. With a view to nucleotide sequences databases as archives, the continuation of natural history by other means is a legitimate claim. Certainly, the collection and recording of organismal fragments has a history that extends beyond the foundation of molecular biology. Accordingly, bioinformatic databases have been described as "virtual natural history museums" Hilgartner (1995a). Here, "[d]atabases, like earlier natural history collections, are not mere repositories; they are tools for producing knowledge." (Strasser 2011, p.63) Taking into consideration such

continuities, I contend that their workings can shed light on how the virtual is continually materialised, how the natural is forever encultured, how history is filled with expectations and futures, and how museums are also always field laboratories.

Laboratory studies shall provide as an important interlocutor for the present chapter. More specifically, it sets the scene for the thesis in its entirety as it approaches the databases via (some of) the conventions of laboratory studies. At this point, I am not undertaking an exhaustive review of the subject of laboratory studies. Rather, I want detail the conceptual (and, in the next chapter, the methodological) consequences of taking this stand. In this sense then, this chapter is about settings and set-ups: I wish to situate bioinformational databases in relation to the literature while exploring the extent to which bioinformational databases themselves constitute specific sites for doing science. Thus, I am articulating certain beginnings for trajectories that allow "points of contrast, comparison or reference for other sites and situations" (Mol 2008, p.9). "Situating" the databases is therefore neither a topographical task nor an objective in itself but a way to unravel questions and, as Mol suggests, to find surprises. For a researcher to be startled requires a prior understanding or perhaps a certain normative expectation which the ethnographic observations subsequently defy, thereby making way for surprises and findings. There is, however, another reason for setting ourselves up for surprises: For the most part, critical literature on specific genetic databases lacks surprises. In the following, I will detail three key analytical frames: laboratory work, laboratory objects and the laboratory's outside, the field. Aside from an interlocutor, this chapter enrols laboratory studies as a foil for highlighting some of the novel challenges posed by bioinformational resources as objects of study. In observing the contents of laboratories, laboratory studies have supplied a pervasive rhetoric for studying the doing of science, which this thesis makes liberal use of. But this rhetoric also serves to unravel its own limitations when applied to databases. Here, as this chapter argues, conventional understandings of epistemic things, boundaries, inscriptions and representations fall short of conveying the kinds of things and interactions harboured by the sequence universe.

Laboratory work: mundane actions

Laboratory studies follow the deflationary conventions of science studies by honing in on the micro-sociological level of the ordinary and mundane. Instead of theories and methods, they concern themselves with everyday practices, *in situ* norms, interpersonal relationships and the material furnishings surrounding the doing of science. Accordingly, they examine individual laboratories *ethnographically*, focusing their attention not so much on the place itself but on the activities and objects contained within (Latour and Woolgar 1986; Knorr-Cetina 1981; Lynch 1985; Traweek 1988). Work in the laboratory comprises the handling of heterogeneous matters, from detectors and Petri dishes to phages and other nonhuman and human collaborators. Contrary to its popular conception as a controlled and sterile place, laboratory studies revealed it to be quite messy (Law 2010). In laboratories, natural and social orders are combined to construe "workable objects" (Knorr-Cetina 1992, p.119) and "'do-able' problems" (Fujimura 1987). Put differently, through a series of "articulation and objectification", (pieces of) the macrocosm are translated into the microcosm of the laboratory (Callon et al. 2009, p.59). In laboratories, rats and other entities are transformed into "analytic objects of technical investigation" (M. E. Lynch 1988, p.266) and local occurrences "into traces" (Law 1986, p.34). How exactly, we may ask, are these translations achieved? What is the nature of work *in* the laboratory which accounts for the work *of* the laboratory?

Lynch's purposefully naïve question upon entering the laboratory, "[w]here was the *action* occurring?" (1985, p.227), anticipates a key insight of laboratory studies which has quickly come to define the core programme of science studies: Laboratories are places of very parochial goings-on like reading, counting, filling out charts, checking temperatures and cleaning. Laboratory life comprises endless uniform and routine micro-processing perhaps more commonly associated with office work.[34] As Lynch put it: "For

[34] Here we encounter another displacement of the laboratory: In what has been called the "new economy" – concerned with supposedly immaterial labour such as marketing, advertising, branding and communications – there has been a very deliberate convergence between the office and the laboratory. Ross' (2003) study of this "no-collar" workplace shows to what extent this convergence harnesses and translates the notion of the "experimental" which manifests itself not just in work practices (working on projects in small transient groups utilising playful arrangements of incongruous objects and being more concerned with generating questions) but, importantly, work relations (the incorporation of leisure time in work hours or precarious, contract-based or freelance employment). Here, the laboratory no longer indicates a privileged site of science but a

long periods of time one or a few individuals would sit silently, tapping at the keyboard of a computer terminal or scribbling notes while viewing data displays. The bodies did not move, the voices were not animated, and an ethnographer's questions were not always honoured with polite answers." (1995, p.228) More specifically, it is concerned with "*observing* or *tending* nonhumans" (Star 1995a, p.25).

Stressing the quotidian character of scientific work furthered science studies' distancing from teleological and deterministic accounts of the pursuit of truth. Yet, the micro-sociological approach to action brought with it certain difficulties, not least of which the reluctance to account for a differential distribution of power and resources that more than often preceded the observed actions. While surprise at the mundanity of the laboratory might say more about one's own expectations, Lynch's question, I would argue, also demonstrates a certain gaze which privileges *visible actions*. One of the innovations of science studies was of course the rendition of the heterogeneous layers that made-up science, the gestures and the shoptalk, the material arrangements, the squabble over merits or the circulation of papers and data. But while these fruits of ethnomethodological study assembled into a repertoire of less-than-evident activities, they also testify to a preference of *actualised* actions over others. As Mackenzie (2005a) and Fraser (2010) noted, concentrating on the circumstantial assembly of a locality, such as the laboratory, does forego engagements with the "overflows" whose "divergent" and "virtual" orders can enact this coming-together very differently.

Working with data

Scholars have argued that the wealth of (bioscientific) data has occasioned novel ways of working amongst scientists (Hilgartner 1995; Hine 2006; Zimmerman 2008; Leonelli 2007b; Leonelli 2012). The rise of bioinformational research has seen the formation of "new disciplinary identities" (Hagen 2011, p.62) and new (non-technical) challenges faced by scientists. Hagen (2011) recounts how in the 1950s biochemists regarded working and

space that renders both work and worker into experimental objects. At the same time, this increasingly reflects employment practices in the sciences (Felt 2009) where the importance of mobility and excellence as well as project-driven (and therefore time-limited) funding forces researchers into more and more precarious work situations.

building databases as "second-rate" science while computational biologists using databases to study evolutionary phenomena were criticised by traditional evolutionary biologists. For Hilgartner biomolecular databases "represent new forms of scientific interaction based on novel and rapidly evolving communication regimes" challenging, for example, the value of scientific publication and its attendant merit system (1995, p.258) . I wish to suggest that changes associated with the increasing importance of data need to be understood in relation to an earlier shift, namely the one designated by the so-called "Allen thesis": the emergence of a divide in biology between naturalist and experimentalist practices during the period from 1890 to 1950 (1975). Here, descriptive and speculative conventions, largely based on morphology, gave way to an analytical tradition that sought to find causal relations through experimental methods. This had considerable implications for the kind of work carried out by biologists as it "came to rely on standardized biological materials", like laboratory animals, entailed the "infiltration of experimental biology by physicists, chemists, and their techniques" and the development of and reliance on complex apparatuses such as "ultracentrifuge, chromatography, electrophoresis, X-ray diffraction, and electron microscopy" which "collectively opened the door to isolating and analyzing biological substances (…)". (Kevles & Geison 1995, p.100)

The rise of experimental biology also occasioned a literal shift from the field into the laboratory. The above quote suggests other attendant developments: Use of laboratory animals required the enrolment of extraneous entities and practices (such as husbandry) while also impacting on the laboratory's internal organisation. Similarly, the reliance on sophisticated machinery demanded closer collaboration with other disciplines and necessitated the import of their epistemic practices. It also suggests a plethora of mundane and practical tasks arising from the new techniques and technologies such as maintenance, verification, cleaning and measuring.

Alongside these wet labs, that is laboratories that handle biological matter, work in so-called dry labs is mostly done with computers and other electronic equipment and involves *in silico* experimentation and hypothesis-testing such as creating and running

simulations and generating, analysing and processing data.[35] This has led scholars to diagnose a clear division of labour between "computational theoretical biologists" working in "dry labs" concerned with "constructing theories of nature" and molecular biologists and biochemists concerned with "confirming" theories in the wet lab (Fujimura & Fortun 1996, p.165). Respondents at EMBL-Bank and GenBank habitually referred to their offices as "labs". Much of the work in these labs, such as the curation of data, is indeed carried out with computers – curators receive, check and annotate submissions on their PCs. Such bioinformational work has much in common with *information work*, the creation and management of formal representations and abstractions (Star 1995b; Star & Ruhleder 1996; Bowker & Star 2000). This information work involves "abstracting (…), quantifying, making hierarchies, classifying and standardizing, and simplifying" (Star 1995b, p.90). Importantly, Star suggests that the formal representations and situated organisation of information work around them are co-constitutive, that is, the "tension between formal representations and local contingencies is recursive" (1995b, p.103). This co-productive exchange between "the ability to make a thing" and "managing information about a thing" (Star 1995b, pp.101–2) makes room for two propositions: firstly, that bioinformational work with data might not be so radically different than the work carried out in wet labs; and secondly, that such work with data – while not actualised through conventionally visible means – is as indeterminate, messy and lively as work with "on-site laboratory colonies of Drosophila, yeast, slime molds, rats, mice, or guinea pigs" (Kevles & Geison 1995, p.100).

Laboratory objects: inscriptions

Much of science studies have focused on the entities populating the laboratory, from laboratory animals (Birke et al. 2007), model organisms (Leonelli 2007a) to maps (Gaudillière & Rheinberger 2004) and different kinds of machinery (Knorr Cetina 2000). Indeed, laboratory studies have considerably advanced the "object turn" in science studies (Marres 2009). Here, objects are constituted at once semiotically and materially, serving

[35] See Palsson (2000) for an overview of *in silico* biology. Moretti (2011) offers a science studies perspective on what "*in silico* experimentation" means based on discourse analysis of recent research articles.

"both as a means to and as a source of knowledge" (Morgan & Morrison 1999, p.35). In materialising earlier decisions and results an entity such as the fruit fly becomes what Rheinberger has termed a "technical object", an instrument for making trans-species deductions about, for example, reproduction. Yet, the fruit fly also remains an "epistemic thing" in that its potential for representation is continuously unravelled and tested (Knorr-Cetina 1981; 1999; Rheinberger 1997).

The inventory of prosaic actions and not-so prosaic objects, once elaborated within formalised practices of producing matters of fact, has resulted in the discipline habitually avowing Latour's aphorism "Give me a laboratory, and I will raise the world" (B. Latour 1983). While many laboratory things described above oscillate between ontological orders (e.g. animal/instrument/human/machine), the most lasting kind of object to emerge from the laboratory are *inscriptions*. These "marks" (Hacking 1992) or "traces" (Rheinberger 1997) are the outcomes of laboratory practices and translate phenomena into diagrams, plans, machine readings, specimen, standards, journal articles, or conference papers. Phenomena and their apparatuses – stem cells, viruses, phages, DNA samples, neutrinos, microarrays, PCRs – thereby become simplified and easier to process, manipulate and transport out of the laboratory. They are rendered into "immutable mobiles" (B. Latour 1990), pliable entities that can circulate beyond the laboratory while retaining their epistemic currency (which can only be challenged in the laboratory).

As such, inscriptions facilitate "act[ing] at a distance" (B. Latour 1987, p.223), whether this distance be temporal, geographical, disciplinary or ontological (e.g. data derived from animals). Here, the laboratory emerges as a "centre of calculation", a site where inscriptions combine, generate equivalences and turn phenomena into standardised units for calculative processes. Centres of calculation quite literally capture (collect, measure and render intelligible) objects, bodies and phenomena. It is perhaps not surprising that both the archive (Osborne 1999) and the DNA sequence database (Flower & Heath 1993) have been portrayed as centres of calculation.[36] The spectre of quantifiable and computable inscriptions processed through centralised calculation certainly finds

[36] It has also found uptake in postcolonial critiques that seek to unravel the role of museums, archives and other memory institutions in the exploitation of colonised territories and bodies (Richards 1993; Nair 2005; Luyt 2008).

purchase in critical accounts of the informationalisation of biology where it conveys a sense of loss – of vibrancy, complexity and the messiness associated with the world out-there.

Yet, the centre of calculation and the typical conception of inscription can "end up making rationality and calculation sound more pervasive and more organized than they are in practice" (Hinchliffe 2007, p.172). Calculation, as anthropological studies of non-Western number regimes have shown, can be playful, situated, distributed and inventive (Verran 2001; Guyer 2004; Czarniawska 2004; Maurer 2005). Instead of regarding "calculation" invariably as an instrument to rationalise difference and make localised truths hold elsewhere, these scholars highlight *situations of calculation*: embodied, heterogeneous practices which seek to temporarily reconcile (with the help of artefacts, numbers, affect etc.) different scales and asymmetrical relations. This also resonates with work that claims measuring and dreaming (Pynchon 1997; Cosgrove 2008), logic and magic (Marina Warner 2011) and objectivity and imagination (Daston & Galison 2007) as intrinsically entangled.

Bioinformational artefacts

The biosciences have generated a menagerie of ontologically ambivalent objects that precariously straddle the natural, the technical and the social such as mutant mice (Davies 2011), GM crops, stem cells (Wainwright, Michael, and Williams 2008) or xenotransplants (Michael & Brown 2004). They constitute *things* in the Latourian sense, no longer of nature but neither fully synthetic. While these things are often difficult to grasp within conventional object lessons (Law & Singleton 2005), they have nevertheless been the focus of many rich studies. Inscriptions generated around them, such as DNA sequence, on the other hand, rarely come into their own analytical focus and if they do, this tends to remain within the context of centralised, rational calculation.

Gere and Parry (2006) have suggested that the contestations which accompany biobanks and tissue collections are premised on a convergence of *material* (tissue samples) and *informational* (medical, environmental and/or lifestyle data) elements. The value of biobanks relies on the combination of materials (tissue samples, blood samples)

and information (medical, lifestyle, genealogical). In this case, a sample is only as useful as the information that accompanies it (e.g. where it was derived or how it was extracted). This produces "an entirely new class of information" that institutes novel valuative relations between data and bodies (Rose 2001: 29).[37] Though the databases which concern the present research do neither contain tissue nor blood samples, I argue that they nevertheless present equally relevant *constellations* of informational and material entities. Importantly, here, inscription does not necessarily reduce complexity. Records in databases like EMBL-Bank and GenBank bring together very different kinds of data and information (see chapter 6) and these records become differently enacted depending on how they are called upon (Mol 2002). In the sciences, the term "artefact" refers to an object that has come into being through methodological, analytical, technical or conceptual inaccuracies.[38] Interestingly, the artefact comes to be indicative not just of a deviation from or lack of a "norm" but also of the genomic landscape and its abundance of polymorphisms, variations and hybrids. Here perhaps, the conventional understanding of "artefact" as an encultured object and its scientific meaning as an unsolicited effect are closer than initially thought. Rather than an arbiter of certainty the database record accommodate many degrees of indeterminacy as well as tentative and hesitant relations.

All that is solid might turn into paper but texts have given us very wild worlds indeed. Using the term bioinformational "artefact" instead of "inscription" captures not just the dimension of construction and re-construction but also highlights the importance of (cultural) situatedness, aesthetics and genre. In that sense, the "graphematic spaces" (Rheinberger 1998) of the laboratory can occasion explorations inspired by Wynne's suggestion that the "question should be not simply how to expose and critique these simplifications and reductions, but, better, how to render them more poetic, modest and human?" (Wynne 2005, p.87)

---

[37] The way in which biotechnologies occasion exchanges between information and the body that continuously re-draw the borders of bodies has been discussed by Mitchell and Thrutle (2004) and most recently Rosengarten (2010).

[38] For example, abnormal cell behaviour in cytology specimen was shown to be an "artefact" because it was caused by contaminated equipment rather than borne out of a particular pathology (Molyneux & Coghill 1994).

The field

Laboratories have conventionally been understood as separate from the natural environment. Indeed, it was in relation to the "field" that the laboratory (and the field) maintained "distinct modes of knowledge production and have distinct political economies" (Kohler 2002a, p.18) characterised by "distinct epistemic virtues" (Gieryn 2006, p.5). It is this very distance which renders the laboratory "epistemologically advantageous" as it allows objects to be placed "in a new phenomenal field defined by social agents" (Knorr-Cetina 1992, p.117). Thus, presumptions about the kind of space constituted by the laboratory have commonly been articulated in oppositional terms to an "outside". Whereas the former was seen to be controlled, artificial, isolated and generic the latter appeared as messy, contaminated, local and unpredictable. As such, the differentiation between laboratory and field can also be read as a differentiation between culture and nature: the field as the place of nature and the laboratory as thoroughly removed, not just from wild nature but messy politics. Since the laboratory and the field are "coinvented and are mutually (and changeably) defining" (Kohler 2002a, p.3), explorations of borderlands, such as Franklin's study (2006) of a *hatch* connecting *in vitro* fertilisation with human embryonic stem cell facilities, can reveal the co-productive exchanges of naturecultures.[39]

The development of experimental biology has often been couched in terms of a retreat from *in vivo* processes.[40] This shift is illustrated by the migration of the scientific community from museums and herbaria to the laboratory. Similarly, the informationalisation of biology could easily be generalised as a flight from nature. Yet the boundaries between field and laboratory have, as Kohler's emphasis on co-invention suggests, never been particularly clear-cut to begin with. Historians of science (Kohler 2002a; 2002b; Gooday 2008; De Bont 2009) have portrayed more *ambivalent* sites like

[39] The traffic observed by and through the hatch is in itself *productive*. It constitutes a transfer of materials (donor eggs) but it is also a passage in which entities switch ontologies much like the "traffic in information as flesh" described by Rosengarten (2009).

[40] See for example the incredulity and suspicion with which fellow scientists encountered Barbara McClintock who stuck with maize and all the elaborate care and tending it required in the field instead of following her colleagues who began working exclusively in the laboratory with phages and other microorganisms (Fox Keller 1983).

marine laboratories, field stations and biological farms (*vivariums*) where scientists study organisms and processes in their natural settings albeit under controlled conditions. Science studies too have shown the extent to which the laboratory can spill beyond its walls through practices such as extracting, analysing, classifying, marking even as they occur in the jungle (B. Latour 1999) or on the beach (Gisler & Michael 2011).

Conceptualising the laboratory in a more expansive way is necessary to capture the trans- and dislocated practices and consequences of scientific research in the form of drug trials, longitudinal studies (such as the Farmingham Heart Study), GM crop trials but also biobanks and genetic databases. These phenomena inhabit spatialities that are at the same time expansive, endemic and liminal. They are indicative of less determinate geographies of science as they very actively contribute to the making and maintenance of political territories, both within nation states and beyond.[41] Studies on transnational science co-operations around issues such as biodiversity (Ellis & Waterton 2005), bio-prospecting and natural resources (Pottage 2006) but also natural disasters have demonstrated the co-constitutive exchanges between the doing of science and the making (and patrolling) of fields and territories which no longer correspond to designations such as local, regional and global. In these cases, it is difficult to ascertain distinct boundaries between outsides and insides as well as presences and absences as they become continuously (re)constituted. For science studies in particular these indeterminate situations have given rise to inventive topologies, such as *fluid* and *fire* spaces (Law & Mol 2001).

Sequence universe

Biological analysis and research not only use data but they are carried out *inside* and *with* a multitude of different bioinformational resources – the sequence universe. Does this then constitute a new "field"? To a degree, this is certainly a persuasive analogy. Like the

[41] Pálsson and HarÐardóttir (Pálsson & HarÐardóttir 2002) have stressed how nationalist assumptions over, for example, the homogeneity of the Icelandic population, have been enrolled as critical "scientific" factors in marketing the IHSD. Busby and Martin (2006) writing about the UK Biobank have also shown how national identity has become enrolled in the legitimacy of large-scale genetic biobanks, particularly in narratives concerned with national competitiveness, the national sharing of risk and responsibility, community and solidarity. Projects such as IHSD, UK Biobank and Singapore's *Biopolis* (Waldby 2009) very explicitly operate as national projects, both in the sense of building a nation and of enrolling resources, structures and actors on a national scale.

field, the sequence universe is an environment, susceptible to environmental factors, that is, dependencies that are difficult to disarticulate and that originate and reproduce *elsewhere*. It is also messy and, at times, uncontainable – populated by bad or incomplete data, widely differing standards and conventions, and a multitude of different entities from algorithms to fungi (see chapter 6 and 7). Davies raises the issue of "emerging cartographies of experimentation" that have "different temporal-spatial imaginaries, define experimentation through alternative analytical or actors' categories, and address themselves divergently to epistemological questions about scientific practice or the ontological politics of technical democracy." (Davies 2010, p.668)

How to account for the "temporal-spatial imaginaries" of boundless experimentation in relation to bioinformational databases such as EMBL-Bank and GenBank? How to map the material-semiotic settings of data practices that span different ontological registers (*in vivo*, *in vitro* and *in silico*), locales and scales? What generalisations, languages, metaphors and analogies are capable and relevant for describing these imaginaries? Greenhough (2006) takes up the figure of the "seascape" as this reflects the "uncertain and unsettling" spatialities materialised by the infamous IHSD project (see footnotes 7 and 41). Given the frequency of aquatic metaphors in accounts of bioscientific data production (streams, floods, torrents, deluge), the seascape might indeed suggest itself as an apt analogy. More importantly, it points to an expansive ecological topography that while allowing for interconnections across spaces (genetic, epidemic, national) also provides researchers with opportunities to assemble passages and trajectories for moving beyond the laboratory.[42] Such ecological situating of phenomena like electricity, worms, office buildings or hurricanes has distinguished a number of recent studies in geography, history and anthropology (Bennett, 2005; 2010; Murphy 2006; Hinchliffe 2007; Alaimo and Hekman 2008). These offer pertinent terms on which to

[42] Incidentally, Pálsson & HarÐardóttir (2002) have shown how previous public debates on fisheries in Iceland have acted as associative frameworks for debates around property and ownership in relation to the IHSD. Given that the IHSD's collapse was perceived to be an "ethical" failure, the seascape framing allows for an interesting take on the material-semiotic space of ethics: Instead of working alongside and supporting the construction of the IHSD, as in the case of the UK Biobank, which has entailed huge efforts to put ethics alongside its development, the discursive space of ethics appeared as an unmanaged, emergent and extraneous entity contaminated by contradictions and fish.

approach not just the spatial formations enacted by bioinformatic resources but also the situation of the researcher vis-à-vis such entities. Like the sea, the figure of the sequence universe conveys the endemic and dynamic expanses projected by hundreds of bioinformational resources and tools that continuously dissolve and redraw boundaries between laboratory and field, inside and outside, site and environment. Ecological approaches to human and more-than-human geographies project a continuum where these constellations become enacted through (work) practices. Here, the sequence universe can be figured as a "shifting material matrix" (Gabrys 2011b) that is variously enacted as archive, laboratory, habitat, work place or political forum.

Raising worlds: Posthuman politics

Latour's claim about raising worlds through laboratory work refers to the argument that in the laboratory, reality "becomes a second-order concept that arises as an attribute at the intersections of alternative representations" (Rheinberger 1997, p.274). The apparatuses that recreate, record and inscribe phenomena are also responsible for bringing them into existence. This remains a core commitment of science studies which, in the decades that followed the publication of the *bona fide* laboratory studies, have produced ever more nuanced (and less solipsistic) interpretations on the topic of "the real". This has come to underwrite a broad non-reductionist and non-representational ethical-political project that is radically redefining the shape and form of subject matters and their politics. No longer discrete elements that can be thought separate from apparatuses, the basic unit of analysis is the phenomenon (Barad 2007), the assemblage or the relation (Haraway 2003) where realities, differences and boundaries emerge performatively as effects of heterogeneous arrangements.[43]

[43] Here, performance is understood in the sense of Butler's performativity as an iterative process of materialisation that enacts specific normative regimes (1993). In moving toward "posthumanist" conceptions, (science) studies have extended Butler's initial concern with representational regimes to take into account the agential capacities of matter itself (Bennett 2010). Posthumanist performativity is premised on "material-discursive" practices (or *apparatuses*) where [t]he relationship between the material and the discursive is one of mutual entailment" (Barad 2007, p.152). More recently, Butler herself has begun to revise some of her earlier conceptualisations of performativity to include matter and nature. See her 2009 keynote "On the Occasion" at the Third International Conference of the Whitehead Research Project at Claremont available at http://itunes.apple.com/itunes-u/school-arts-humanities-audio/id439752557?mt=10#ls=1. Last accessed: 6 July 2012.

The notion of the laboratory as "theatre of proof" (B. Latour 1988; Shapin & Schaffer 1989) highlights the role of *performance* in laboratories while also making explicit the political import of laboratory space. This has been articulated through the dual meaning of representation, what Laclau has called the "double movement" of representation (2006, p.297). The laboratory's experimental apparatus creates inscriptions as *representations of* phenomena while making *representations for* its constituents. How representations work or are put to work in the laboratory offers a line of enquiry that has installed the laboratory (and its apparatus) as a paradigmatic figure in the realm of political associations and deliberations (Stengers 1997; B. Latour 2004; Barad 2007). For Stengers, experiments confer upon nonhumans the ability to speak which in turn confer onto the scientist the power to speak in their name (1997, p.89). It is not just that populations are rendered visible and thereby calculable, but that a population is given a way to "speak" and perhaps challenge their inscriptions. Here, the laboratory emerges not just as a site for "demythify[ing] science" (Stengers 2000, p.14) but as a model for a new kind of politics no longer premised on "modern" dichotomies of nature and society and the primacy of exclusively human agency.

The entanglements of naturecultures and the decentring of the human projects an ecology of "encounters and connections" (Stengers 2008, p.48) that affords better (modest, faithful, partial, situated) ways to engage with the common world. The encounters that happen in sciences between all kinds of entities (*in vivo*, *in vitro* and *in silico*) "can give rise to new modes of relation with humans, i.e. to new political practices" (Paulson 2001, p.112). It is this conviviality of incongruent beings which defies what in Sloterdijk (1999) so doomfully (and controversially) appears as the impotence of humanism – the failure to properly account for whatever and whoever does not fully correspond to the human. Serres makes a similar case when he argues that the entangled, random and unbridled violence that our actions unleash on and through nature (climate change, loss of biodiversity, oil spills, etc.) can only be properly redressed when all beings and objects are endowed with "legal" status through a "natural contract" (Serres 1995b). Bioinformational datascapes and resources and the *in silico* life they harbour further contribute to the menagerie of incongruent entities. Rather than widen the "ontological

gap" (Barad 2007), the sequence universe can offer further models for imagining posthuman politics. In having become a vital site for bioscientific research, the sequence universe too provides for speculative and inventive questions towards cosmopolitical worlds.

Surprises towards a database imaginary

The laboratory has provided a central site for articulating some of science studies' most lasting claims including the work of inscriptions and performative representations, the role of nonhuman materialities and the co-productive exchange between practices and objects. In doing so, it has transcended its (porous) boundaries to become, like the archive and the database, a paradigmatic figure. Like the archive and the database, the laboratory has been explored not just as an ethnographic site but a key epistemic arrangement. These figures seen twice occasion not only double-takes but equally make room for partial visions. EMBL-Bank and GenBank are many things at once – sequence databases, workplaces, institutions, archives, laboratories. In relation to the significance of *site* in science studies, these multiple configurations entail practical and analytical problems. GenBank and EMBL-Bank engage in different spatial formations (or "spatial practices" after Lefebvre) that appropriate and order spaces: They occupy distinct sites (the Wellcome Trust Genome Campus and the NIH campus, respectively), constitute a global space (INSDC), provide sites for experimentation (the *in silico* "discovery environment") and enrol individual laboratories, genome centres and sequencing facilities. While the following chapter will address some of the resulting practical challenges, this chapter has provided some orientations for situating EMBL-Bank and GenBank – not in terms of absolute coordinates but as an assemblage of objects, work practices and texts.

The laboratory has afforded many critical lessons to science studies and beyond. The present chapter has reviewed some of these lessons – concerning work, objects, boundaries and politics – and applied them to bioinformational resources. In doing so, it has revealed some limits of laboratory studies in accounting for entities such as EMBL-Bank and GenBank. Calculation as imagined through the figure of the centre of calculation requires a more context-specific, perhaps more cultural, dimension in order to reflect the

way that scientific curators process data. Similarly, “inscriptions” in the form of the EMBL-Bank and GenBank records should not only be understood as a reduction and abstraction but as a *bioinformational artefact*. Lastly, the proliferation, expanse and centrality of bioinformational infrastructures project a sequence universe that, just like the laboratory, can serve as a realm for posthumanist connections and encounters.

Chapter 3: Meeting the sequence universe

This chapter presents the methodology and methods applied in the present research. It sets out the methodological challenges posed by examining the databases as a *sequence universe* while also explicating the benefits of choosing such an expansive form. The chapter begins by introducing "diffractive methods" – as developed by Haraway (1992) and Barad (2007) – and "inventive problem-making" (Fraser 2006). I illustrate my understanding and use of these methodological strategies through the Agnes Varda's film *The Gleaners and I* (2001) before reviewing some of the key issues that present themselves in ethnographic encounters with information infrastructures. In particular, this chapter discusses the difficulty of achieving the ethnographic prerogative of co-location and the problems presented by multiple sites. I suggest that the notion of co-presence (Beaulieu 2010) and multi-sited enquiry (Marcus 1998; Hine 2007) are better suited for exploring the sequence universe. Elaborating on the concept of diffractive methods, I offer a discussion on the viability of imagination, form and duration. The chapter concludes with details about the "method assemblage" (Law 2004) undertaken in the present research.

Introduction

> [F]eminist objectivity means quite simply *situated knowledges.*
>
> Donna Haraway 1991, p.188

Haraway's laconism betrays a big project, one that remains as topical as when it was first articulated in feminist science studies. How to *account for the world* without reinforcing the very natural and social orders which we seek to question while retaining a sense of the world that remains sufficiently agential to resist our attempts at total capture and interpretation? Feminist science studies proceeded to reveal the social and political biases behind scientific claims. They also crafted methodological and epistemological arrangements with which to build and strengthen alternative claims (Fox Keller 1983;

Fausto-Sterling 1985; S. G. Harding 1986; Haraway 1991).[44] One such device, "situated knowledge", insists that all knowledge claims arise from a particular site and situation that is, in turn, reflected within these claims. This by itself does not necessarily guarantee a responsible and responsive analytical programme: merely privileging localism runs the risk of reproducing the very dichotomies one wishes to abandon. Yet by attuning oneself to the particularities of situations "one may get a sense of what is acceptable, desirable or called for in a particular setting" (Mol 2008, p.9).

Importantly, Haraway's project and that of other feminist (science) scholars such as Mol is not concerned with giving up on the bigger picture, disposing of the bathwater *and* the baby. What it seeks to do instead is vacate the God's eye view and claim new pastures that facilitate "partial visions" and the making of more *relevant* (Fraser 2009) accounts which remain receptive to the affordances, demands and desires mediated by one's environment. Partial vision requires an extension of senses, not relying on a disembodied eye but on a veritable *ecology* of devices and sensors.[45] Situated knowledge and partial vision have readily been incorporated in science studies as they continue to challenge perceptions of truth, universalism and ontological divides through anchoring scientific knowledge claims in concrete contexts.

Cosmic encounters

How to partially see or connect with a universe? In Italo Calvino's story "A Sign in Space" (1968), the protagonist Qfwfq, travelling through outer space, is having troubles placing a marker. Given the infinity of the universe and the fact that it is forever changing, setting a permanent sign becomes an impossibility. He can of course *set* a sign but it will in an instant loose its purpose *as a sign*. Instead it will just be an empty signifier, or perhaps more accurately, a *floating* signifier. This is how Calvino's story is usually read – in relation to an anxiety of semiotics, the worry about the vagaries of meaning and that all our

[44] Feminist critiques of biosciences are too exhaustive to list in detail. An excellent compendium is Keller and Longino (1996). For a more recent review of feminist science studies see (Mayberry et al. 2001) and Subramaniam (Subramaniam 2009).

[45] Senses here are "something accomplished through the competent deployment in a relevant setting of a complex of situated practices" (Goodwin 1994, p.627). A similar argument is made by Michael (2006) who suggests that [p]erception (…) cannot be separated from actions in the world" (ibid. 115).

mental, technological, physical efforts are forever bound to the capriciousness of words. In relation to my research, Calvino's story invites a much blunter analogy: Like Qfwfq, I found myself faced with a universe that appeared formidably amorphous, inhumanely large and multiple. How to find and employ an analytical handle on the INSDC, the world's biggest database? What methods to use? How to approach the sequence universe or in Barad's terms, how to *meet* it?

When faced with the sequence universe, Haraway's instruction for situatedness has to go beyond the recognition of space and place as important variables in the construction and shaping of technologies and knowledge. Recent work in the philosophy of science, such as that of Isabelle Stengers (1997; 2000; 2005; 2010), articulates this "beyond" where *spacetime* refuses the comforts of position: Here, space (and time) are themselves "reified entities that are to be explained by the contingent, changing, but nevertheless concrete elements and events from which they are abstracted" (Fraser 2006, p.130).[46] So while situating our knowledge claims is an important step towards more responsible and ethical objectivity, we have to acknowledge the non-innocence (Haraway 1997) of space and time. If our analytical context is as fickle as our objects of study, how then to grasp either one without robbing the other of its own particularities and interventionist capacities?

Inventive diffractions

In her film *The Gleaners and I* (2000), Agnes Varda follows the forgotten practice of gleaning, the gathering of leftover grain and produce after the harvest. At one point, she films her hand as it rummages through a pile of potatoes, discarded because of their failure to meet the aesthetic requirements of a global food industry that expects its customers to recoil at the sight of anything less than perfect. As one hand holds the

[46] Fraser and others continue the project of "process thinking" that has begun with Whitehead (*Process and Reality*, 1929) and Deleuze (*Difference and Repetition*, 1968). Though they differ in their analytical focus, they share a commitment to the process of becoming ("actual occasion" for Whitehead) as the constitutive substance of the world that continuously and iteratively invents and re-invents matter and relations. The principal occasion for process thinking is the relation: subject and object are constituted in relation to one another. For Whitehead, objects provoke activities and, in turn, respond to any such activity. This thereby becomes an activity undertaken *by the object* (a very similar argument is brought forth by Barad [2007] in relation to the experimental apparatus). Agency here is thoroughly de-centred and non-exclusive.

camera, capturing the image, the other makes its way through the potatoes, picking the ones that are shaped like hearts. On another occasion, Varda forgets to switch off the camera which, slung across her shoulder, idly continues to capture the images and sounds as Varda trudges through the muddy landscape. In tracing and recording the gestures, stories and routes of the gleaners, Varda herself turns into somewhat of a gleaner: pursuing, discovering, collecting images and stories of poverty, waste, French law and survival.

With scenes like these the film portrays situations that are not unfamiliar to the ethnographic researcher: Finding oneself *in the field*, turning one's attention to *wondrous* things, research equipment not quite sticking to the script, balancing one's immersion within the field with the required mediatory and analytical distance, or facing the unexpected. More specifically, Varda's film draws out the uneven topographies of a problem space that retains responsiveness to emergent, not wholly resolved issues (what Callon would call "overflows") but remains committed to a frame – in Varda's case, the very literal frame of the camera.[47] She traces a figuration of gleaning that encompasses European agricultural policy, rotting potatoes, land degradation, digital filmmaking, 19th century painting, poverty, ethnic strife, international trade, high-tech machinery, weather cycles, French nationalism, family histories as well as filmmaking traditions and conventions, her own and that of others.

Varda's film points to two methodological techniques pertinent to the present research. Firstly, Varda presents a series of *disproportioned* and *disproportional* means and methods: Investigating the history of gleaning by letting the camera study a painting for example, or using personal, poetic reflections to grasp the inequities of EU agricultural policy, or, indeed, explicating the absurdity of global food industry through heart-shaped potatoes. This, I would argue, constitutes what Barad, after Haraway, calls "diffractive methods" (Barad 2007). In an interview, Haraway gave the following account of her use of "diffraction":

[47] For Callon, a frame denotes not a context but a pattern that momentarily distinguishes (disentangles) particular actors (1998).

> I am talking about the particular interest and respect I have for well-designed field experiments, in the study of a baboon troupe in a particular ecological setting, for example. That way of knowing intersects with the skills of reading a novel. Those two sets of skills - reading the experiment and the novel - condition the way each gets read so that I can't approach a grant proposal, a scientific paper in primatology (…) without carrying with it the ways that I know how to read a poem, a short story, a novel, a museum display or a painting. (J. W. Schneider 2005, p.149)

If, as Butler (2005) suggests, narrative capacity is a precondition for accounting for oneself and therefore making oneself intelligible to others, then Haraway's capacity for the kind of *generous reading* outlined in the quote above, is equally necessary. Both Haraway and Butler write against universality that refuses to be responsive to more-than-human assemblages and historical condition, respectively. A diffractive methodology reads, applies or interprets observations and perspectives on issues from one frame (for example, history) through other frames (for example, agriculture). It does not content itself with reflexively studying issues but instead seeks to make a difference, which is neither as simple nor as difficult as it sounds. This is a methodology which "is respectful of the entanglement of ideas and other materials" (Barad 2007, p.29). It is a "critical practice of engagement" (Barad 2007, 90) where science is read through fiction, and fiction (and poetry, art, film, etc.) is an integral part of the repertoire of scientific instruments. Particularly, the present research seeks to diffractively read a number of objects that have emerged in interviews and observations through texts and contexts extraneous to the field and discipline. In the context of the present research this means adopting what Russell calls a "documentary gaze" which resides in the real, a particular situation, but also claims stakes in the imagined, or as Russell puts it, in "a narrativity that functions as an eclipse of the real" (1999, p.86).

The second methodological concern posed by Varda's film relates to what Fraser (2006) calls "inventive problem-making", which also goes some way in figuring the kind of *difference* that diffraction makes. Inventive problem-making recasts an issue in such a way as to establish new grounds for enquiry, engagement or address. Hence, Fraser contends, "[t]he best (…) that a solution can do is to develop a problem." (2010, p.78) These two concerns are connected and are, I would argue, required by the new sites that have

emerged in the biosciences. The following section will detail the sequence universe's methodological challenges before describing fieldwork in the universe and the methods used in this research.

Co-presence amidst infrastructural assemblages

Databases do not dwell, they *function* – much like infrastructure. Large-scale scientific data collections such as EMBL-Bank and GenBank bring out many questions related to the situation of science. How to account for the reflexive constitution of scientific practices based in the realm of bioinformation and the notion of particular sites? How to attend to the epistemic culture and sociality (Knorr-Cetina 1999) enacted within the sites of EMBL-Bank and GenBank? How do global bioinformational infrastructures configure the relationship between the laboratory and *the field*? Do they constitute a novel discrete locale? Most pertinent for the present chapter, however, is the question of accessing, literally and figuratively, "sites" such as sequence databases that occupy multiple places, that encompass very different scales and that are not always physical, making a "being in the field" very difficult.

"Where is the database?" I asked many of my respondents at EMBL-Bank and GenBank only to be met by utter perplexity as if I had asked for the location of the internet itself. It was a deliberatively facetious question yet one that points to an important issue. Global science data infrastructures are both endemic and expansive and escape conventional notions of *presence* and materials usually required for situating entities. These are familiar issues in the study of information infrastructures (Star 1999; 2002) where the "first barrier to using fieldwork is seeing infrastructure" (Star 2002, p.108). Science studies have wrestled with less than visible entities such as sick building syndrome (Murphy 2006) or stem cells (Michael et al. 2007) demonstrating that the means of locating/accessing phenomena cannot be divorced from how these phenomena are made to appear. Indeed, one of the primary concerns of these and other studies relates to *making things present* (Law 2004).

As mentioned in chapter 1, this thesis was prompted by Manovich's claim that the database has replaced narrative as the cardinal symbolic form of our times. In thinking

about analogous or comparative spaces for situating EMBL-Bank and GenBank, spaces that make room for stories and materials, the laboratory readily offered itself. And the laboratory, as discussed in the previous chapter, has commonly been studied *ethnographically*.[48] Social scientists went into laboratories, talked to scientists and observed their interactions with other humans and nonhumans. They entered into a sustained engagement *in situ* – spending an uninterrupted amount of time in a locale defined so as "the field" (Atkinson et al. 2010). Observation here is crucial. Lingering too. Most fundamental, however, is *co-location*: The researcher's body has to coincide with, or at least be near, her object of research. Thus, while ethnographic methods allow the rendering of objects and materials that would resist the "database logic", they also pose challenges for the present study, especially concerning site and co-location. These are addressed in the next sections.

Multi-sited co-presence

In the course of my research it became more and more difficult to maintain a narrow definition of either EMBL-Bank or GenBank. They were part not only of the INSDC. Within their respective institutional settings, they formed but one layer of a highly complex, continuously expanding web of bioscientific resources and devices, some produced in-house, others built through outside community efforts. This cascade of resources and tools opened up the possibility of using EMBL-Bank and GenBank as gateways into exploring an extensive realm, a veritable universe of DNA, RNA and protein sequence, genome maps, biological resources, fungi, oceans, environmental samples, mine drainage and controversies.

[48] There are at least two good reasons for using ethnography to study laboratories. Firstly, it provides the appropriate empirical setting for deflation (M. Lynch 2008), or in the words of Latour and Woolgar, avoidance of "epistemological concepts" (198, p.153). Their "examination of the microprocesses of laboratory work" based on "observation of *actual* laboratory practice" is "particularly suited to an analysis of the *intimate details* of scientific activity" (ibid., emphasis added). Daily encounters and routines, gestures, flippant remarks, notes and spreadsheets are crafted into presence while others like cleaners or corporate financing are made absent. Secondly, ethnography offers a convenient historical precedence. Ethnography's origin in anthropology lets laboratory studies in on one of its most potent tricks (though not without chagrin, see Strathern 2004b). Laboratory studies, like anthropology, make us of thick descriptions (Geertz 1973) while exploring purposefully estranged sites. This creates accounts that weave together the familiar and the unfamiliar and that, if done successfully, neither concede (too much) to social constructionist nor defer (too much) to technological determinism.

Here, most clearly, conventional ethnographic co-location was not possible. Aside from being not practicable, conventional notions of ethnographic site can become obstructive in figuring the radical relationalities of artefacts and practices within the sequence universe. Digital environments, infospheres and infrastructural ecologies have prompted critics to ask if what we need is a different conception of site, ethnography and, indeed, locality (Hine 2000; 2007; Mackenzie 2003b). Spending time looking at, reading through and trawling amongst websites, Beaulieu (2010) makes a case for *co-presence* as a valuable mode for ethnographic investigations, a mode that can be enabled by physical co-location but often encompasses other forms of engagements. I too spent long amounts of time *inside* the suite of digital resources provided by the EBI and the NCBI, finding my way around the various levels, interfaces and surfaces and learning how to "read" the database records.

Hine (2007) notes that the multi-sited studies that have emerged in recent STS writing often fail to adhere to the conventions of methodological traditions. According to the anthropologist Deborah Heath, whose work includes studies of the HGP and epigenentics, the study of genomics and its attendant disciplines requires an "agile ethnographic practice" and "a readiness to hyperlink between diverse fieldsites" (quoted in Hine 2007, 662). The research presented in this thesis is multi-sited: I have spent time at EMBL-Bank and GenBank, in their laboratories as well as in canteens, meeting rooms and landscaped gardens. I have also spent time on online discussion forums, in conferences about scientific data management, on blogs and comments sections. But the most time was dedicated to wandering around the digital discovery environments sustained by the databases, exploring database records, following the links contained within them to other resources and moving even further beyond. Like in Hine's study of systematics, "[l]andscapes of interconnected institutions and initiatives emerged on the internet, providing a territory of their own to navigate with ethnographic sensibilities" (Hine 2007, p.666).

Imagining methods

While co-location is one challenge in ethnographic engagement with (digital) information infrastructures, visibility is another: "The labor-intensive and analysis-intensive craft of qualitative research, combined with a historical emphasis on single investigator studies, has never lent itself to ethnography of thousands." (Star 1999, p.383) Not only is there very little to see, but infrastructural things are often boring. My first look into EMBL-Bank and GenBank revealed *offices* and people working on computers, typing, scrolling and moving and clicking their computer mice. Star and her collaborators demonstrated that by focusing on simple tasks, organisational routines, trivial documents and all manners of "articulation work" (Strauss et al. 1985), infrastructure can indeed be *read* and thereby reveal vibrant landscapes. While analysing technologies as texts (Woolgar 1991) has its limits, it does make room for *imagination*. Accordingly, as Mackenzie (2003b) argues, "imagining" plays a crucial role in the meeting of infrastructure and individuals: Engagement with these systems involves activities that purposefully entwine human and nonhuman bodies. Here, "imagining, understood as an experience of other bodies in relation to our own" constitutes an integral aspect for understanding how infrastructures work (2003b, p.367).

By changing the question from "Where do I go?" to "How can I establish co-presence?" (Beaulieu 2010, p.457), the researcher becomes more aware of the devices and set-ups deployed to establish *somewhere to go* in the first place. Whereas some places, such as EMBL-Bank and GenBank offices, readily offered themselves as relevant sites, others, like the mumps genome described in the next chapter, emerged in the course of my speculative exploration of the sequence universe. In this case, "imagining" took a more literal turn. "So what I have to tell in the present book does not just relate to the events that figure in my stories. It also relates to other texts. Lots of them." (Mol 2002, p.2) There are, in the present research, a lot of *other* texts, some explicitly referenced, others only in the form of subconscious traces. Persistent companions throughout my fieldwork with continued appearances in my field diaries and research notes were, among others, Henry David Thoreau, W. H. Auden, Marcel Proust, Thomas Mann, Thomas Pynchon. Countless artists whose work I have looked at in the course of this thesis also found their way into

the text: Nancy Holt, Hanne Darboven, Trisha Brown, Yvonne Rainer, LTTR, Mierle Laderman Ukeles, Sophie Macpherson. They, as much as the interviews, informal conversations, documents, photographs and database records analysed in this thesis, have shaped my research's problem space.

"Fusion", Haraway writes, "is a bad strategy of positioning" (1991, p.192). My thesis comprises stories of other things which might not ostensibly have anything to do with bioinformatics, with nucleotide sequence and databases. But as the literary scholar Barbara Herrnstein Smith contends: "Incommensurability is (…) neither a logically scandalous relation between theories, nor an ontologically immutable relation between isolated systems of though, nor a morally unhappy relation between sets of people, but a contingent experiential relation between historically and institutionally situated conceptual and discursive practices" (B. H. Smith 1997, p.262). In some cases, practices never meet, in others they do but inconsequentially and still sometimes they meet and become mutually transformative. Smith suggests that perhaps come "Judgement day" we won't be able to tell "who finally won" or "even to tell which was which" (B. H. Smith 1997, p.262).

Idiotic pace

Beaulieu's "co-presence" mediates generous relations as it suggests a mutuality that can be quite out-of-bounds, for example, in that establishing co-presence does not necessarily rely on conventional observational capacities. On one hand, this allows for an incongruously populated problem space. On the other, multi-sensory collection and reflection makes for an interesting research persona. The extension of senses returns me to another ethnographic trope, the idiot. Etymologically, idiot refers to a person not versed in the idiom (Latin) and hence not able to communicate properly. This can be seen in the figures of the idiot in literature where the idiot is often described through their awkward and very present physicality. This is, however, often evoked in connection with the idiot's propensity for different and often inventive insights.[49] Here, a privileged observational capacity comes from being out of sorts with the environment: Perhaps the idiot doesn't

[49] See for example Benjamin "Benjy" Compson in William Faulkner's *The Sound and the Fury* (1929) or Ignatius Jacques Reilly in John Kennedy Toole's *Confederacy of Dunces* (1980).

speak the local language, lacks the historical context or cannot partake in the indexicality of local settings. Ethnographic convention very much instrumentalises this notion of idiocy where it not only helps to enchant the familiar but also prevents the ethnographer from "going native" (Ybema & Kamsteeg 2009). In approaching my research sites I knew very little about bioinformatics, DNA sequence, molecular biology or proteins. In my interviews with participants this lack of knowledge was a constant companion. On occasion, it helped prompt participants' self-reflexivity where my "idiotic" questions would make them pause and look at an issue differently.

Isabelle Stengers' cosmopolitical proposal (re)appraises the figure of the idiot as someone who *slows things down* (2005).[50] Stengers' equates this deceleration or pause with an incapacity to proceed along agreed routes and well-rehearsed arguments. More than a refusal to fall in-line with a canon, for Stengers the idiot enacts a particular challenge to the order of things. Here, the idiot introduces the prospect of a totally different order, in other words, the idiot enacts an "interference" (Haraway 1997, p.163). With reference to Melville's *Bartleby*, Stengers suggests that Bartleby's refusal to work, go home or even eat, escapes our comprehension because it fundamentally goes against commonsense. Here, the idiot becomes a foil for thinking things differently, perhaps in the same way as the "fool" of the picaresque novel was used as a rhetorical device with which to give voice to dissent and opposition. Stengers introduces a temporal quality to our thinking, something which ethnographic tradition has enshrined as a key to its success.

In contrast to the brief moment of ethnographic co-location, this thesis has taken me a very long time to research and write. I myself have taken things slow and things have sometimes presented themselves in an untimely fashion. Stengers' idiot allows me to reflect on my pace, its adequacy in relation to institutional conditions but also in relation to my research objects and questions. A commitment to time and duration is a key element of the ethnographic mode. Hess identifies ethnography as a "considerable amount of time" spent in the field (2010, p.238). Durational lingering makes visible routines, renders the

[50] In French and German this connection between idiot and delay is very clear: *retarder* (French) and *retardieren* (German) both mean "to slow things down". In German theatre studies, the *retardierendes Moment* (Gustav Freytag) refers to the moment just before the play's conclusion when alternate endings are given renewed attention – heightening suspense and reminding us that things can still turn out very differently.

everyday tangible and by doing so betrays patterns. It is not the event but the *longue durée* that is the habitat of the ethnographer. I cannot lay claim to a relevant interval in my fieldwork. Whereas effective fieldwork took up (only) about one month, my research, thinking and writing took almost 7 years. Carrying out a doctoral degree part-time is in many ways a durational exercise. *Sticking with it* turns into a mental and physical task and time becomes very visceral as "the PhD" has a very real presence in one's life. The experience was comparable to the many durational artists' films and videos I have watched (or more appropriately "sat through" and "sat out"). There, the duration was actively sought and construed. It was a deliberate entanglement with the viewer. And as a viewer, one goes through motions: At times hypnotised, at others bored and then at others stimulated with thoughts and associations seemingly appearing out of nowhere.

On form

Bringing together data, literature and analysis, my research follows Bowker's instructions that

> a database should be read both discursively and materially; they are a site of political and ethical as well as technical work; and that there can be no a prior attribution of a given question to the technical or the political realms." (Bowker 2006, p.123)

The databases, like alcoholic liver disease, water pumps and atherosclerosis, become different objects in different contexts: They are coherent, invisible infrastructure at one point and messy, lived sites at others. They are multiple or as Law (2004) calls it, "fractional", and there are many strategies by which coherence or "singularity" is achieved. In taking this approach, this thesis assembles many objects, shapes and forms: texts such as scientific papers and textbooks, guides and manuals, websites, field notes and diaries, database records; visual depictions such as photographs, diagrams and website layouts; maps; devices and software; conversations and interviews; architectures and landscapes; and "human apprehensions" (Law 2004, p.146), such as curatorial senses, care and feeling for the organism, frustration over data accuracy and curious probing in digital discovery environments and oceans.

A central aspect of this thesis' *method assemblage* (Law 2004) is the notion of the "sequence universe", which forms yet another figure seen twice: As I style my ethnographic study of the databases in terms of explorations of the sequence universe, the sequence universe, recursively, suggests a specific *form* of study and analysis. Thus, the sequence universe performatively enacts certain "metaphysical commitments" (Verran 2009, p.173). Such commitments can, as Greenhough (2006) has demonstrated, become constraints. Examining the plentiful accounts of the IHSD controversy, she criticises the construal of and subsequent commitments to what she calls the "island-laboratory": Social scientists and anthropologists that had descended upon Iceland to assess the IHSD thereby unhesitatingly reproduced the "relevant" site as suggested by the IHSD venture. This venture had, by seeking to harvest the relatively "clean" gene pool of Iceland's population, turned the island into a laboratory. Such fateful projection found little reflection in critical accounts of the IHSD, which, for the most part, continued to regard Iceland as a "laboratory". This not only set the limits for the IHSD (and its consequences) but for their own analytical contributions.

Another, more visual, take on "form" is deployed in Mol's *The body multiple* (2009) where the literature contiguously runs alongside, or rather below, the main text of her study. Set in a different type and using a smaller font size, her literature forms a concomitant narrative on every page. It establishes a visible dialogic component that performs associations between entities in her ethnographic account and entities in the literature, turning observations into surprises and vice versa. The connections and passages here cut across all kinds of boundaries (body, genre, discipline, expertise, language, chapters). They traverse many spaces (in the hospital but also in medicine and the body) while enfolding many (but not too many) *actants*. Here, form serves as an important vehicle in transporting translations from one site to another.

Form, as the two accounts above demonstrate, can mediate relations in the text and beyond the text, making connections while also rendering limitations. In this respect the sequence universe invokes clumps and clots of matter and matters (Verran 2001), excess and inflation, constellations and orbits, folds and topologies (Murphy 2009), cosmic deflations (Hird 2010), cosmopolitics (Stengers 2010), wormholes (Haraway 1997), black

holes (Michael 2010) and angels (Serres 1995a). The sequence universe here is one of the thesis' key "working imaginaries" (Verran 2009, p.173) and as such points to the material shape of things but also retains a performative possibility (in its verbal manifestation, *to form*). Whereas it is not locatable as such, the sequence universe is sustained by many *in situ* practices: tending to organic matter, running sequencers, creating clone libraries, submitting sequencing data to databases, reviewing and verifying submissions, blasting sequences, building alignments and scaffolds. These make possible connections, or rather "partial connections" (Strathern 1991) and rapid changes of scale – from molecules to the NIH Campus, from Los Alamos to the protein landscapes of Margaret Dayhoff's atlas, from Singapore's *Biopolis* to the "tangled mess" of fungal symbiosis in chapter 7.

The present research

This is a qualitative non-participant research comprising ethnographic explorations of research settings. Primary data for the present research is derived from interviews with staff at EMBL-Bank and GenBank, observations from the field, including field notes as well as from documentary sources and scientific literature, mostly in the area of genomics and bioinformatics. The "field" here includes the buildings and offices occupied by EMBL-Bank and GenBank, their wider institutional settings and, importantly, the world of bioinformational data and tools accessible through their respective portals.

In my search for a suitable database to visit and study, I was immediately drawn to GenBank, often referred to as the biggest database in the world. Only after some initial probing did it emerged that GenBank was in fact one part of a much larger set of databases, the INSDC which includes EMBL-Bank and the DNA Database of Japan. Because of EMBL-Bank's proximity – located on the Wellcome Trust Genome Campus in Cambridgeshire – I approached EMBL-Bank first. Through the EBI website I contacted the head of the EBI division responsible for EMBL-Bank, introducing my research and enquiring about access to study the work of the EMBL-Bank division.[51] My message was passed to a "group leader" (EBI mirrors a standard laboratory organisation in being split into teams headed by group leaders) and occasioned negotiations over purpose and extent

[51] The EBI website is at http://www.ebi.ac.uk/.

of access that lasted 7 months. I appeared to be the first social science researcher to request access to EMBL-Bank although EMBL headquarters in Heidelberg have been running a science and society programme since 1998.[52]

My interest in EMBL-Bank was met by great curiosity from people working there. It seemed no one could quite grasp why a social scientist would be interested in studying their work. My initial desire to carry out "fieldwork" and spend regular amounts of time with EMBL-Bank staff over a longer period of time had to be severely curtailed. I can only speculate as to the reasons for the reluctance that I had encountered during my negotiations for access: EBI and particularly EMBL-Bank were undergoing some organisational changes at that time. Also, renovations were ongoing and the resulting re-arrangements and displacements made for a volatile working situation. Importantly, I came to this task as a novice and most likely failed to instil the kind of confidence required to make institutions feel at ease about letting in a stranger, in this case, a *non-professional* one (Agar 1980). After we had agreed on a series of interviews with members of staff willing to participate, access arrangements were handed to a member of the administrative team who became my guide and point of contact for the duration of my visits to EMBL-Bank.

In requesting access to GenBank I was faced with the inverted problem: Their website did not provide any staff details, neither names nor job descriptions or contact details were listed. Geographically, GenBank did not allow for a committed engagement as I could only take 2 weeks off work and my funds were limited. After having sent an email to a general email account for the NCBI, I received a message from the head of the curation team who became my gatekeeper and guide. Though surprised, the institution showed no unease in welcoming me. Access was negotiated within a couple of weeks and once I arrived, more and more staff agreed to participate. Upon arrival, both institutions had arranged for me to give a presentation to staff. This had been specifically scheduled at EMBL-Bank only for staff who had agreed to participate. At GenBank, my presentation followed a general curation group meeting. I introduced my institution and myself and

[52] This initiative organises seminars, symposia and lectures as well as publications promoting the public understanding of the life sciences. See http://www.embl.de/aboutus/science_society/ for further details.

presented the purpose of my research. A shared sense of curious wonder accompanied both presentations – something which lingered not just through the following interviews and observations but the entire research process. I will return to this sense of wonder in chapter 8.

Observations and interviews

Arriving at EMBL-Bank I was met by the data submissions assistant. She showed me the facilities and grounds and arranged for rooms to carry out interviews in. In contrast, at GenBank all interviews were carried out at the respondents' workstations. Before commencing interviews, I handed the participants consent forms and explained that the research was carried out in accordance with the ethical guidelines set out by the British Sociological Association and the Department of Sociology, Goldsmiths. In discussing the consent form, conversation often turned to the purpose of social research and, in particular, the relationship between social research and science. The consent form was an object familiar to most scientists and as such it offered a useful translational object between disciplines. It also facilitated the transition into the interview. Whereas the interviews carried out at EMBL-Bank were semi-structured, following a number of pre-defined themes, the interviews at GenBank were open-ended and mixed observations with emergent questions. Thus, the research combines qualitative interviews and ethnographic interviews, that is, interviews conducted on-site, and distinguished by a friendly and conversational rapport between research and interviewee (Heyl 2010). I interviewed 6 members of staff at EMBL-Bank, including curators, web developers and an administrator. Conversely, at GenBank I carried out 22 interviews with curators, software developers, dataflow programmers, and taxonomists. This provided for very different materials, alternating between very structured and very open data, and accounts for a much stronger presence of GenBank in the primary data. Retrospectively, the interviews carried out at EMBL-Bank served probing purposes and established more concrete themes for the remainder of the research.

At GenBank I carried out interviews while observing my participants at work. I would meet my respondents at their desk and after an introductory question ("How did

you come to work for GenBank?"), they resumed their task, providing me with an explanatory commentary. As I do not have any formal science training, I acquired a basic knowledge of genetics and bioinformatics so I could follow my respondents' activities, stories and answers. At times, respondents would mention concepts and processes that I was unfamiliar with. I dealt with such moments in two different ways: Where the concepts appeared to offer insights into a realm which I considered relevant, I would ask my participants to explain them further. But more often than not, I would use these occasions to spark off investigative trails during transcription.

The acquisition of basic knowledge of the most important terms and processes around biological data often happened *en passant* in my explorations of the bioinformational resources. In developing the objects that appear in the course of this thesis – viruses (chapter 4), biocuration (chapter 5), metagenomics (chapter 6) and biological annotation (chapter 7) – I consulted considerable amounts of scientific literature, mostly journal articles, gathered through the PubMed gateway and direct sources such as *Science*, *Nature* and the *Public Library of Science*. Both EMBL-Bank and GenBank make manuals and guides available which detail not just instructions on how to use their resources and tools but often also contain descriptions of the underlying science and specifications about the inner mechanisms of certain tools. I also consulted EBI Train online (Beta version) and the NCBI's comprehensive "bookshelf".[53] In addition, the interfaces, especially the portals by which to enter the sequence universe, have, in the course of this research project, become much more user-friendly and visually coherent. As the discovery journeys described in the following chapter make evident, the sequence universe can give rise to some wondrous encounters and discoveries, even for non-scientists.

[53] The EBI training portal, which contains documents, videos and interactive learning, can be found http://www.ebi.ac.uk/training/online/. The NCBI bookshelf, which comprises comprehensive descriptions of all NCBI resources as well as a collection of biomedical textbooks, is at http://www.ncbi.nlm.nih.gov/books.

Chapter 4. Viral and valent trails: a visitor's guide to the sequence universe

This chapter presents a partial topography of the sequence universe, detailing some of its key landmarks, such as the Entrez interface, and sites, such as the Wellcome Trust Genome Campus (WTG campus). To do so, it recounts two kinds of journeys through the sequence universe. Each journey features a distinct traveller: the social scientist (myself) and the biologist (Sandra Porter). My journeys to the WTG campus in Hinxton, UK and the National Institutes of Health (NIH) in Bethesda, Maryland are relayed through field notes. The second traveller, the biologist Sandra Porter, is on a discovery path through the sequence universe, a journey which she has documented on her blog *Discovering Biology in a Digital World*, published on *ScienceBlogs*, an invitation-only network of science blogs. All quotes are taken from Porter's 5-part series "Do mosquitoes get mumps?" published between 21 and 25 September 2008.[54] Other voyagers will make occasional appearances, such as researchers from the Institute of Nephrology at Peking University and an as yet unnamed paramyxovirus.[55] The figure of the *doubtful guest*, named after the odd visitor in Edward Gorey's illustrated book of the same name, *The Doubtful Guest* (1957), shall serve as a common guide through these varied travels, functioning both as a literary device for exploring narrative, virtual and ethnographic tropes and as a metaphor in analysing and interpreting these.

Introduction

EMBL-Bank is located within the European Bioinformatics Institute (EBI) which occupies two interconnected buildings on the Wellcome Trust Genome Campus (WTG campus) in Hinxton, near Cambridge. The campus is set amidst a historical estate dating back to the

[54] See http:// scienceblogs.com/digitalbio. Last accessed: 20 March 2012.
[55] An RNA virus causing acute respiratory diseases transmitted in an airborne manner. Among the agents of paramyxoviruses are mumps, measles, Newcastle disease, parainfluenza, Sendai virus and Hendra virus (Shiel 2008). Unlike other viruses, it replicates in the cytoplasm and not in the nucleus of the host cell.

early 16th century, encompassing 95 acres of parkland, situated on the banks of the River Cam. Aside from the EBI, the campus is home to the original Hinxton Hall, now the Wellcome Trust Conference Centre, as well as the Sanger Institute and its suite of attendant structures (such as the Data Centre, the Cairn Pavilion, and mouse facilities). Unless you have a car or are eligible to use the shuttle bus from Cambridge, getting to the campus can be somewhat difficult. Having discerned that the closest railway station was not in walking distance from the campus, I decided to bring my bicycle with me on the train from Liverpool Street Station. Aside from putting my destination in easy reach, it gave me an opportunity to survey the Cambridgeshire landscape that provides the setting for the WTG campus. Cycling there, I thought, would also heighten the sense of incongruity that has been haunting the idea of an ethnography of the sequence universe.[56] Surely, *visiting* a database is strange enough, doing so on a bicycle would put me firmly in the realm of the absurd?[57] For two weeks in the spring of 2008, I would take a train from Liverpool Street Station to Great Chesterford, a station of which even the campus' website warned: "Please note that Whittlesford and Great Chesterford stations do not have a taxi rank or telephone." From there, I cycled for 20 minutes – through Great Chesterford, passing Ickelton, towards Hinxton. I travelled along picturesque country lanes, lined by quaint houses. I occasionally crossed paths with sheep and 4x4s. Rabbits, however, where everywhere. Glancing into gardens, fields and even the ditches that framed the roads revealed busy masses of them.

This brief scene foreshadows some of the themes which shall be discussed in this chapter: the "surrounding" environment, multi-sited proliferation, strange scales, and perception-as-movement. In describing my travels to the WTG campus, I attend to some of the ostensibly *spatial* challenges posed by my object of research, nucleotide sequence databases, and, more generally, the sequence universe. Yet, the spatial cannot be articulated as a discrete or sensible domain in the journeys recounted in this chapter. In

[56] Incidentally, recent literature (Spinney 2006; Spinney 2009) has suggested cycling as a valiant ethnographic method of research. It can afford a *mobile* sensing of environments that captures an experience of landscape more in keeping with its laboured, temporal composition ((Ingold 2000).
[57] In fact, the bicycle turned out to be quite an apt means of transport given the history of the Hinxton estate. Tube Investments, the previous occupants of the campus, owned the British Cycle Corporation which looked after the production of, among others, Armstrong, Brampton and Raleigh bicycles. Hence, arriving on bicycle could be seen as an incidental homage to the campus' history.

particular, the notion of scale, usually an ordering moment in reference to space, emerges as a challenge that does little to assuage the undoing of spatial logic. Whereas the molecularisation of biology (discussed in chapter 1) has often been associated with ever diminishing scale, the informationalisation of biology is easily characterised in terms of a multiplication of scales. "Biologists", a recent article in *BMC Bioinformatics* informs us,

> must constantly traverse across micro-, meso-, and macro-levels of biological knowledge to gain insight into the workings of the cell. Moreover, our current understanding of cellular phenomena is also highly multi-layered, organised as assemblages of several -omic spaces such as the genome, transcriptome, proteome, metabolome, and biochemical pathways. (Arakawa et al. 2009)

Here, scale is neither an elective frame imposed by the enquirer, nor a consensual, if temporary, common ground. Instead, it becomes part of that which one seeks to know in the first place. The "-omic spaces" too are not plains to be discovered or containers for analysis but are actualised differently in relation to interventions of all sorts, molecular, medical, technological, biological, literary and otherwise. The topography of the sequence universe assembled on the scale of the human social scientist involves, amongst other things, buildings, travel arrangements, institutional encounters, participants, ethics forms, digital recorders. The same topography brought into view on the scale of a viral sequence engages the Center for Disease Control and Prevention, a sick goose from the Guandong Province, China, casualties (human and avian) in Hong Kong, haemagglutinin and the National Institute for Medical Research. Yet, a viral RNA sequence, consisting of a simple string of letters representing nucleic acids, operates on a molecular scale that might be considered "smaller" than the human scale of the social science researcher.

For social sciences, multi-scalar, peripatetic phenomena are posing similar challenges (Woolgar et al. 2009) to the extent that "what is blocking the whole interpretation of the social is the macro and micro distinction" (Gane & B. Latour 2004, p.84). Here, scale or rather, the commitment to a binary conception necessitating foregone conclusions (e.g. macro-big-general-global vs. micro-small-particular-local), often hinders conclusive accounting for the multi-scalar traversals and encounters that make and mark our research and the social. One of the first things I was told by my respondents was that

"[y]ou really can't talk about GenBank out of context. GenBank is the archive – the importance of GenBank has to do with how it's integrated with feeding these other things." (GB11) Hence, while GenBank might be the largest database in the world, it is just one part of a manifold assemblage.

When faced with implacable binaries, the strategy of *symmetry*, taken from the sociology of scientific knowledge (Pinch & Bijker 1984), has proven a potent perspective, and this chapter too seeks to call upon this strength by treating scales alike. Thus, in the following, Porter's explorations of (and *on*) the mumps genome are recounted alongside, and on equal terms to, my ethnographic travels to EMBL-Bank in Hinxton and GenBank in the US. The constituents that at once carry and populate the journeys vary in degree and kind. Here, the determinant features of Euclidean space, including location and scale, do not suffice in cohering travels, spaces and encounters. Instead, narrating the journeys the chapter assembles ways for figuring some of the *topological concerns* suggested by the sequence universe.

The doubtful guest in the sequence universe

Sandra Porter's discovery journey begins on a viral genome (mumps) and details its proteins and their emergences elsewhere – first in what appears to be a human cell, then in a rat before settling in a mosquito. Parallel to Porter's quest, which wants to bring to light how exactly the virus came to reside in a mosquito, this chapter recounts my own ethnographic journeys into and within the sequence universe. In recounting Porter's journey, I inhabit her story as I follow the trails documented by her on her blog. Porter, and by extension I, are attempting to reconstruct how a mumps protein got itself mixed up with a mosquito. In the course of this reconstruction, other travellers and voyages are rendered present. Appearances unfold along paths and thereby unravel a series of varied places, like genomes or the NIH Gateway Center. The transient gaze assembled in these accounts has no unequivocal mastery over and comprehension of the phenomenon it seeks to describe, the sequence universe. Ethnographic convention poses the traveller as a privileged cognitive figure, wandering through the field and engaging in conversations which are brought back home and subsequently mined for insights (Kvale 2007). This

chapter features different travellers, Porter and I but also more-than-humans such as viruses and mosquitoes. Here, symmetry exceeds commitment and becomes narrative method in a series of tableaux that present travels which move in and out and through very different spaces. The sequence universe is thereby chartered as a realm for both voyagers and guests, *doubtful* ones at that.

In Gorey's illustrated poem, *The doubtful guest* (1957), the eponymous guest first appears at the top of an urn, part of a neo-classicist balustrade belonging to a stately manor. Wrapped in a striped scarf and wearing lace-up shoes, it gazes at the manor's residents who have come out to answer the bell. The doubtful guest does not belong to any known species: Its arms resemble the wings of a penguin, it has a beak yet it appears to be covered in fur. Throughout the poem, the doubtful guest comes into view in strange locations and postures – "with its nose to the wall", "inside a tureen", eating a plate at the breakfast table or standing in the fireplace – while causing havoc, inconvenience and general bewilderment to residents and readers alike. No one knows what it is, where it came from or why it is there: "It came seventeen years ago – and to this day, It has shown no intention of going away".

The guest denotes dwelling, albeit of a precarious kind, though her appearance as a guest would suggest travels of some kind. Bound by conventions that are not of her making, the guest continuously negotiates, in gestures, comportment and words, her presence and, where decorum dictates, her absence. Indeed, a good guest's presence should never impose. Gorey's doubtful guest is doubtful in multiple ways as its presence oscillates between welcome and invasion: One cannot be certain if it is indeed a guest, appearing uninvited with none of the usual deference extended by those seeking hospitality. At the same time, its odd but quietly brazen demeanour implies that its disposition might be primarily inquisitive, studying its hosts and habitat with insistent puzzlement. One of the privileges enjoyed by the guest, doubtful or not, is that she is not just physically "let in" but "let in on": conversations, private rituals and intimate spheres can reveal themselves to the guest, often inadvertently.[58] It allows the reader together

[58] This has made the guest such a rich figure in literature – think of Hans Castorp's extended visitation of the Berghof sanatorium or Proust's receptions in Parisian salons.

with the protagonists to gradually fathom the more complex patterns – cultural conventions, socio-economic standings, personal histories and political beliefs – governing the world they have entered. In that sense the figure of the guest can become a scalar device.

The doubtfulness of Gorey's guest also pertains to its obscure nature, recognisable as neither human nor nonhuman, and, by extension, to the nature of the environments it visits. Are they in fact of this world? In this chapter, the travellers are also doubtful guests, providing sufficiently entangled perspectives for travelling through the sequence universe and making connections that confound outsides and insides, local sites and global processes, specific actors and the practices they constitute *and* trouble. In doing so, this chapter adds topological concerns to the notion of site discussed in chapter 2 while also anticipating matters, such as presences, absences, vision and the agonistic-affective spaces of controversy, that will appear in the following chapters.

Into the Wellcome Trust Genome Campus

Passed Ickleton, the road winds to the right in a gentle uphill curve. As I reach the top of the slope, I can see some low-rise modern buildings in the distance off to my right. Though they represent a marked departure from the cottages and churches dotting the path through Saffron Waldon, they nevertheless integrate into the landscape. Even the chimneys of what later turns out to be the “mouse facility” have to vie with treetops for airspace. There are no signposts to the WTG campus, which sits inconspicuously amidst rolling hills and picturesque clusters of country life. It occupies its site very discreetly. The buildings quickly disappear from view as I make my way down Ickleton Road towards Hinxton. After clearing a level crossing, I see that, to my right, a brook is now accompanying my journey. As I dismount to inspect this further I realise that I must have reached the south-west edge of the campus. There is a low stonewall just beyond the brook and more rabbits cavorting in the sunlight. I continue along the road and come to a small intersection. To my left is a road leading into Hinxton, to my right the brook has given way to a sizeable wall. There is a small gate for pedestrians, ostensibly leading into the campus. I see some people emerge in groups from the campus heading down the road

into Hinxton. It is lunchtime. Though the gate and its surrounding area look as if not much has changed since the original estate, the ID cards prominently displayed on peoples' clothing betray a state-of-the-art security system. I certainly was not able to enter the campus from the side gate and had to continue along the road, for another 150 meters until it ended in a single carriageway. I turned right, following alongside the campus' wall for another 300 meters when a roundabout and toll-gated driveway marked the entrance to the campus. There was nothing "to see", no spectacular sculpture, arch or other ostentatious feature that would let the visitor know that she was about to enter the place from where the human genome had been decoded.

The security guard manning the tollbooth directs me to the visitor centre, which was set to the left of the driveway and fronted by a small car park. The centre, where I pick up my visitor's pass, takes the shape of a freestanding shed, front side all glass, tilting skyward. This is more congruent with the architecture I had expected to find and, as it turns out, the visitor centre does herald the kind of architecture to come. The further I make my way into the campus, the more the mundanity of the entrance area gives way to an arrangement of modern and idiosyncratic buildings.

Landscape with database

The campus is a product of three rounds of major developments (completed in 1998, 2005 and 2007, respectively). A group of portakabins, set to the side of the EBI building, suggests ongoing works. After the site's acquisition by the Wellcome Trust in 1993, Robert Myers Associates oversaw the transformation of Tube Investments' metallurgy labs (see footnote 57), the restoration of the gardens and parkland. The architects were keen to create "a new landscape setting for the laboratory buildings, whilst restoring the fabric of the 18th and 19th century landscape".[59] They explicitly sought to minimise the buildings' impact on the parkland by configuring location, views and buildings so as to create a coherent architectural and environmental unity. This original vision of minimal impact certainly appears intact to the first-time visitor. My first impression of the campus – from the slopes of Ickleton Road affording an expansive panorama – remained true despite my

[59] See http://www.robertmyers-associates.co.uk/projects/Hinxton.htm. Last accessed: 20 March 2011.

current proximity: The weight of the buildings dissipates into the landscape and the juxtaposition between landscape and research facilities is seamless – even the car park, sunk into a mount, appears strangely wholesome. The vertiginous vanishing points of the financial district that has been my departure point (Liverpool Street Station) have settled on a more human scale, allowing a manageable and horizontal vista.

I leave my bicycle in the designated bike section of the underground car park. As I make my way out of the car park and come around the building, I can see that the front, which is slightly elevated, is entirely glassed and opens to a large terrace overlooking the Sanger Institute. Opposite the Sanger Institute ("the Sang" as people call it), stands the European Bioinformatics Institute (EBI) which houses EMBL-Bank. The EBI, like the Sanger Institute, was built in the course of the second re-development, completed in 2005, and led by NBBJ whose portfolio includes headquarters for Amazon, Starbucks and Reebok as well as The Bill and Melinda Gates Foundation and numerous healthcare facilities around the globe. It comprises space for 2,000 staff, further laboratories, the Data Centre (called the "ice cube") with its floating meeting room and supercomputing facilities, research support facilities and the Cairns Pavilion, a multipurpose building for dining, meeting and sports which also houses the car park I had emerged from.[60] From the EBI reception, I proceed through the building and cross the threshold into the East Wing. This was a result of the third round of development, prompted by the success of the Human Genome Project, concluded in 2007 and once again designed by NBBJ. The East Wing, connected to the main building through two walkways (one on the ground floor and another on the first floor), houses training facilities, a communal kitchen and dining area as well as additional EBI initiatives focusing on human variation, chemoinformatics, and a more consolidated and extensive effort to integrate scientific literature.

Like the second campus re-development, the East Wing was not just an expansion in terms of space but also staffing, from 70 when it first opened to 350, and scientific remit (Wiegler 2007). An EMBL-EBI press release states how the integration of architecture and landscape was meant to translate into the research activities carried out amidst its spaces:

[60] The expansion has attracted accolades from the architecture and design community, the research community (*R&D Magazine*'s "Lab of the Year" in 2006) as well as from environment assessments.

> Continuity in design across campus reflects the scientific complementarity of the campus members and adheres to the principles of sustainability, innovation and connection between the people who work there, the environment and the surroundings. (EMBL-EBI 2007)

The architectural developments and the considerable investment they necessitated bear testimony not only to the advancing of bioinformatics but also to the increasing physical and conceptual entanglement between the generation of bioinformational data, its management, curation, distribution and translation into downstream products and resources, and laboratory benchwork. The continuity between research carried out at the Sanger Institute and the bioinformational efforts of the EBI finds expression not only in their architectural similarity and spatial proximity. Visiting the Sanger Institute's website (http://www. sanger.ac.uk), the distinction between products of dry and wet laboratories becomes difficult to discern. Many of the resources I had initially encountered during my research on the EBI here appear in the domain of the Sanger Institute: the Ensemble browser, which generates and maintains annotation on selected eukaryotic genomes; Wormbase, the database on the biology and genome of *C. elegans*; TreeFam, which provides phylogenetic trees of animal genes; Artemis, a genome viewer (see below); and the sequence analysis tool BLAST.

Performative integration

Sitting on the terrace of the Cairns Pavilion, I overlook a sequence of landscaped spaces proceeding from the parklands to the central plaza. This, as one architectural review remarked, was designed "in the tradition of a European market square" (Walz 2006). The analogy is apt in more than one way. While people congregate in the formal and informal meeting places around the square, the Sanger Institute also claims an imaginary space, replacing the transcendental *axis mundis* suggested by churches and cathedrals with a double helix which is arguably no less transcendental (Fox Keller 2000).

Between the Sanger Institute and the EBI sits a carved tree trunk (*Oak Spiral* by Richard Bray, 2000). Laying on its side and skewered unto three vertical steel bolts, the trunk is carved into the shape of a screw. The torsions that run along the entire length of the trunk gradually narrow. While it is perhaps more immediately reminiscent of a drill, it

is supposed to resemble a helical structure. The trunk's grooves bear carvings of the letters A, C, G, T as well as names of species (such as "Arabidopsis") whose genome had been decoded by 2000. The sculpture seemed odd in relation to its surroundings. Its material and technique, wood and carving, were very distinct from the technological aesthetic of the surrounding architecture and, more generally, the highly technologised science happening on campus. Its drill shape pointed towards a functional value, a device to bore into and probe bodies. Listing "decoded species" serves to commemorate a milestone but also retains a strange sense of memorial. And, lest we forget, it was a dead tree. Coincidentally, the EBI's logo consists of a grand oak tree, which also features as a large mural-like display inside the EBI's reception. It depicts the actual oak tree that can be found growing next to the institute's entrance (passed the tree sculpture). Further inside the EBI, in the East Wing, a contemporary art installation discretely set within the ground floor area seems to act as a prescriptive reminder of the performances the building is meant to encourage. Delicate plates, ingeniously light to cast shadows onto a grid of small projection planes, bear words such as "network", "exchange", "function" and "standardize".

Later on that day, my guide shows me around the parkland. As we tread on serpentine paths, she points to different buildings, telling me about the research carried out inside of them. In the same breath, she comments on the features in the garden and park. On the south-west end of the campus, she leads me into a walled-in orchard populated by century-old apple trees. In contrast to the manicured integration of buildings and landscape, these appear deliberately freakish, more uncontrolled yet contained like the managed ruins of some neo-romantic theme park. The integration with the landscape, the "minimum impact", is also indicative of the campus' environmental charter which assumes a prominent presence on the campus' website. There, mention is made of the campus' "other" inhabitants, like a flock of Canada geese. What about the other "other" inhabitants, like the mice in the mouse facility? There is no public information available on the animal facilities. In response to my question about the nature of the building with three chimneys, one scientist replies: "Animal building, mice, rats and frogs. No primates." (EB 6) But it makes me wonder whose impact is being minimised?

For Galison and Thompson (1999), the magnificent post-War laboratory buildings such as Louis I. Kahn's Salk Institute (1966) illustrate a shift in the doing of science from small group-based work to a style more akin to factory work characterised by hundreds of staff, highly centralised authority and industrial methods. Thrift (2006) too sees space and its architectural manifestations as indicative and in the service of the dominant mode of production. Here, the built environment is both an expression and an operator of the new genomic governance which for Thrift is primarily characterised by an avalanche of information that demands to be processed in ever more viable ways. The bioscience building here becomes a "spatial prototype" for a "space of invention (Thrift 2006, p.292).[61] These buildings are *performative* in that they "are clearly meant to manipulate time and space in order to produce intensified social interaction so that all manner of crossovers of ideas can be achieved" (ibid.).

The buildings that converge on the central square certainly convey a particular image of bioscience but instead of ostentatious connections, the buildings and the careful massing appear to be signalling integration: green sedum roofs, timber brise-soleil, composting of disposable vending cups, an orchard and the Hinxton Wetlands Nature Reserve are just some of the ways in which the campus integrates the demands of data-driven bioscience with Cambridgeshire flora and fauna. Instead of an imposing physical presence, the buildings of the WTG campus, though certainly distinct, seep into the surrounding landscape. This is no *Biopolis*, the purpose-built science quarter in Singapore, described by Waldby (Waldby 2009) as a high-tech version of Piranesi's *Carceri* complete with suspended walkways connecting gleaming steel-and-glass constructions rising into the Southeast Asian skies. The buildings converging on the central plaza of the WTG campus feel proportioned, somewhat diffident even. If, as Waldby and others (Thrift 2006) have suggested, the architecture of places like *Biopolis* betrays some of the aspirations instructing the ventures they give room to, then the WTG campus casts a curious ambience, both highly technologised *and* pastoral, spectacular *and* inconspicuous.

[61] Thrift lists the Centre for Life (2000) at the University of Newcastle and the Wellcome Trust Biocentre (1997) and the Centre for Interdisciplinary Research in the Life Sciences (2006), both at the University of Dundee.

Entrez: Playground with mumps

A manageable vista also welcomes Sandra Porter, a "digital biologist, teacher and entrepreneur", upon her arrival in the sequence universe though her route differs considerably from mine. Porter authors a blog entitled "Discovering Biology in a Digital World" that documents developments and experiences in the expanding realms of digital biology. In September 2008 she began a series of blog posts that recounted a research discovery, a possible new paramyxovirus (see footnote 56) in the yellow fever mosquito. This series of posts can be read as both journal and lab book in that it narrates a journey through the sequence universe as well as details her (digital) methods. Like my journey, hers had an exploratory purpose. Initially intending to write about immunology and vaccines, Porter instead "decided to play a bit with the sequences in the mumps genome":

> I did what I usually do when I want to learn about a new virus. I went to the NCBI and searched for the mumps virus genome among the 1600 or so eukaryotic viral genomes that have been sequenced.

The entry point for her journey, which also provides for the pleasantly manageable vista, is the Entrez interface. The simple layout, a centrally aligned search box on top of a table whose 3 rows and 2 columns list the accessible resources, conceals the world's largest and most complex collection of biological data. Entrez is an integrated database "retrieval system" (Benson et al. 2010, p.D36) and a suite of resources that provides access to a set of 40 molecular and literature databases containing over 350 million records (Sayers et al. 2010). Put differently, Entrez assembles genetic data for 290,935 archaea, 5,659,573 bacteria, 3,225,572 eukaryotes and 110,055 viruses.[62] It is accessible to anyone through a web browser and brings together, among other things, nucleotide and protein sequence databases (GenBank and Protein), taxonomy, gene expression, conserved protein domain (CDD), biological material description (BioSample) and short genetic variations (SNP) databases and the publications database PubMed. In short, Entrez constitutes a central entry point, a nave into the sequence universe albeit one reproducible on screens across the world. As a cosmic axis, it traverses multiple spheres and by doing so allows the user

[62] There are also 2 viroids and 1,099 "other sequences" (all figures as at 22 April 2012).

to *make relationships* through the integration of data and resources as well as the use of hyperlinks.[63]

The ability to occasion "unanticipated" relationships is a particularly noted feature in the documentation provided on the NCBI web portal. And it is this prowess which turns Entrez into "an engine for scientific discovery". NCBI continues "to add new discovery components that assist researchers in finding particular Entrez links and using them to discover interesting relationships *within* the NCBI databases" (Sayers et al. 2010, p.D40). While it provides Porter access to the place where the mumps genome (and others) resides, Entrez is an active environment in which the coming together of data from different corners of the sequence universe can harbour unknown or yet to be discovered elements, surprises even. A simple search for mumps reveals, among other things, that bats are hosts to major mammalian paramyxoviruses (in Drexler et al. 2012) via PubMed); the J. Craig Venter Institute is working on a project to map the genomic characteristics of viral agents that are controlled by vaccination ("Sequencing of Vaccine Preventable Disease Agents" in BioProjects); that a lumphocyte cell sample from a Caucasian ADA (adenosine deaminase deficiency) patient shows a delayed hypersensitivity to mumps (via BioSample); that mumps virus infection (or more specifically the expression of mumps virus V protein which we shall get to know further down) can induce the degradation of the STAT3 protein, a transcription factor known to be overactive in certain cancers (Ulane et al. 2003) via *Gene* database); and that Andrew Wakefield, of the MMR vaccine controversy (for an overview, see Casiday 2007), still has a publishing career (*Waging war on the autistic child: the Arizona 5 and the legacy of Baron Munchausen*, Skyhorse Publications, New York, 2012, via the NLM catalogue). Entrez is where Porter, not unreasonably, goes "to play a bit". There, she locates and visits the single-stranded mumps genome in the Genome database, the NCBI database resource providing genome information including sequences, maps, chromosomes, assemblies, and annotations.[64]

[63] Relationships are established through computationally derived associations within a database and relationships based on information present in the records themselves. Related sequences can be identified through similarity searches with BLAST (see chapter 1) whereas related structures can be ascertained via the Vector Alignment Tool (VAST). PubMed citations are produced on the basis of algorithms that compare words and phrases in the abstract of scientific papers.

[64] A genome is sequenced in bits and then re-assembled to create the full contiguous sequence. Each piece of sequenced DNA is called a "sequencing read" or "read" (these are stored in the Trace

Travels to the NIH

Like Porter, I too "went to the NCBI" though I took the metro instead, disembarking each morning at Medical Center station serving the NIH campus, the world's largest medical research facility. An escalator ride and a few steps take me to the Gateway Center, the public access point to the NIH campus. Despite our different means of travel, Porter and I are both in explorative spirits as we pass our respective gateways, Porter the Entrez gateway and I the NIH's Gateway Center. Yet my own travels eschew much of the ludic disposition that characterise Porter's discovery journey. While we share a sense of exploration, I am more aware of my status as a guest, casting critical and self-reflective glances throughout my travels. To begin with, my way of accessing both EMBL-Bank and GenBank was less carefree and required in parts extensive prior negotiation. Going to the NCBI in my case entailed communicating with senior staff at NCBI, making travel and accommodation arrangements and taking an 8-hour flight from London to Washington, DC. From there, it was another 20 minutes on the metro to Bethesda, Maryland where the NIH campus is situated. Similar to the Entrez interface, which offers an assembly of different sites to search and visit, the NIH campus is a sprawling sphere of institutions and buildings. Unlike the discrete embeddedness of the Wellcome Trust Genome Campus, sunk into its Cambridgeshire landscape, the NIH campus makes no secret of its very large and imposing presence. Upon entering the Gateway Center, I am met by security arrangements: Two metal detectors, flanked by x-ray machines and security staff greet the visitors. I am told to put my bag and coat in a tray to be x-rayed, which I do before walking through a metal detector. Once cleared by the machines, I am allowed to proceed to one of the "check-in desks". There, a woman behind what appears to be security glass confirms my identity and purpose via my passport and a cross-check with GenBank's visitor register. I am issued with an ID card, valid for the day, and proceed up a flight of stairs into a non-descript corridor leading to a glass door. This finally releases me unto the NIH campus.

Archive for Sanger-based sequencing and in the Short Read Archives for sequence reads from next-generation sequencers). In order to reconstruct the entire sequence of larger molecules, several thousand (in some cases million) reads need to be produced. More on this process can be found in chapter 5.

Ways into the sequence universe

I found my way from Great Chesterton to the WTG campus with the help of a printed out map of the Hinxton area generated by Google Maps. Incidentally, searching Google Maps for "Wellcome Trust Genome Campus" would have erroneously led me to Little Chesterford. There is no (public) map of the campus – the website does list the main institutes and organisations that reside within the campus but does not provide any indication as to where these are located. In contrast, the NIH campus is covered by a multitude of maps: There are differently annotated visitor maps for the campus and for individual buildings on the campus, parking and shuttle bus maps, a Gateway Center map and Employee Access maps. This serves as a reminder that, unlike the NIH campus, the WTG campus is not meant to be accessed by the public. My journeys to and from the campus as well as the protracted negotiations which I had entered into in order to gain access to EMBL-Bank's facilities and staff in the first place, all stressed this point of no public access. In contrast to the round-the-clock virtual access to the databases, physical access proved difficult. While this draws attention to the fact that, although ostensibly a distributed and virtual resource, the databases still occupy a particular site, it also suggests that access very much depends on the means of access and the reasons for access and, importantly, on who or what is seeking access. Viral sequences, for example, certainly have an easier time getting there than social scientists.

Once inside the WTG campus, I am entrusted to EMBL-Bank's data submissions support assistant, a trained librarian, who shows me around the campus and, for the

Figure 1: Booklet made by the data submissions support assistant at EMBL-Bank

duration of my visit, acts as my guide. But it's not just me who she leads into the EBI and, subsequently, the offices (or "labs" as they are called) of EMBL-Bank. Her job is to guide sequences into EMBL-Bank, or, as she puts it, "moving the data to people who are appropriate to deal with it" (EB 4), also guides sequences into EMBL-Bank. She is the entry point to the sequence universe – for both sequence and social scientist: "I am the gateway. Which is amazing if you think I have no biological knowledge whatsoever but I don't need to."[65] Single-stranded RNA viruses and other more-than-humans have to pass through her before they can find a place in the database. How, I ask her, does she accomplish such a feat? She picks up an unassuming self-made booklet (Figure 1). Each page contains the name of a data curator, in no particular order.[66] This booklet, she explains to me, determines the distribution of nucleotide sequence submissions received. I sit next to her as she looks at the direct sequence submissions received that day on her computer screen.[67] She begins by allocating the first submission that appears in her submissions' queue to the curator named on the first page of her booklet. The next submission is forwarded to the curator named on the second page and so on. She allocates a maximum of 20 submissions per curator per day.

[65] GenBank deploys a different approach to handling incoming submissions. This is described in chapter 5.
[66] Data curators are the scientists employed at EMBL-Bank and GenBank who verify and manage the incoming data. The next chapter (chapter 5) details their work practices.
[67] Once again, details about computer programmes used to manage sequence data are provided in the next chapter.

Her work also entails tacitly monitoring curators' workloads and distributing submissions accordingly. To her, allocating submissions by assigning them to curators in the order of their appearance in the booklet ensures an even and fair distribution of tasks. The informal exchanges that occur through the shared facilities on campus as well as through activities such as the 3-o'clock coffee meeting, held every day on the landing above the EBI reception area, give her an opportunity to keep up-to-date on any personal or professional responsibilities that may affect the capacity or availability of individual curators. This information is important to her as she uses it in tandem with the booklet to ensure the appropriate delegation of work for each curator. Hence, the booklet enables the distribution of sequence data to take into consideration tacit personal developments as the support assistant bases decisions on her empathic and compassionate assessment of curators' dispositions (sometimes ignoring the order stipulated by the booklet) and not on their scientific expertise or any other more apparent principle such as the nature of the incoming nucleotide sequence (something she would not be able to discern given her lack of biological knowledge).

In the offices at GenBank, another paper-based object catches my attention. Almost every curator there has a diagram of the "coding table" on their desk (Figure 2) or affixed to their cubicle wall. Similar to the data support assistant's booklet, the printout assists them in ensuring appropriate matches, this time between amino acids and codons.[68] This A4 printout features a series of concentric circles, with the inner most circle divided into 4 slices of equal proportion featuring the letters A, C, U and G, representing the four nucleotides (in RNA the DNA's T is replaced by U). The subsequent circle depicts slices containing all 4 letters in each segment while the next circle is again sub-divided into 4 segments containing all 4 letters. The outermost circle features the translations of the codons. Therefore, codons are read from the centre along the circles' radius in outward direction. Prompted by my question about what that diagram represents, a curator describes, somewhat offhandedly, how it works:

[68] Scientists "read" the genetic code in triplets of nucleotides called "codons": AUG is "read" as Methionine, UGG is "read" as Tryptophan and so on (there are currently 64 combinations in total).

You got the four nucleotides and 3 nucleotides make up what's called a codon. And a codon encodes one amino acid. So in a first position if you have an A followed by a G in the second position and a G in the third position, AGG, that would encode an Argenin amino acid. Whereas if you had CAG that would encode Glutamin. GB8

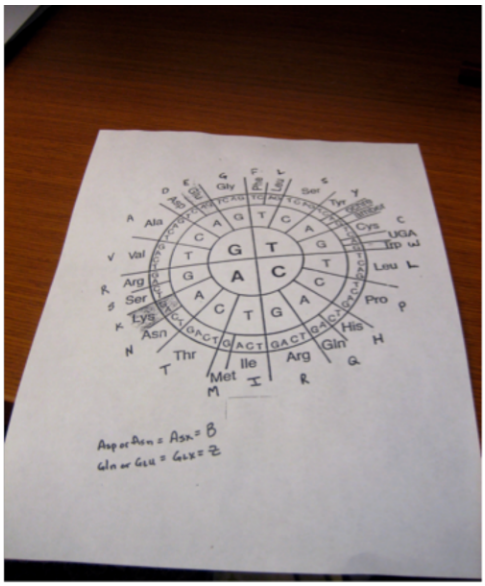

Figure 2: Coding table used by curator in GenBank office

This coding table, as he points out, is "just a little way… I sometimes use this in my work as a reference" (GB8). For the curation efforts then, this diagram serves as a guide for reading the codons from the sequence. Accentuated by the curator's laconic explanation, it represents a very straightforward process of following a predetermined, unilateral pathway commencing at a central departure point and moving in a straight line towards one of a possible 64 conclusions.

Unlike the diagram of the coding table, the booklet is an odd object to find in the sequence universe. On one hand this is due to an apparent incommensurability of scale: Here we have one of the largest sequence data resources in the world, yet data are routed into the database by means of an object that lacks any technological or scientific sophistication. It is made out of paper (there is not much paper around in the offices of the EBI), it can only be made sense of by one person and it uses *ad hoc* orderings whose logic lie beyond the grasp of scientific method and reproducibility. Yet, it *works*. Rather than a disruption or disparity, the booklet and its attendant entanglement of non-expert, affective knowledge would suggest that the sequence universe does not necessarily heed to one spatial logic or scale. Such entanglements of affective and epistemic registers also emerges in the work of curators which will be discussed in the next chapter.

The coding table has prompted me to enquire about other reference devices used in curatorial work. As the curator lists the other reference tools guiding his work, he concludes

> Everybody indexes different things and it's amazing how different people's view of the database is based on what they index. Even after having been here for years, I think different people have slightly different visions of the world based on simply what they've come across. GB8

The initial certainty gives way to "different visions of the world". Despite the same coding table in every cubicle, the same codons, the same amino acids, the same data format and record layout (see chapter 6), there remains room for imaginations and differences. Slightly startled by his change of tone, I enquire about the curator's own "vision of the world". Mirroring my own bewilderment, he quickly clarifies that by "visions of the world" he meant the way in which curators approach "complex biological situations". My disenchantment following this revision dissipates in light of the relations that our exchange revealed: the work on database record brings together different kinds of reading, from the factual (the table) to the prosaic ("complex biological situations") to the poetic ("visions of the world"). The complete coherence of the coding table, a perfect circle with an finite set of combination, stands in stark contrast to the empathic bricolage suggested by the booklet. Thus, the orientation devices for the sequence universe encountered so far, Entrez, the booklet and the coding table, betray different, at times incongruous, orders.

Porter's orientations

Porter has located the record for the mumps genome in the Genome database. There, like in Entrez, the information is brought together from various sources, such as the taxonomy database and GenBank. Following her instructions I too find myself looking at the mumps genome. The screenscape that presents itself here contains a number of features. We can

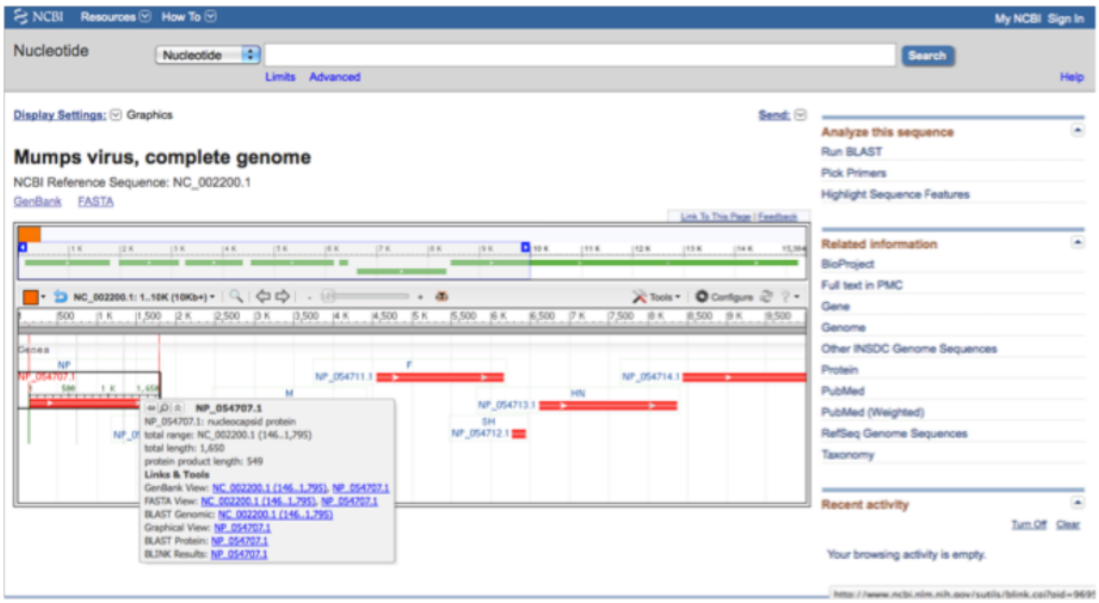

Figure 3: The mumps genome in GenBank (Graphics format)

see the virus' taxonomic lineage from "Viruses" and "ssRNA viruses" to "Rubulavirus" and "Mumps virus". All of these can be followed, via hyperlinks, to their location in the taxonomy database. Underneath the lineage, a table looms bearing various bits of information, some hyperlinked: The number of genes is given as 7. There are 8 proteins coded for. The RefSeq status appears as "Provisional" whereas the sequence status is reported as "Complete". One publication is listed as well as the sequencing centre, the "National Institute of Infectious Diseases, Viral Diseases and Vaccine Contorol (sic)" in Tokyo, Japan.[69] Beneath the table a diagram resides, bearing 8 blue arrows of various lengths pointing to the right along a horizontal axis that records from 1nt (unit of nucleotides) to 10,001nt. This is a simple genome browser, a graphical interface that visualises the genome in a linear diagram showing distance in bases and the location of genes (here in the form of blue arrows). Various routes suggest themselves to Porter. From the genome view she moves on to the taxonomy link where she "could see that other

[69] This is perhaps not surprising given the disease incidents in Japan which had stopped administering the triple MMR vaccine in 1993 (Kawashima et al. 2005; Sasaki & Tsunoda 2009).

paramyxoviruses, related to mumps, have been found in fish (salmon), snakes, dogs, sheep, and pigs". We are, it appears, collating quite the kinship.

Porter inspects the mumps genome record further: "It was then I noticed it. There was a brand new, itty bitty link below the graph of the genome." The first discovery then in our explorations is not a novel gene but a *link*. Retracing Porter's steps, I can see that this link is still there, embedded in the sentence that runs underneath the mumps genome. It reads: "Click here for Sequence Viewer presentation (base sequence and aligned amino acids) of selected region". Like Porter, I take this as a direction and am promptly moved into a new window that displays the NCBI SeqViewer (sequence viewer), which turns out to be the "graphics" display setting of the GenBank record for the mumps genome (Figure 3).[70] By clicking on the link we therefore moved from the Genome database into GenBank. Below the now familiar genome diagram sits another version of the diagram, this one containing an additional range of displaying and processing tools. This now also includes annotated data showing the location and nature of genes, proteins and chromosomes. Like in a traditional genome map view, the coding regions are laid out continuously in stacked rows. It is a useful view for establishing, at first glance, the length of the coding region producing the protein as well as the protein's neighbours. More than that, it offers an annotated display. Porter explains:

> You can see the second gene encodes two different proteins. That's kind of neat. I found, too, that when I held my cursor over the sequences, menus appeared with links to various things that I could do. It turned out I could get FASTA sequences, GenBank records, and pre-computed BLAST results. [71]

Porter marvels at the discovery of the SeqViewer and the menus revealed by the hovering cursor. An equal measure of marvel is directed at the fact that the second gene produces

[70] GenBank records can be viewed in different formats, from the sober "flat file" view to more visual formats such as the one described here. Further details on the way in which records are presented are discussed in chapter 6.

[71] FASTA is a text-based format for representing nucleotide and peptide sequences. It starts with a single –line description and is followed by sequence data in standard amino acid code (single letters representing acids). It derives its name from the FASTA sequence comparison software written by Pearson and Lipman (1988). Since 1989, David J. Lipman has been the director of the NCBI.

two different proteins. To Porter, the SeqViewer evokes the adventurous explorations of computer games:

> Web pages at the NCBI are oddly reminiscent of the games that my kids used to play. (…) [I]n [the game] Millie's Math House, you had to click objects to find out what they would do. The pages at the NCBI are designed the same way. There's no way of guessing ahead of time, you just have to take the plunge and either move your mouse over things or click on random objects, just in case.

Moving her cursor over one of the protein sequences, she finds "lots of links." As the space on screen unravels the features of the mumps genome, it also reveals a range of routes. Put differently, the proteins of the mumps genome turn into places to visit. The next chapter shows how the computer screen, in interactions with the sequence universe, often confounds our expectations of a two-dimensional plane. Here, too, the screen turns into an environment more akin to an online virtual world than a database. And like in classic adventure games, the most innocuous things turn can turn into valiant routes.

Traces in the sequence universe

I have so far proved to be a rather doubtful guest myself. Arriving on a bicycle with a purpose not readily comprehensible to my hosts, my immediate reflections on the environment in which I found myself were full of hesitancy. Yet, in the course of further observation, other odd presences revealed themselves to me, such as the booklet, the coding table and the peculiar but nevertheless instructive relations unravelled by searching *Entrez* and clicking links. At times, travels in the sequence universe can cause the guest to stumble, come to a halt or re-route. This is described in the following paragraphs.

Back in Bethesda, despite the initial restrictions in gaining access, I did not forfeit the possibility for extemporaneous exploration. After exiting the Gateway Center I find myself on a pedestrian pathway leading through the landscaped surroundings of the campus' southeast section. The campus, which forms the NIH's headquarter, is vast. Officially dedicated in 1940 by President Roosevelt, it now houses over 50 buildings and over 15,000 employees. The further I follow along the path, the more structures come into

view. In contrast to the glass and steel aesthetic of the WTG campus, the NIH campus offers a more traditional scenery which nonetheless holds some conspicuous buildings. Set to my right is the William H. Natcher Building (Building 45). Completed in 1994 it comprises a low-rise curved vestibule which houses the conference facilities and a 7-story office block. This is home to GenBank as well as the National Institute of General Medical Sciences, the Natcher Conference Center, the Nobel Laureate Exhibit Hall and the NIH Visitor Center. As I make my way toward the Natcher Building, the path winds to the right, opening up the campus' north side to my view. There the vista is almost entirely taken up by the NIH Clinical Center, the heart of the campus and the largest hospital devoted to clinical research in the US.

To my left sits the National Library of Medicine (NLM), a bold low-rise (five storeys, three below ground) of thick limestone walls that are topped by a seemingly floating hyperbolic paraboloid roof in stressed concrete. A cross between a tricorne and a stealth bomber, it does indeed look ready to lift off.[72] Designed by O'Connor and Kilham and completed in 1962 it is an example of mid-20th century International Style, a style not unusual for large-scale post-war science buildings in the US (Galison & Thompson 1999).[73] Next to it rises the Lister Hill Center for Biomedical Communications, the NLM's 15-story annex that seats its research and development division, including the NCBI. Designed by Carroll, Grisedale & Van Alen (whose Philadelphia airport had inspired Eero Saarinen's designs for Washington Dulles International Airport which I had landed at), the annex opened in 1980. The Lister Hill Center was formally established as a research division of the NLM in 1968 to support and carry out biomedical informatics research. Named after Senator Lister Hill (1894-1984) who had supported its foundation, its homepage prominently bears one of Senator Hill's quotes: "We must develop a communications system so that the miraculous triumphs of modern science can be taken from the laboratory to all in need." As part of the NLM, NCBI pursues a mandate, somewhat implicit

[72] The bomber, it turned out, is more than visual analogy: Inside the library, a massive circular court is cut through the centre of the building. This was to protect from the impact of a bomb (equalising pressure from within) as the building had also served as a designated disaster shelter (Mohrhardt 1962).

[73] Aside from Louis Kahn's Salk Institute, I.M. Pei's Mesa Laboratory of the National Center for Atmospheric Research (1961) in Boulder, Colorado and Robert R. Wilson's Fermilab (1967) outside Batavia, Illinois are stellar examples.

at times, with cognate sorts of concerns: collection, preservation and presentation of, as well as access, to materials pertaining to medicine, health, disease and biomedical sciences.[74] While there are clearly sections remaining that are dedicated to science *communication*, most notably the library and PubMed, these systems have become difficult to disentangle from the products of scientific enquiry and, as the databases make evident, have turned into triumphs in their own right.

The NIH campus displays all kinds of histories: Deliberate traces of the history of medicine, architecture, public health, biotechnology, and the surrounding area, Bethesda, can be found in almost every building I visited. The NLM, Lister Hill Center and the Natcher Building all contain special areas showcasing technological inventions, scientific breakthroughs and individual scientists and projects. Wandering around the ground floor of the Lister Hill, I find a series of small, interconnected rooms displaying various artefacts and histories related to some of the "miraculous triumphs" Senator Hill had referred to. I am the only visitor in a small darkened room and am drawn to the room's centrepiece, an illuminated set of large plates. I had encountered the Visible Human Project (VHP), one of Lister Hill's most famed (and perhaps notorious) research effort.[75] I was familiar with the VHP as this object had corralled sustained technoscientific critique around the gendered and racialised normativities embedded in medical imaging. It remains a striking example for the politics of visibility and readability at work in biotechnological interventions and the resulting economisation of bodies (Waldby 1997; Cartwright 1998; Kember 1998). As such, it established a critical trope where the fragmentation of bodies and the collecting, imaging and naming of these fragments is *always already* a "productive mislocation" (Goodeve & Haraway 1999, p.92), the taking of the abstract (e.g. the gene) for the concrete (life). The encounter with the VHP brought into relief one of the reasons that had made me

[74] Incidentally, yellow fever played a role in the origins of the NIH which emerged in the late 19th century from a one-room laboratory (later known as the Hygienic Laboratory) within the Marine Hospital Service (MHS). In the 1880s the MHS was responsible for screening passengers on arriving ships for infectious diseases, especially yellow fever and cholera (Fredrickson 1978).

[75] The VHP provides a digital image library of volumetric data representing an adult male and an adult female. Data is complied from magnetic resonance imagining (MRI), computed tomography (CT) and anatomical images. The library was released in 1994/5. It was conceived as a "reference for the study of human anatomy, to serve as a set of common public domain data for testing medical imaging algorithms, and to serve as a test bed and model for the construction of network accessible image libraries" (see http://www.nlm.nih.gov/pubs/factsheets/visible_human.html).

embark on this journey: While feminist technoscience critique profoundly resonates with my own politics, I wanted to write against a trope that compulsorily equates data and databases with reduction, abstraction and the preparation of "biovalue" – the database logic (see chapter 1).[76]

Porter's pause

Porter too encounters the past and like me, she is forced to pause for further reflection. She had concluded the first part of her discovery journey which has taken her to the mumps genome. More specifically, we left Porter at the SeqViewer which presents a graphical view of the mumps genome. We can see the order of the genes and their protein products and the spacing between them on the chromosome. From there, Porter begins the second part of her journey, entitled "What do mumps proteins do? And how do we find out?". While not entirely abandoning playful exploration, she now commences a more systematic probing. Behind each protein lies a number of as yet unknown locales. Moving to the first protein, a nucleocapsid protein with a length of 549 base pairs, Porter clicks on it. This unravels a small menu, "Links & Tools", from which she chooses "BLink Results".[77] I do the same and am taken to a list of 100 results (out of 987) hits. BLink is tool developed by the NCBI for protein similarity information. Like BLAST, it compares a (protein) sequence to existing protein records and retrieves the (statistically) most similar. The result list produced by BLink and comprises 5 columns including a graphical representation of the distribution of BLAST-hits and conserved domains over the query sequence, the score, the accession number, the length and the protein description. "Not surprisingly", Porter comments on the results, "I found that the nucleocapsid protein only matched proteins from other viruses."[78] She clicks on the accession number for the first protein appearing in the BLink results "to see if I could learn more from the protein

[76] Waldby describes this as an abundant form of worth that materialises "from the calibration of living entities as code, enrolling them within bio-informatic economies of value which converge with capital economies" (Waldby 2000, p.33).

[77] Other tools and links include GenBank View, FASTA View, BLAST Genomic, Graphical View and BLAST Protein.

[78] In contrast to her search result, mine, carried out 4 years later, show the nucleocapsid protein appearing in 2 bacterial records, which turn out to be one and the same record of a conserved hypothetical protein derived from *Actinosynnema mirum*, an actinobacterium: the first submitted in June 2010, the other in August 2011 (both by the US DoE Joint Genome Institute). The latter record being the proposed RefSeq version but pending approval by NCBI.

record. This part involved a bit of trial and error. Some records had information, some did not." Roaming through the flat file views of the NCBI Protein resource, Porter gleans pieces of information and puts them together to know more about the function of each of the eight mumps proteins.[79] For example, she learns that "the mumps nucleocapsid protein, NP_054707.1 [accession number], protects the viral RNA genome" as well as some details about the structure of the protein. Another "helps make other proteins" and yet another is involved in fusing the membrane of the virus with the membrane of the cell.

Porter always returns to the SeqViewer from where she commences the explorations anew: blinking the proteins, following the results' accession numbers and assembling different kinds of information provided in the record views into a narrative:

> The next two proteins are encoded by the same gene: V/P. The V protein is the smaller of the two proteins. And, the P protein shows us that GenBank is missing a spell check function. The P protein should be listed as a "phos**p**hoprotein" but the name in the menu is "*phoshoprotein*.

I follow in her footsteps and, to my surprise, find the spelling mistake still there. Like my explorations of the campus, Porter's travels too involve encounters with times past. I will not dwell on the mistake, though in a world where individual letters (A, C, T, G) determine kinships, susceptibilities, capacities and worths, such slips can be anything but trifling.[80] Instead, I wish to take this error as a feature or formation in the sequence universe that goes some way in figuring the temporalities enfolded by the universe.

The spelling mistake noted above as well as the wrongful species designation discussed below point to sedimentary layers that betray incidental *histories* within data and the database. On some levels, EMBL-Bank and GenBank remain governed by an archival rationale, which dictates that records, regardless of their level of accuracy or completeness, remain in their place once they have been entered. The next chapter will show that links in the sequence universe can only be broken through tricks, by "telling the database to forget". Here, I am more concerned with how this archival element translates into the topography of the sequence universe. In an interview with a curator at GenBank,

[79] The flat file refers to the bare database record: lines of data and descriptors for each sequence. This is described in chapter 6.
[80] Chapter 7 shows how feelings can run high when annotation is inaccurate.

the matter of withdrawn records and so-called "secondaries" arose. These are produced when previously sequenced pieces of an organism are superseded by the sequence of the organism's whole genome. Fragments retain their accession numbers but when called up (or "plugged in"), the complete, newer sequence will appear. This, as the curator says, is important because

> [y]ou have to track the history of things that are in there. Again it's similar to things that are withdrawn. We have to have a comment. Because if it's just gone, people will have this accession number but ask what happened? So in this way it's tracking this accession number and knowing what happened with it. GB15

Thus, guests and travellers in the sequence universe who linger and look closely can come upon data histories – traces of "lives and works of past generations who have dwelt within it, and in so doing, have left there something of themselves." (Ingold 1996, p.59) Data in the sequence universe is always changing: Short sequences become assembled into larger fragments, which leads to a reduction of visibility for obsolete entries (for example, they are no longer indexed by *Entrez*). Or, annotation is added or changed. Yet, it is important "to track the history of things that are in there" and the provisions made for recording iterations and their trails can reveal multiple temporalities at work at any one time.[81] While I faithfully step in the tracks left by Porter, further traces reveal themselves which once more betray ways in which times become enfolded within the universe. Wandering through and dwelling in the sequence universe affords sights of past doings and future works.

[81] There are two databases that record the "original" raw data output from sequencing efforts: The Trace Archive, established in 2001 by EMBL and NCBI, contains the raw chromatogram data derived from gel/capillary platforms while the sequence reads from next-generation sequencers are archived in the Sequence Read Archive (SRA). The ENA Sequence Version Archive provides access to older versions of EMBL-Bank entries, allowing the user to specify "snapshots" of records at a given date. Similarly, so-called "GenBank Releases", the entirety of the database published for download via FTP since 1996, are available online dating back to 1990 when releases were distributed on CD-ROM by the US Government Printing Office. This is to show that while generation and accumulation of new data often renders data obsolete, they are seen to harbour a future efficacy not yet articulated.

A habitat for doubtful guests

Together with Porter, I return to the mumps genome. It has become our vantage point and the place we come back to and reconvene for setting out anew. She moves on to the next protein "M", a membrane protein, and blinks it. "Interestingly", Porter writes, "this protein matches 468 viruses and *one metazoan sequence*. This is cool!"[82] This odd metazoan is identified by an accession number (AAK76747), its length (340bp) and the protein description "Angrem52 [Homo sapiens]".[83] Like Porter, I click on the accession number and am taken to a record in the Protein database for Angrem52. However, quickly move on to the actual nucleotide sequence from which this protein was inferred through following the link to the "database source". We are transported straight to the GenBank record "Homo sapiens AngRem52 mRNA, complete cds [coding region]". Among other things, this contains links to a paper, via its PubMed identifier, detailing the research that had led to the inference of the protein. Porter follows this and finds herself in the PubMed reference for a 2006 paper published in *Virology* by researchers from the Institute of Nephrology at Peking University, entitled "Beilong virus, a novel paramyxovirus with the largest genome of non-segmented negative-stranded RNA viruses" (Z. Li et al. 2006).[84] The authors describe the discovery of a new paramyxovirus, the "Beilong virus". This previously unknown virus appeared during the course of research that had originally sought to identify the genes switched on by angiotensin, a hormone causing constriction of blood vessels and elevated blood pressure. With the last sentence of the abstract, Porter's journey comes to an unexpected pause:

> Although the exact origin of BeV [Beilong virus] is presently unknown, we provide evidence indicating that BeV was present in a rat mesangial cell line used in the same laboratory prior to the acquisition of the HMC [human cell] line, suggesting a potential rodent origin for BeV. (Z. Li et al. 2006)

[82] The BLink results indeed still show this *one* metazoan record but 919 viruses and 3 results in a group entitled "The Others" whose entries all refer to synthetic constructs.
[83] The protein name Angerm52 stands for "angiotensin II-induced, renal mesangial cell gene 52". Its location in the protein database is at http://www.ncbi.nlm.nih.gov/protein/20384698. Last accessed: 18 March 2012.
[84] The Beilong virus remains an unclassified paramyxovirus.

"Hah!", she writes, "[t]his wasn't a human sequence at all". For Porter, this "goes to show, it's not enough to look at the databases, you need to read the papers or at least the abstracts." Instead of learning more about the expression of a gene (AngRem52) in human mesangial cells, Li et al. "found" a virus, or in fact, "construed" a new paramyxovirus. Yet this virus, as the research paper concludes, did not originate in the human cell lines of their experiment but was most likely a remnant of an earlier experiment that took place in the same laboratory.

The paramyxovirus has lingered in the laboratory and somehow inserted itself into a human cell line. We do not know if the virus is actively replicating and fully functional. Neither do we know whether there is a virus infection present in the cells or if the virus is replicating but otherwise defective. There are other routes by which the virus could have found its way into human cells: the viral genome could have become assembled into the cellular genome or the PCR reaction used to identify Angrem52 could have been contaminated. Also, the cells could have been infected *in vivo*. We do know, however, that the mistake has propagated (more on the kinds of problems this causes in chapter 7) and by doing so, it has opened up odd affiliations, something that viruses are very adept at establishing as the next section will show.

Viral presences

Porter moves onto the next protein in the mumps genome, a fusion protein. The BLink results again show matches to metazoans. Following the relevant links we once more encounter the "Angrgm-52 from homo sapiens". In addition, we now see entries for the genome of *Trichoplax adhaerens*, a primitive marine multi-cellular animal that lives in tropical waters. It quickly becomes evident that the human Angrgrm-52 sequence can indeed be attributed to the same set of sequences of supposedly human mesangial sequences discovered above. Therefore, "we can be pretty confident that this is a viral sequence and *not* a human sequence." Regarding the *Trichoplax adhaerens* sequence, Porter suggests that due to the short aligning region (only 100 of the 538 base pairs), the "paramyxovirus sequence may have gotten included in the Trichoplax genome assembly, the evidence isn't as strong though, as it was in the rat cell line."

Here, then, we encounter a menagerie of doubtful guests. *Trichoplax* is considered as representing the most "basic and ancestral state of metazoan organization" known to date and therefore provides an important species for "unravelling metazoan evolution" (Schierwater 2005, p.1294). It possesses no organs, axis of symmetry or basal membrane and has only 4 somatic cells but according to the Blink results, it apparently can catch mumps. The viral sequence itself has travelled from a rat in a laboratory in Peking to human mesangial cells and continuous to make appearances in likely and unlikely places. Viruses, of course, depend on voyages and hosts to stay alive. This makes them difficult to classify as they change rapidly in time and from host to host. Indeed, their "nomenclature and taxonomy has become tremendously complicated and controversial, even puzzling for some." (Fauquet & Fargette 2005) As we have learnt through Porter's travels, mumps can jump taxonomic classes: The taxonomy link in the mumps genome record revealed a number of trans-species crossings, among them, fish and dogs.

Mumps is a negative-sense ssRNA (single-stranded RNA) virus.[85] In contrast to other viruses (and other kingdoms), its RNA polymerase – a key step in gene transcription and gene expression – does not involve a DNA template but catalyses straight from RNA (it is "RNA-dependent"). This means that the mumps virus, upon entering its vector, must bring along the RNA-dependent RNA polymerase (called "replicase") enzyme. This makes it prone to errors because there is no repair mechanism or (DNA) back-up. Instead of thwarting viral advancement, the resulting error rate has made it extraordinarily adaptable. RNA viruses "are thought to exist within a host as a genetically heterogeneous mixture of variants that differ from a consensus sequence" (Jerzak et al. 2005). The master sequence of the virus is therefore entangled within a "complex mutant spectrum", non-identical but nevertheless related viral variants, which are also referred to as "quasispecies" (ibid.). Because of this genetic diversity amongst the virus, it is exceptionally fit and thereby difficult to treat, especially hepatitis C treatment and AIDS disease progression can be put down to this quality (ibid.).

[85] Other viruses in this group, apart from mumps, include the Borna disease virus, measles, rabies and Ebola virus.

It is curious to observe how the virus retains this capacity in the sequence universe, ostensibly propagating *in silico* and affecting new species. The uncertainties integral to viral lives have indeed caused some strange patterns in GenBank. The 2005 report of the International Committee on Taxonomy of Viruses (ICTV), the official body of the Virology Division of the International Union of Microbiology Societies, listed 6,000 viruses which were classified in 1,950 species in over 391 different higher taxa (Fauquet & Fargette 2005). Yet, at that point, GenBank contained sequences belonging to 3,142 species of viruses not present in the ICTV master list (ibid.). This discrepancy, as the authors point out, is a product of confusing a concrete entity, the viral sequence, with a conceptual one, the species – an artefact of submission and retrieval system as it conflates a new virus name to a new species. So names are given to virus isolates that have produced the sequences in GenBank and are below the species level. The problem is further exacerbated by the fact that "many virus species names and virus names are identical in the words that compose them, except that species names are written in italics" – a stylistic disambiguation which the databases' ASCII character set does not permit (ibid.).

Faced with the experimental residue of the rat cells, it may be pertinent to ask, with reference to Davies (2010), "*when* do experiments end?". Though the first experiment, involving rodent mesangial cells, had concluded, its traces inserted themselves into the new experiment, propagating into new inscriptions. Yet, despite the researchers acknowledging this possible tangling, their sequence, archived in GenBank, clearly states a human origin. The record turns into a bioinformational palimpsest as it documents more than it should. While it still *works*, appearing in search results and, away from sight, contributing to calculations and inferences, its incommensurate excess gives pause to the travellers. Here, quite suddenly, a laboratory space comes onto the scene and superimposes itself, not as a sterile environment but a viral vector – carrying wrong information and potentially contaminating other environments.

Viruses can be very good at travelling on different scales, traversing through individuals and entire populations. They can also overcome large distances, both in space (pandemic) and time (latent). Yet, they are hapless travellers on their own, needing the

company of hosts, vectors, vehicles and favourable happenstance to live and replicate. Viruses "make patterns in time and space" that demand "a political imagination that links together, rather than holds apart, the various controversies that together make viruses into difficult things" (Hinchliffe 2004, pp.229–230). In the sequence universe too, the viruses effect certain trails that come together and weave a pattern in which incongruent scales and unlikely positions sit side-by-side.

Blinking, sightseeing and jumping to conclusions

In Porter's journey we have now moved to the last remaining protein on the mumps genome, the L protein, which is engaged in the production of replicase, the RNA-dependent RNA polymerase essential for viral replication. Once again, the BLink results include odd metazoans, this time referring to a hypothetical protein (predicted but not experimentally confirmed) in *Aedes aegypti*, the yellow fever mosquito. We pay a visit to the Conserved Domain Database (CDD), another resource provided by the NCBI, in order to establish whether the mosquito protein sequence does indeed belong to the same protein family as the one on the mumps genome sequences.[86] A graphical summary presents an axis, this time from 1 to 853nt, and a coloured bar, spanning from 170 to 840nt. The latter indicates the position of a protein superfamily (cl15638) common in paramyxoviruses. The similarity between the hypothetical mosquito protein and this domain responsible for replicase is based on an expected value (e-value) of 2×10^{-57} and therefore significant.[87] But instead of instilling a sense of certainty and orientation, this statistical confidence puts us firmly out of place: The observation does not make sense because, as Porter incredulously asks, "*[w]hy should a mosquito have an RNA-dependent RNA polymerase?*" They do not, as far as Porter knows, require copying of antisense RNA.

Faced with the incongruous case of replicase in mosquitoes, Porter decides to make sure that mosquitoes and other insects really *do not* contain such a protein. From the list of BLink results thrown up by the L protein, we choose the yellow fever mosquito

[86] The CDD offers protein annotation and other information on conserved domains, spatially separated units of the protein structure that can function and evolve independently of the protein chain and are responsible for key cellular processes such as enzyme activity (Malek & Haft 2001).
[87] This is a statistical significance measure and describes the number of hits one can expect to encounter by chance when searching a database of the same size. It decreases exponentially as the score of the match increases.

result, “rhabdoviral-like polymerase, partial (Aedes aegypti)” and blink this sequence to see what else matches the replicase sequence. This reveals matches in 559 viral proteins (and one metazoan protein), the most significant ones relating to replicases from plant viruses affecting, among others, orchid, maize, rice, strawberry and lettuce. To further investigate, Porter follows the “Multiple Alignment” tab and we find ourselves in a new window, the NCBI Phylogenetic Tree Multiple Alignment Viewer, which displays stacked rows of the same protein sequence in different species. Looking at the stacked protein sequences one can distinguish regions of similarity but we need a further tool for rendering the rows of letters into a more intelligible form. On the same page, a menu offers to “Build tree” and Porter does just that. This turns the assembled protein sequences into a phylogenetic tree. This tree offers a visual representation of the genetic relationship between all the organisms sharing considerable parts of that particular protein sequence, allowing the study of differences and similarities within and across kingdoms, phyla, orders and species. Inspecting the tree, the mosquito sequence remains closest to plant viruses. What is interesting here, Porter maintains, is the fact that “mosquitoes pollinate certain kinds of orchids”. She muses:

> “I don't know if mosquitoes pollinate strawberries, but they definitely pollinate blueberries. So, maybe finding a viral RNA polymerase in a mosquito that's most similar to the Strawberry crinkle virus or Orchid fleck virus makes sense. The curious thing now, is how did a viral sequence end up getting assembled into the *Aedes aegypt*i genome? Does it belong there?”

Concluding mix-ups

Porter suggests three possibilities for the match. There could have been, she argues, an oversight in the genome assembly. As will be described in the next chapter, the assembly of genomes is a complex process that entails multiple heterogeneous procedures, from laboratory work to computational modelling and curatorial efforts. To check the integrity of the genome assembly, we return to the Genome database from which we had begun our discovery journey, only this time we visit the mosquito genome. This master record tells us that the genome was sequenced as part of a collaboration between TIGR (now the J. Craig Venter Institute) and the Broad Institute, a genomic medicine research facility at

MIT and Harvard. It also details the actual mosquito whose DNA had been sequenced, a specimen from the Liverpool strain of the species. The accompanying paper listed in the genome record was published in *Science* and explicates the method by which it was assembled. Reviewing the literature, Porter concludes that the virus cannot be attributed to a mistake in the genome assembly.

Another explanation for replicase in the mosquito evokes the possibility of an error, this time in the constitution of the mosquito genome. Perhaps yellow fever mosquitoes do, in fact, contain a replicase gene? As usual, this is a question of *where* to look: Porter stops by VectorBase, part of the bioinformatic resource suite provided by the National Institute of Allergy and Infectious Diseases, once again part of the NIH. VectorBase holds information on and tools for invertebrate vectors of human pathogens, that is, genomes and DNA molecules responsible for carrying and transmitting infectious agents of diseases that affect humans. Naturally, mosquito genomes can be found here (as well as the human louse which we will encounter in the next chapter) but none of them contain the replicase gene. This disqualifies Porter's second explanation and leaves us with her third and last hypothesis, namely the assumption that "the replicase gene might have ended up in the *Aedes aegypti* genome through the actions of a retrotransposon, and the presence of this sequence might be unique to the strain of mosquitoes used for the genome sequence."[88] It appears that the master genome record for *the* yellow fever mosquito is a master genome record of *a* yellow fever mosquito.

And since we are now dealing with an *individual* delinquent, Porter avers that the time has come "to play Hercule Poirot and use those little gray cells to try and reconstruct what happened". Like Poirot, Porter assembles the affected (or infected) parties and begins a demonstrative narration of events:

> I think an ancestor to the Liverpool mosquito was buzzing around one day and sucked some nectar from a plant and got a snoot full of a plant virus. I don't know much about insect reproduction or how the virus ended up near the newly forming germ line cells,

[88] Retrotransposons are the RNA equivalent of the jumping genes discovered by Barbara McClintock (1902-1992), mobile segments of sequence that can insert themselves in a new location on the genome potentially causing mutation as well as an increase or decrease in the amount of DNA in the genome of a cell (Fox Keller 1983).

> but these viruses can make cells fuse together, so I can imagine this happening somehow. When the mosquito cells were dividing, a retrotransposon copied part of the viral RNA and caused it to get integrated into the host genome.

And so Porter explains the discovery of mumps in mosquitoes. This, she suggests, was in fact some plant virus that had gotten itself mixed up with a mosquito a while ago.

Jumping scales

The retrotransposons featured in Porter's final explanation for mumps in mosquitoes are also known as "jumping genes". This indicates a central movement enacted by the travellers featured in this chapter, namely that of "jumping scale" (N. Smith 1992). Recounting her research on maize chromosomes, which would lead to her discovery of jumping genes, McClintock noted that as she was looking through the microscope: "I was *right down* there *with them*, and *everything got big*." (Fox Keller 1983, p.117) Her, her body and its comportment in relation to the microscope together with her acute visual and empathic senses allowed her to inhabit, if briefly, an environment on the same scale as the chromosomes which thereby revealed themselves to her. Such jumps and disproportions have also accompanied my ethnographic travels, allowing me to literally *enter* the databases and materialise selected aspects of their global information infrastructure. Doing so made present boundaries, objects and concerns that too negotiated different scales: the booklet by which data are routed into the database; the coding table which offers orientations on nucleotide sequences; architectures that performatively seek to integrate experimental, data and Cambridgeshire landscapes; and openings (Entrez and the NIH Gateway Center) which connect humans and more-than-humans with vast digital and physical expanses. Porter's journey too betrays many scalar jumps along the mumps genome, unto proteins and into laboratories and scientific literature before leaping to a conclusion that crosses temporal (an ancestor buzzing around one day), ontological (from mosquito to plant to virus to data) and epistemological boundaries, from scientific reasoning to speculative narrative.

Connecting the sequence universe

Gorey's doubtful guest is presented in a sequence of tableaux, showing it set amidst different domestic scenes, accompanied by brief captions detailing its exploits. This chapter has also sketched a number of tableaux – the WTG campus, the genome browser, the NIH campus – and described some of their features such as buildings, objects and proteins. Following the appearance of Gorey's doubtful guest throughout the mansion's rooms and occasions, one is moved to ask questions not dissimilar to the ones posed by Porter: Does it belong there? How did it acquire such human attire if it is very clearly not very human? Can it possibly be of this world? How did it get into the tureen? What does it do laying on the floor? There is also a similar inflection of wonder and incredulity in investigating the doubtful guests inhabiting the sequence universe.

The doubtful guest casts as much doubt over its identity as it does over the space it finds itself in. Like Gorey's creature, the mumps virus that Porter finds in the yellow fever mosquito confounds with its presence. Having inhabited the sequence universe prior to both my and Porter's visits, its appearance is still unexpected and disturbing. And like the creature, it will remain there and elsewhere plaintive, unsettling and potentially lethal. The virus and the doubtful guest are visitors, not unlike the ethnographer. In contrast, however, they have come to their current locale through prior travels we can know very little (if anything) about. For Serres, voyaging and visitations are linked to seeing, scrutinising even (Serres 2009). Bodies of all scales move on all scales, sometimes with bicycles, and with it, our senses move too. Visiting, therefore, is inextricably linked to "the sensible", which in-between all the "paths, crossroads, interchanges" and amidst "changes in dimension, sense and direction (…) *holds together*" (Serres 2009, p.305, emphasis added).

At first sight, our travels appear to have little in common: Our modes of transport vary considerably. So do our motivations, obstacles and maps. Neither do we enjoy the same sights nor visit the same sites even as we follow trails left by prior visitors. Our respective sojourns mount very different challenges: Borders and boundaries, navigation and orientation as well as access become differently enacted. Yet, in the course of our journeys we all prove adept at encountering each other and manoeuvring widely differing

scales and ontologies. In fact, such variation is not an obstacle to travels within the sequence universe, on the contrary. We both use diaries to track and recount our journeys and we both drift into unexpected spaces: While I find myself in front of the VHP, Porter moves from GenBank to a paper (Li et al. 2006) and ends up in a Chinese laboratory. Visualisations, metaphors and models can be seen as part of Porter's repertoire as well as mine: They help scale and transport objects and arguments. While there is no social science equivalent to BLink, the formal operation at its heart – matching similar strings of letters in a continuously changing archive of sequence – is perhaps not unfamiliar to the ethnographer who seeks to discover patterns in the activities and environments she observes that match patterns observed and/or described elsewhere. There too, homologues, orthologues, neighbours and kin are of interest and provide important references for making arguments.

Through recounting journeys and the sites and objects they unfold, this chapter has drawn up an incongruous topography, rendering present different sites, incommensurable scales and unlikely affiliations.[89] Landscapes gave way to datascapes and libraries which returned to laboratories before resolving someplace where a mosquito allegedly sucked on a strawberry. Here, the sequence universe resembles "not a network of connected points, but a meshwork of interwoven lines." (Ingold 2011, p.62) Certain things appear in places assigned to them. Others appear out of place because of spelling mistakes, wrong species designations, sequencing errors or erroneous translations (where the product stipulated by the submitter is not the product of the submitted sequence). Algorithms, in the form of BLAST and BLink, can still process such misplaced sequences because once data has been accessed, it has become and remains *connected*. While this raises the huge issue of data quality (which is discussed in more detail in chapter 7), the primacy of connectivity points towards decidedly topological formations. In these travels the sequence universe emplaces new kinds of constellations: ethnographers, spelling mistakes, viruses, histories, laboratories. Instead of a reference, the databases provide a

[89] There are, however, many absences which these travels have not accounted for. The playfulness of exploration that Porter is undertaking hides some uneasy sights such as the suffering of animals hiding behind the genome browser record. Similarly, the environmental sustainability of the WTG campus, the concerns for preserving the wetlands around the campus stand in stark contrast to the activities involving laboratory animals carried out in research facilities on campus.

habitat where viruses confound and ethnographers and scientists alike can make different kinds of discoveries: links and hidden views as well as functions and tools; unexpected relationships and kinships between organisms; diagrams and graphs and interactive trees; mistakes and histories.

To narrate Porter's journey, I have also taken the journey myself. Adhering to her directions, I found myself in the same places she had described, mostly being greeted by the same vistas and meeting the same entities. Sometimes, however, things had changed and I found myself in unchartered terrains. Retracing someone else's steps is a particular kind of travel, not a style of journey I would have expected in the sequence universe. Following someone else's steps and encountering some of the same sights makes for strangely affective spaces. There is a curious joy in encountering the same mistakes and spelling errors. Here, the journey does not just unfold in space but also, very concretely so, in time. Porter's discovery journey here becomes a trail. Given the rate of queries preformed within the sequence universe, the amounts of sequences blasted and searched, there are millions of trails created each day. Aporta (2004) has described how trails enact the indigenous geography of the Arctic, a space deemed empty and featureless by outside explorers. Trails are stories as well as maps and archives of the landscape, recording sights and sites but also change. At the same time, they are meeting places, veritable "homes", and sites for important exchanges and discovery. As we were blinking our way through the sequence universe, Porter and I were following pre-computed paths determined by a multitude of algorithms that analyse, compare and match strings of letters, guided by complex statistical models. But measuring and dreaming are part of this integration. It was a narrative that in the end explained mumps in a mosquito: Imagination is not opposed to database logic but integral to it. This database imaginary does away with other perceived distinctions too. In the next chapter, we will see how practices are manual (human) *and* automated while chapter 6 examines present absences and likely presences.

Chapter 5. Curating sequence: Visions of the universe

This chapter examines the work of curators and developers at GenBank. On the basis of observations and interviews, it describes some of the heterogeneous activities that afford the timely and accurate handling and publishing of nucleotide sequence data. The description of biocuration is framed with reference to the specimen-making practice of Joseph Grinnell (1877-1939), first director of the Museum of Vertebrate Zoology, Berkeley, and thereby maps traces of continuity between curating natural history and curating sequence data. The chapter focuses on three distinct routines: The *triage* of direct submissions, the handling of whole genome shotgun (WGS) submissions, and the maintenance activities of dataflow management. It suggests that these activities enrol multiple *ways of seeing*, which, in turn, co-produce different figurations of such data. Here, vision is understood to exceed ocular capacities and involve human and more-than-human bodies, machines and their interactions.

Introduction

> To secure a really practicable scheme of arrangement [of specimens, card indexes and data on specimen labels] takes the best thought and much experimentation on the part of the keenest museum curator.
>
> Joseph Grinnell, (1968[1910])

Joseph Grinnell (1877-1939), first director of the Museum of Vertebrate Zoology (MVZ), summarises the requisite characteristics of a museum curator as detailed knowledge, a propensity for experimentation, and enthusiasm.[90] Evidence that such qualities remain

[90] Grinnell directed the MVZ from 1908 to his death in 1939. He was a prolific naturalist of his time and left a substantial body of work on ornithology, conservation and wildlife management as well as on zoological methods. Grinnell also undertook large-scale faunal surveys, such as *Animal Life in the Yosemite* (with Storer, 1924) and *Vertebrate natural history of a section of northern California through the Lassen Peak Region* (with Dixon and Linsdale, 1930). His work was re-appraised by historian of science James Griesemer (1990; Griesemer & Gerson 1993). Most importantly for STS

vital even for today's curation of biological data (biocuration) was abundant in my observations and interviews with curators at both EMBL-Bank and GenBank.[91] This may, at first, appear surprising given that museum curators tend to deal with a very different *matter*, namely natural specimens. But the Grinnellian specimen was not like any other specimen. It was a product of somewhat ornate curatorial practices, including meticulous standards for recording, collecting and describing specimens, devised by Grinnell himself (1968; Grinnell & Storer 1924; Herman & Grinnell 1986). And it was intimately connected to the scientific knowledge produced at the MVZ (Star & Griesemer 1989, p.393). This knowledge, the authors note, was looking to accommodate two newly emerging concerns: On one hand, classical descriptive natural history based on collection, classification and identification was giving way to laboratory-based experimental biology (Kohler 2002a; Kohler 2006). On the other, Grinnell embraced an emphatically ecological approach, adamant about considering habitat and organismal interactions in his faunal surveys.[92] His curatorial method concentrated on *specimen-making* and at once reflected and supported these two developments. Grinnell valued *documentation* in relation to specimens for

> rendering what we do obtain as permanently valuable as we know how, to the ecologist as well as to the systematist. It is quite probable that the facts of distribution, life history, and economic status may finally prove to be of more far-reaching value, than whatever information is obtainable exclusively from the specimens themselves. (Grinnell 1968[1910], 39)

Labelling specimens, attaching information about their provenance, means of extraction, *in vivo* dimensions, behaviours, life history and more indirect measures, were as integral as tending to the physical specimen (Carson 2007).

I suggest that the curatorial practices at EMBL-Bank and GenBank, concerned with *making database records* in relation to a particular environment, offers insight into a similar moment: The last twenty years have seen an exponential increase in scientific data

---

scholars perhaps is Star and Griesemer's text "Institutional Ecology, 'Translations' and Boundary Objects" (1989), which relates the use of "boundary objects" in Grinnell's MVZ.

[91] Biocuration is still a fledging discipline. References to "biocuration" have only started appearing in scientific publications (e.g. *Nature*, *Science*, *PLoS Biology*) from 2006 onwards.

[92] Describing faunal distribution in the Yosemite region, Grinnell refers to "associations", minor units of habitat that are not bound by zonal restrictions (life zones such as "arctic alpine" or "Hudsonian Canadian") and that can contain different "assemblages of birds" and other animals (Grinnell & Storer 1924, p.24).

production, an intensified reliance of scientific discovery on data and a multiplication of data uses for science. At the same time, the concentration on the production of data has been accompanied by ever more sophisticated ways of making sense and working with this data. Moving away from decoding genes and genomes, new *ecological* approaches to biology such as systems biology (O'Malley & Dupré 2005) or integrative biology (Wake 2003) have emerged. Similar to Grinnell's rigour in recording the organism through its relations with a wider environment and inscribing those relations into the specimen, biocuration (also termed "data curation") strives for the integration of different modes of being within the database record: *in vivo*, *in situ*, *in vitro*, *in silico*. As will be detailed in the next chapter, these can include biotic and abiotic factors, institutional affiliations of researchers and fungal strains as well as molecular occurrences (such as coding regions). Similar to Grinnell's specimen, the record is thereby assembled with a comprehensive view to presenting the nucleotide fragment in relation to wider contexts. In order to assemble and correspond elements on such divergent scales (e.g. local, molecular, regional, institutional, temporal) curation has to look inside, beyond and between specimens. In this chapter I suggest that both projects, Grinnell's specimen-making and today's assemblage of database records, enrol certain kinds of visions – ocular, imaginative and technologised ways of seeing and projecting. Again, the work of biocuration expands on these provisions while challenging some of the conventions that have been articulated in relation to the "molecular gaze" (N. Rose 2007; Nelkin & Anker 2003).

Biocuration

Curation at EMBL-Bank and GenBank is carried out by "curators" (EMBL-Bank) or "indexers" (GenBank). Despite the different designations, I will refer to curators and curation throughout. Instead of specimens, curators at EMBL-Bank and GenBank handle digital data and information. They spend much of their time in front of computer screens, using custom-made computer applications to review, validate or edit incoming sequence submissions and correspond with submitters. But they also present at conferences, write papers, devise guides and user handbooks, deliver training provisions, and design

standards (for data handling, workflows, metadata provisions and so on). At both sites, curators work very closely with technical developers and software engineers who maintain and develop the database management systems as well as a host of groupware tools, applications and software suites (discussed below). A large percentage of curators at GenBank are women (the ratio of women in EMBL-Bank is considerably lower).[93] All curators had PhDs though in different scientific subjects, including genetics, cell biology, biochemistry and molecular biology. Asked about their reasons for working as curators, scientists highlighted the non-hierarchical and collaborative work ethic and the relative ease by which one could attain a sense of achievement as well as stable working hours. Most of them had previously worked in wet labs.

Certainly, with the progress and diversification of genomics, the role of the curator has developed a new urgency while also testifying to the intrinsic interconnection between collecting and experimenting, practices that have commonly been disarticulated and placed within distinct spaces (the field and the laboratory, respectively). After all, attaching "labels" in the form of contextual information to nucleotide sequences is essential for making sequence data meaningful and usable. Curation has become and continues to be a critical activity in bioscientific research, evidenced by mounting literature (see chapter 7) and resources dedicated to its pursuit. Nevertheless, it remains largely "infrastructural" and therefore mostly invisible work: There is no formal scientific merit attached to curating data. Neither is it regarded as a discipline *per se* (there are, at writing of this thesis, no degree programmes for studying biocuration) though the biocuration community has recently begun to organise itself professionally (the International Society of Biocuration formed in 2010) and to contribute their very own perspectives to the literature.[94] At the same time, calls for more sustainable curatorial

[93] There is great scope in studying biocuration with attention to the gendered aspects that still bear on science careers and hinder the progression of women scientists. Due to its focus on the processing of data, it could prove a fertile ground for challenging notions of "human computers" and "blue-collar science" (Grier 2005).

[94] On the curation work undertaken at the model organism database FlyBase see, for example, St. Pierre & McQuilton (2009).

practices are mounting in the face of the increasing reliance on scientific data for discovery and experimentation.[95]

What has become known as "biocuration" generally describes "the transformation of biological data into an organized form." (Bateman 2010) Turning raw data into an intelligible, available and useable resource primarily requires *annotation*. This denotes the gathering and appending of contextual information in order to qualify, in the case of EMBL-Bank and GenBank, the string of As, Cs, Ts, Gs that has been generated by sequencing machines.[96] Such annotation can, for example, include descriptions of the function of certain genetic regions within the genome of an organism. Curators are responsible for ensuring the presence and quality of annotation. GenBank curators check submissions to a certain extent but given the exorbitant rate of submissions, they cannot manually edit the annotation which is for the most part assembled through bioinformatic methods and machines (called "annotation pipelines").[97] Data standards, correct names and unambiguous classification schemes are of utmost importance in annotating sequences, especially considering that errors will propagate, as we saw in the previous chapter, in all analyses taking place after sequencing (this includes all annotation, gene recognition as well as identification of phylogenetic relationships). In practical terms, as Pierre and McQuilton (2009) note, curators have to identify relevant literature and translate results into a standardised language. In addition, curators need to be able to

[95] There are also, a growing number of initiatives, usually constituted around individual model organisms, that carry out and promote the curation of data relevant to their community. These have given rise to databases built around specific (model) organisms such as TAIR, Fish, E.coli and Flybase.

[96] Chapter 7 shows how annotation can turn into a somewhat controversial issue.

[97] For the first-level of annotation (also called "one-dimensional annotation" by Reed et al. 2006, and "nucleotide-level annotation" by Commins et al. 2009), methods are commonly based on prediction and similarity (to experimentally confirmed genes or proteins). Here, gene-finding algorithms and sequence-homology search tools, such as BLAST, compare and predict on the basis of similar sequences that have already been annotated. The outcome of this procedure is a "map" of known genes, markers and landmarks as well as predicted gene locations. This routine annotation is based on the assumption that similar sequence has similar function, an assumption that itself has become challenged. Likewise, "two-dimensional annotation" or "protein-level annotation" is based on predictions that take into consideration gene neighbour, gene cluster or phylogenetic profiles in order to establish patterns. In this step "[g]enes are named and assigned functions mostly by means of comparison to already annotated genomes" (Commins et al. 2006, p.60). This often involves placing proteins "into 'unknown function' or 'hypothetical protein' categories until experimentation provide light on the purpose of the gene at hand." (ibid.) The third step of annotation, process-level or functional annotation, provides details about the biological processes affected by the gene(s) in the sequence such as cell cycle, metabolism or immune response and, as the previous steps, is determined by means of comparison with available information.

confirm the quality of data and develop or extend standard formats, semantic support and data processing applications for biological data. Hence, curation requires high-level knowledge of genetics and molecular biology as well as competency in navigating and using a vast suite of bioinformatic resources, to wit, an understanding of the role and functions of the sequence universe.

At the databases, curation is concerned with handling and checking sequence and other primary data. EMBL-Bank offers the following description of the role of its curators:

> The curation team guides submitters through the submission process. They take the unique opportunity to obtain directly from submitting researchers exact provenance information on the sequenced sample and on the methodology surrounding its preparation for sequencing. Curators sort submitted data, fix errors and resolve taxonomy issues. They provide a helpdesk and generally mediate communication between the scientific community and ENA software engineers. (…) Curators also maintain the annotation guidelines and are involved in the data integration. (ten Hoopen et al. 2010)

Curators gather provenance and methodology information as well as address issues, such as taxonomic inexactness. Curation, as this description reveals, encompasses not only strictly "scientific" tasks but involves decidedly social, pedagogical, even psychological ones: Curators *guide* researchers through their submissions and offer a helpdesk for researchers preparing submissions. This involves continuous dialogue – most curators sustain multiple correspondences at any one time while also continuing conversations established elsewhere (when submissions move from triage to indexing for example, see below). Here, curation requires not only patience but also the ability to translate questions and concerns brought forward by submitters unfamiliar with or indifferent to the system. Observing curators' work, many instances were concerned with reading and re-reading submitters' emails in an attempt to properly understand queries and, more importantly, respond to such queries accurately. The rhetorical capacities of curators must encompass explanation as well as in some cases assuagement and respectful probing. The description above highlights the latter by speaking of "extracting" information from submitters – a curious inversion where instead of the organisms and DNA fragments, scientists themselves are explored. After successfully eliciting information, curators engage in

sorting, fixing and resolving data and "issues", taxonomic and otherwise. They are veritable "plumbers" as one respondent put it (see below) in maintaining the flow of information coming through the pipes but also engineers, teachers and diplomats in taking on a mediating role between the database, submitting scientists and the entities contained with each submission seeking admission into the sequence universe.

Looking into curation

As suggested earlier, curation is an activity which requires the combination of different ways of seeing. The following will elaborate on this proposition by means of depicting the handling of incoming sequence data at GenBank. At the same time, I wish to draw a relation between curatorial visions and the particular landscapes it manifests. It is here that I return to Grinnell for an instructive precedence in investing curatorial work with ecological concerns.

Submissions to GenBank are processed in two different ways, depending on whether they are Whole Genome Shotgun (WSG) submissions or "direct submissions". WSG curators handle high-throughput submissions produced by WSG projects, that is, projects that generate large amounts of overlapping reads that are subsequently assembled into genomes.[98] This will be described in more detail below. In contrast, direct submissions refer to smaller submissions of nucleotide sequences that are received via the two automated submission tools, BankIt (at EMBL-Bank an equivalent program is called Webin) and Sequin. Whereas BankIt is a browser-based workflow for submitting single sequences, Sequin is a stand-alone menu-driven program created by the NCBI. It features a graphical interface and facilitates submission *and* editing of multiple, long and complex sequences. At its most basic submission level, it supports the combination of a nucleotide sequence with a five-column table of feature locations and qualifiers (e.g. protein product). GenBank direct submissions pass through three curatorial levels: "triage", "indexing" and "on-call". This process is described by a curator as follows:

[98] The WGS method was famously championed by J. Craig Venter and his Celera team in decoding the human genome and remains the preferred method due to expediency and cost-effectiveness. Grinnell too favoured the "shotgun method" though for entirely different reasons. Here, the "shotgun method" refers to the obtaining of specimens from the field ("skin records") and, unlike mere field observation by the "opera-glass student", guarantees "precision and accuracy" (Grinnell 1968[1915], 65).

> Triage takes a quick look at the sequence and the annotation to make sure that it's got all the nuts and bolts that are required to be able to get an accession number. In other words, they're looking for the minimum requirements that need to be fulfilled such that an indexer can finish process the record successfully. In indexing you double-check all the annotation again using BLAST analysis. And then, after you're done indexing, you're supposed to put it aside for a day and then come back to it the next day and give it a second look. Once you've given it a second look and deem it ready to go to on-call, you send it to on-call. And on-call is where a third member of the crew double-checks your work. Every direct submission record goes through triage, gets a quick look to check that the nuts and bolts are there, goes through indexing, for a little bit of a polish on it, and then goes to on-call to make sure that the indexer didn't miss anything. (GB8)

"That's indexing", he curtly concludes having laid out the curatorial trajectory as characterised by specific passage points: triage, indexing, and on-call. His description suggests that curation is very much a "looking after" submissions. As I will show, this looking after encompasses *care* as well as, more literally, *vision*. Interestingly, it enrols different kinds of sight: from *eyesight* in triage where one "takes a quick" diagnostic look to *oversight* in on-call where the submission receives its final once-over. Taking a "quick look" and "looking for" the presence of minimum information also requires *foresight* as to the demands of the wider system – how does it, for example, fit into established categories and if it does not, how to handle it? This is followed, in the indexing stage, by a more probing gaze in "double-checking" that often calls for *second sight* in anticipating future problems and conditions. Then, after a period of not looking (*blindsight*), the curator returns to "give it a second look". Once the submission meets approval of this second look, it is "double-checked" again and inspected for any omissions. Here, the look extends beyond the submission to take into view the work done in previous stages.

Each of the passages thereby is associated with a different way or intensity of *looking*. Whereas *triage* calls for a glance to ensure the "nuts and bolts" – a superficial scanning of the submission – the next stage, *indexing*, demands a more endemic and invasive gaze. This is supported by the BLAST tool, introduced in chapter 1, facilitating a look that not only penetrates the visible surface but also distributes the inspecting gaze to other entities, other database records that have already acceded into the database and

now form approved sequence records. The last passage point, *on-call*, is manned by curators with at least one year of curating experience. On-call demands a look that takes into its purview not just the submission but also its previous passage points (this includes other curators, their correspondence, personal dispositions, expertise and so on) and any modifications that might have occurred in its course.

The transitions in the biosciences that marked the times of Grinnell – the move to academic institutions and laboratories, the replacement of societies and amateurs for research scientists, the shift from natural history to academic biology – was accompanied by changes to the vision *of* and *in* the biosciences. The study of classification and morphology gave way to an interest in process and function. Instead of "observational and comparative approaches, biological methods came to include experimental, manipulative and quantitative techniques." (Star & Griesemer 1989, p.394) I suggest that the molecular turn has prompted another transformation in bioscientific visions, further unsettling the primacy of ocular vision while revealing and exploring novel sites, scales and dimensions. The previous chapter has portrayed some of those sites but we could also consider the wondrous shapes produced by folded proteins, the visualisations of molecular pathways or the odd couples established by phylogenetics as well as new screening and diagnostic techniques which extend our field of vision not just spatially but also temporally.[99]

Vibrant visions

Seeing, therefore, is always entangled with broader visions, imaginations and technologies and has served as a rich object through which to explore not just particular spaces but conventions and aspirations (Berger 1972). Observing the densely forested mountain ridge from my vantage point in the Grizedale Sculptural Park, I once asked a fellow visitor who happened to work for the local Forestry Commission what it was that he did. He followed my gaze and pointed to a tree that stood slightly taller than its surroundings, breaking the gently curved tuft topping the Lakeland Fells around us. "We cut those down," he said unceremoniously. A forest warden's task was to look after the forest but at the same time to look out for a vision of the forest and by extension, a vision of the Lake

[99] Falcons are closer related to parrots than to other birds of prey (Suh et al. 2011) while crocodiles share more in common with chicken than turtles or other reptiles (Larhammar & Milner 1989).

District, its landscape a naturecultural heritage. Here, as Ingold (2000) suggested, landscape is more accurately understood as *taskscape* because it is the product of accumulated activity (while co-producing the very directions of this activity). The sequence universe could be grasped on comparable terms if we take into account the work of biocurators: sorting, fixing and resolving data inscribe themselves into a bioinformational topography which, in order to be divested of the "natural" character that often clings to landscape, could do with untangling some of these activities.

One curator described his work as a process to (accurately) bring out a record's talents and capacities:

> Assuming something is known about a piece of sequence data, we'd like to have some sort of biological annotation on the record that gives the user some idea of what this thing is. For example, is there a gene present in this sequence and if so what is that gene? Where is that gene? Does that gene encode a protein, if so what is that protein? Where is that protein? What's the amino acid translation of the protein? What's the biological function of that protein? (…). When we have that kind of information associated it makes it a much more rich and *vibrant* and useful entry for the community. (GB7, emphasis added)

GB7 explains that one of his main tasks consists of enriching the sequence data, of making submissions "*vibrant* and useful". Just like the museum curator in Star and Griesemer is seen to resurrect the dead specimen, the database curator re-animates the sequence. What is more, it is this vibrancy that makes it useful: harbouring a gene, identifying its name and location as well as tracking its products and allowing associations to higher-level realms such as biological function.

In GB7's account the record's vibrancy becomes linked to its intelligibility and usefulness in relation to the community: Enriching it allows the record to effectively work together with other sequences, experimental results, specimens, scientists and environments. At the same time, appending inaccurate information obstructs certain functionalities while permitting other kinds of cooperation (some of these are detailed in chapter 7). In the respondent's quote, making the record vibrant means asking questions and assembling responses in the form of annotation. But as the succession of her questions shows, such responses do not come into being as discrete elements. Instead,

they take the shape of a structured progression, they unfold as a narrative: From first establishing the presence of a gene to its characteristics and location, it moves to the gene's products (such as proteins) and their subsequent biological functions. Once this narrative is assembled and associated with the record, the record has gained in vibrancy. And it has become a better "spokesperson" since now it is also more "useful (…) *for the community*".

In examining curation and its multiple visions, this chapter then takes another step towards furnishing the sequence universe. What are the ways of looking and seeing demanded by the curation process? If vision, as Cosgrove (Cosgrove 2008) suggests, is physiologically and historically specific, then what kind of vision does the sequence universe offer? Whereas chapter 4 has explored the experimental topographies afforded by the sequence universe in relation to travel and discovery, the following sections examine how this topography is built and maintained.

Triage: diagnosing sequence data

> Honey, I rearranged the collection to remind everyone that the original definition of a curator was: A guardian of a minor, lunatic; a person who has a cure of souls.[100]
>
> Allen Ruppersberg (1999)

Allen Ruppersberg's installation "Honey, I rearranged the collection while you were gone" (1999-2002) offers not just a laconic swipe at the curatorial regime that dominates the contemporary art world but also points to the oftentimes absurd routines that accompany collecting. These might be understood in relation to exaggerated affective investments – as evidenced by Ruppersberg's expression of affectionate attachment to both his partner ("honey") and his collection. Curation as I have observed it at the databases evinces similarly intuitive processes. In particular, guarding and looking after data should be

[100]Quote from the American conceptual artist Allen Ruppersberg's work entitled "Honey, I rearranged the collection while you were gone" on show at greengrassi gallery, London, 26 May – 3 July 1999. The work consists of a series of post-it notes stuck unto framed photographs and screenprints that depict library interiors, bookshelves, advertising images, and documents. Each note bears a pencil-written statement that commences with the words "Honey, I rearranged the collection" followed by a description of the logic informing the order of the rearrangement, e.g. "Honey, I rearranged the collection to showcase the work we got before anyone else even heard of the artist." Or, "Honey, I rearranged the collection so that it represents my secret life. I'll be back in 2 weeks."

understood with reference to the word's etymology (curare), as *taking care of data*. The following section describes the *triage* area in GenBank, the first passage point for incoming direct submissions, where the cure of data finds its most caring expression.

In triage, curators review submissions, check whether they meet minimum criteria (such as minimum length) for incorporation in GenBank and issue an accession number to each sequence within 48 hours of receipt. Accession numbers are unique to each sequence and where submitters fulfil a deposit mandate stipulated by a journal or funding body, the accession number serves as proof of deposit. In triage, as one respondent put it, the "nuts and bolts" are inspected:

> Did they [the submitter] run the program correctly, did they put in all the required information, have they sent everything that they were going to send, does it look like something that fits into the criteria of what we accept, are there any huge obvious problems with their submission? (GB7)

Curators take turns in staffing triage. Unlike the general curating section, which occupies a large open office space divided into individual work places by low partitions, the triage area is housed in a separate office with three work stations. The triage office is quieter than the rest of the floor. Also, the lighting appears to be slightly dimmer than in the main space. It occurs to me that this atmosphere befits the kind of work undertaken in the process of *triage*. In its original context, triage refers to the assessment of the wounded on battlefields and in emergency rooms. And just as in those sites, triage at GenBank determines the further course of action, combining a diagnostic gaze with prognostic outlooks. Given the ambience of sombre efficiency that dominates the office, there too, the *triagers* take the condition of their (sequence) patients very seriously.

I take my seat next to one of the two curators currently staffing triage. After a few introductory remarks, he resumes what he had been working on prior to my arrival. For my benefit, however, he provides a running commentary to his routines. We are looking at submissions through two customised groupware tools developed by NCBI programmers. These integrate the different elements of the submission and display them on the computer screen. The first program, called "Smart", facilitates the management of email correspondence in relation to incoming sequence data. Selecting a particular record in this

window will open up another program, called “Sequin”, which functions as the main editing tool. This lets curators manage, review, update, annotate and validate incoming sequence. Sequin’s Spartan interface bears resemblance to the classic grey dialogue box premiered with the Windows 95 operating system, revealing some of the stability and complexity built into the tool. One curator estimates that Sequin allows almost 10,000 manipulations to be carried out on the submitted data. I am reminded of what a software engineer had told me the previous day: “In software if it works, you better not touch it. Maybe it's not efficient, maybe it becomes slower but if it still works then don’t touch it!” (GB15) Sequin’s interface certainly looks like it hasn’t been touched in the last ten years. Yet, this visual throwback betrays an ongoing struggle between legacy structures, present demands (for example, for accuracy) and future expectations, such as mounting data volumes, that plays out as much on the level of software provisions as it does on the epistemic plane. On an institutional level this tension becomes most evident in the avowal of the databases’ archival rational (see chapter 1) and comes to a head in chapter 7, when the archive is faced with the “chaos” of “wikification” according to David Lipman, Director of the NCBI.

The physique of data

Rapidly taking in the various bits of information on screen, strategically zooming in on certain features, the curator quickly establishes whether it “looks right or wrong”. The speed by which he navigates the submissions is dizzying. Hurried successions of mouse clicks are followed by relentless scrolling up and down, left to right, back and forth. I enquire about the speed and he acknowledges that it is “pretty quick, yeah. Unless it's something that catches you, that you need to go back and go slower.” (GB7) For the moment though, nothing seems catching as he races through the windows and dialogue boxes on the screen. Watching him move through submissions, clicking buttons and scrolling across screens at a breakneck pace, the spectre of battlefields and emergency rooms re-appears. Once again, a sense of urgency is played out. I resume my casual questioning.

TN: You have to look through every submission?

GB7: Yeah, every submission has to be looked through as they come in.

TN: How many do you check per day?

GB7: It's a lot. (…) We're giving out in excess of 50,000 accession numbers a month. It's a lot of sequences you look at in a day.

This amounts to an average of approximately 2,500 accession numbers issued *per day*. Granted that a considerable part of those pertain to one submission (with multiple sequences) and WGS submissions, this is still an improbable amount of data to *look at*.[101] Such hurried workings impel questions pertaining to the kind of diagnostic gaze enrolled by the triager. As Saunders has shown in his study of computed tomography (CT) diagnosis, the diagnostic gaze is not "a simple, coherent, or merely visual experience" but encompasses "a multiplicity of gestures" (Saunders 2008, 18). Here too, the gaze involves a continuous shuttling among windows, looking back and forth between different pieces of information, running validation checks, clicking and scrolling through emails, annotation files and taxonomic trees. Like in medical triage diagnosis (Olszewski 2003), the curator's gestures combine certain signs (symptoms) with appropriate narratives and standardised classifications and terminologies.

Curators in triage review multiple parts of submissions: They read the accompanying correspondence, they scan through the flat file view, skim through the list of attached files and even glance over submission files in the ASN.1 (Abstract Syntax Notation One) format, an international standard notation that defines data types transmitted by telecommunications protocols – from simple integers to more complex ones such as sequences or sets.[102] ASN.1 institutes a crucial level of abstraction that lets different parts and systems, from the physical hardware to end-user applications, communicate with each other. Where I in search for the database's key normalising mechanism – the moment where all complexity is compelled into computable categories and concepts – ASN.1 would certainly offer itself as a convenient moment. Instead, the

[101] The speed of triage also underscores the fact that for submitters, accession numbers are much coveted articles – without one they will not be able to publish their findings as many biology and biomedical journals require accession numbers *before* publication.

[102] ASN.1 has been an international standard since 1984. See http://www.itu.int/ITU-T/asn1/introduction/index.htm. Last accessed: 18 June 2012.

curator's description of format suggests a different kind of moment, one that will be discussed in more detail in the next chapter and that sits at the heart of this thesis: Databases like EMBL-Bank and GenBank do not abrogate mess and heterogeneity but impart it. In his words, ASN.1 is "all this stuff [waves at the piles of data on his screen] which you do learn to read after a while but it's more designed for a machine than a person to read." (GB7) Even at its deepest and most abstract end, there was room to "read" in the database, room for narratives, uncertainties and skills to matter. Another curator notes that ASN.1 allows for a level of variability that affords a "global look" as it "generates [the] different views" (GB16): It makes the GenBank flat file, the table format as well as the many graphical views, such as taxonomic trees or the SeqViewer explored in chapter 4. Each view becomes a stage for specific questions, narratives and interventions to take shape.

In triage, we move on to the next submission. The curator remarks:

> And then we look at the next one, which is a different set of sequences. So he's put them into groups where he has 37 or whatever the number is of the same type of sequence. So he's doing comparative sequencing analysis. GB7

Based on the overall form of the submission (and not its specific contents) the curator infers the purpose the sequences served while still in the laboratory. The submission consists of sequences of the same region across different organisms, the repeats an indication that we are looking at an instance of comparative sequencing.[103] We read through the information provided by the submitter and learn that the sequences derive from a viral structure protein of the canine distemper virus. The curator tells me that we could be looking at different isolates from the virus, either from different animals or different geographical locations. They could, he muses, even be from the same source before and after drug treatment.

We are now viewing a submission from a laboratory in China that had been received the previous week lacking any annotation. The initial triage had resulted in a message to the submitter asking for further information. So at present we are faced with

[103] Comparative sequencing is used for example in determining phylogenetic distribution.

an updated version of the same submission, this time containing (some) annotation. The curator opens the attached files generated by Sequin:

> Let's see what he did. So here we have some of the information although not the ideal information [laughs, exasperated]. He used a very generic thing called a "misc" [miscellaneous] feature instead of… Ideally we'd like him to define each. There's three regions, there's several genes and there are some regions which we call spacer regions that are not a gene according to the official definition of that. And he says "oh it contains all these things" but he doesn't tell us where each of those things is. GB7

Evidently, all is not well with these sequences but the curator very quickly establishes the extent of the damage and the level of intervention required. The curator's concern with proper and thorough definition is a minimal concern for most submitters who are more interested in communicating the findings which the sequence helped achieve. Listening to and watching the curator as he diagnoses the submissions, the familiar sequence of letters (A, C, T, G) morphs into something very different. Under the curatorial gaze, genes, Chinese laboratories, sloppy scientists, spacer regions and database classifications appear. Mol (2002) argued that "bracketing the practicalities" of disease, that is, rendering the apparatuses of diagnosis (screening, measuring, extracting, etc.) invisible, locates the disease inside the body. It is a necessary practice that coalesces the heterogeneity of different enactments into an *object* thereby facilitating the passage from diagnosis to treatment to management. Triage, and curation more generally, is very much concerned with *unbracketing* practicalities, rendering visible the ways in which sequence data was produced. The moment marks in fact several passage points: In terms of location, the nucleotide sequence data produced in China leaves the confines of the laboratory and moves into the globally accessible domain of GenBank. Yet, the curator's evident concern also manifests an internal transition. No longer an expendable upstream product or institutional demand, the nucleotide sequence data attains a worth in its own right while assuming a new valence. Sequence data is obviously produced by sequencing machines but it is only *made* once it becomes part of a bioinformational resource.

From the curator's exasperation as well as the extensive amount of correspondence in his mail window, it becomes apparent that a considerable part of

submissions do not correspond to "the ideal information".[104] A peculiar physique emerges from the screen in the course of the curator's narration: The sequence of letters and bits of data recede from view while an entity appears that is still too indeterminate to sit comfortably amongst its peers and that looks out to the curator for help in easing into its new habitat. "These guys", the curator calls the sequences that await his attention: One "guy" is not long enough. Another needs to be grouped with "these other guys". Yet another "guy" seems to have duplicates. This emphasises the impression that he is dealing with *patients* while also animating the data on screen. In the quote above, the "guy" has *things* that remain undefined and not properly located. Despite the inordinate and certainly not very humane number of sequences passing through his screen, the triager does not revert to machinic or industrial metaphors. Instead, he *treats* the ones that catch him in his tracks as ailing "guys". Anthropomorphising the string of As, Cs, Ts and Gs, the curator's narration suggests that a DNA or RNA sequence has proclivities and makes demands while also suffering from ostensible shortcomings. Anthropomorphism can do away "with ontologically distinct categories of beings (…) but with variously composed materialities that form confederations" (Bennett 2010, p.99). Curators tease out the practicalities in the course of which scales and ontological boundaries become re-arranged (see below for a "huge louse"). Here, certainly, the "categories of beings" refuse any conspicuous order. Instead, the curatorial gaze and its gestures dissolve scales and ontological boundaries by holding very different entities in a common diagnostic topography.

The diagnostic visions deployed by curators in triage – the first looks, the BLAST searches, the interpretative perusal of communications – constitute global, haptic and caring ways of seeing. On one hand, browsing through the submission in triage, curators "read" their story, its shape and form tells them something about the generation of the sequences. On the other hand, the triage process itself enacts a story or an encounter with a narrative. Here, the submission is no longer treated as a discrete object but the act of looking renders the context and contours of the data on screen. The effortless move

[104] "Ideal information" here means both comprehensive coverage (every thing should have its own definition) and precise mapping ("where each of those things is"), all required to make these sequences into useful and relevant constituents of the sequence universe.

between windows, programs and tools is also an effortless move between different kinds of data as well as different, texts and textures, orders of intelligibility, and ways of looking. At that stage, the object at the centre of these moves, the sequence submission, is not quite well enough to be admitted to the database. The next sections examine how sequence submissions are looked after once they have become database records and entered the sequence universe. Again, it demonstrates that the kind of looking entailed by curation enrols intellectual, affective, creative and corporal facilities.

From sludge to scaffold: discerning differences

> As a matter of routine, each specimen as it is obtained in the field is at once tagged, the label being inscribed in India ink with the exact place of capture, date, collector and field number. The original field number is the same as that under which the animal is at the same time recorded in the field notes.
>
> Joseph Grinnell (1968[1910])

Grinnell was very exacting in documenting the methods by which museum specimens were to be created. Separate-leaf notebooks were to contain records of observations ("with carbon ink" he specified). These were to comprise details about floral surroundings as well as animal behaviour. Once obtained, specimens then entered an elaborate system of cataloguing which involved three sets of cards that accounted for the specimen's relations with its *in vivo* extraction site, its place within the museum and its purpose amidst the body of zoological knowledge. Thus, the specimen became entangled in an intricate world of materials whose composition and textures were of great curatorial concern as they contributed to fixing the specimen *qua* specimen. Curators are involved in making not only very tangible objects (e.g. specimen, cards) but also stories that connect very different places and times. In triage, curators are concerned with the nucleotide sequence in relation to its context of production. "How was it obtained?" and "What processes and materials were used?", are just some of the questions through which they *unbracket practicalities*. Taking into account also the curatorial gestures, we find ourselves in a manifestly material world not unlike Grinnell's, despite remaining in front of curators' computer screens. The following passages unwind some more layers of the sequence's multifaceted physique in the sequence universe.

I have joined a curator responsible for whole genome shotgun (WGS) submissions. A WGS submission refers to a submission whose sequences were derived via whole genome shotgun method. They account for one the most rapidly growing dataflows into GenBank. Whole shotgun sequencing allows the sequencing of long strands of DNA, entire genomes even, by shearing them into smaller fragments, sequencing the ends of those fragments and aligning them via software tools called "assemblers".[105] WGS is an accumulative process where genomes are built over a period of time. Most commonly, GenBank receives incomplete genomes (with or without annotation) in parts which are then grouped by genome project. These projects are, in turn, organised by organism, the most data rich ones being two human genomes submitted by the Beijing Genomics Institute (BGI) in 2009, a "male African individual" and a "male Asian individual". The third largest WGS project is the marine metagenome from J. Craig Venter Institute's Global Ocean Sampling project, which is discussed in chapter 7. Unlike direct submissions, WGS submissions are not triaged, requiring different workflows. Once received, a WGS curator will review the submission and contact the submitter by email. If an accession number is issued, the WGS submission will then undergo a more thorough review. What emerges in the course of observing the WGS curator deal with incoming submissions, is a very physical handle on data. Rendering WGS data intelligible and assembling genomes turns out to involve a lot of *construction work*.

Once again, I am watching the curator as she navigates through her multiple dialogue boxes and windows while trying to reconcile her commentary with what I can see on screen. We are looking at a metagenomic record from an acid mine drainage project.[106] The master record identifies the isolation source and location for the project's metagenome: "Pink biofilm microbial community collected from flowing acid mine drainage" taken from the Richmond Mine at Iron Mountain in California. From the master record, we move a level up to "Acid Mine Drainage Biofilm" (currently 1,039 nucleotide

[105] WGS was used by J. Craig Venter and colleagues at the Institute for Genomic Research (TIGR) in 1995 in sequencing the first complete genome of a living organism, the bacterium *Haemophilus influenzae* (Fleischmann et al. 1995). In contrast to genomes derived from clone-by-clone sequencing, genomes assembled via WGS are only complete in "a statistical sense" containing multiple gaps and discontinuities (Galas 2001).

[106] Acid mine drainage, the flow of sulfuric acid into ground and surface water from mines, has proven a fertile environment for metagenomic analysis (Tringe et al. 2005).

sequences and 2,543 protein sequences) and from there to the umbrella project, "Iron Mountain Acid Mine Drainage Project" (currently, 2,672 nucleotide sequences and 11,004 protein sequences). This is indeed a very different environment from the discrete records generated by direct submissions. Here, nucleotide sequences are nestled within projects, sub-projects and umbrella projects, and we are accordingly switching between GenBank views and the NCBI's BioProject pages, a collection of biological data from single initiatives or organisations.[107] The curator brings up the list of nucleotide sequences associated with the project and notes that "from acid mine drainage they were able to assemble it into these different organisms." (GB10) She says this while demonstratively scrolling through the project's myriads of nucleotide sequence accessions (almost 4,000). I might as well be looking at the sulphuric sludge itself for I find it difficult to make any sense of the enigmatic record titles, let alone behold *organisms*. Yet, from the unintelligible and supposedly hostile sludge, numerous DNA fragments have been emerging, pointing to an abundant organismal presence.[108] In order for those nucleotide pieces to arrange into organisms, considerable curatorial oversight and intervention is required because, as the curator pointed out, WGS data needs to be *assembled*.

As mentioned earlier, the WGS method sequences random chunks of DNA. This produces large amounts of raw data or base reads. In order for these environmental fragments to differentiate into individual species, they have to be processed through algorithms that identify overlaps, sections of identical sequence on different fragments. Putting together such overlaps produces a contig, an assembled read, which is no longer in pieces but forms a continuous whole. WGS projects in GenBank consist of such contigs, contiguous, overlapping segments of genomic sequence, and scaffolds, which denote assembled contigs.[109] At the end of 2011, there were over 3,400 WGS sequencing projects in GenBank, comprising more than 9 million scaffolds for genome assembly (Benson et al.

[107] See http://www.ncbi.nlm.nih.gov/bioproject.

[108] The Iron Mountain project, which pioneered environmental sequencing in the late 1990s, contains mostly bacterial sequences but there also appear to be viral, archaeal (*Thermoplasmata*), fungal and protist life present. It remains a key site for discoveries in the realm of microbial evolution, ecology and, more recently, the microbiome. For a history of discovery at Iron Mountain, see Denef et al. (2010).

[109] The human genomes submitted by the BGI contain about 5 million contigs each, JCVI's marine metagenome comprises about 4 million contigs.

2010). Because scaffolds are in the right order but not necessarily connected in one seamless stretch of sequence, they are best understood as a "gapped mosaic of assemblies" that allows genes to emerge "like the picture on a reconstructed Grecian urn" (Galas 2001). Hence, the difficulty of WGS projects lies in the assembly of sequenced fragments. All assembly, from contigs to scaffolds and entire genomes, is done using assembler programs that match the sequenced ends of the DNA fragments.[110] Although these programs append annotations, much of it requires curatorial oversight as the GenBank's submission instructions for genome submissions state: "Many genomes are annotated by automatic prediction programs and since these programs do make mistakes, it is up to all of us to try and ensure the information being presented is as accurate as possible."[111] This touches upon a very tenacious distinction within the sequence universe, the distinction between automated and manual annotation methods, and, conversely, between curated (manual) and non-curated (automated) resources. As will become evident, however, this distinction is not as clean-cut as it may first appear.

In the present case, we have progressed from sulphuric soup to mosaic fragment: Out of the Richmond mine, DNA fragments appear neatly organised as contigs in the sequence universe where each contig has its own individual database record.

> And they got their own numbers. So if you wanted to look at this *Leptospirillum*, it's 29595 which is ... in this list. We click on this and these are the scaffolds that were built from the base data. GB10

Here we meet one of the many organisms that have surfaced from the sludge by virtue of sequencing technologies and analyses. *Leptospirillum*, an iron-oxidizing bacterium, is found in biofilm of acid mine drainage but also near deep-sea hydrothermal vents.[112] Since

[110] There are two types of assembly programs: One is based on the "overlap-layout-consensus" principle, which matches each nucleotide fragment to all available fragments in order to find correspondences ("overlaps"). The other approach, more favoured nowadays as it is less prone to making errors in matching very short reads (next-generation sequencing produces very short reads, usually no longer than 75 base pairs), uses algorithms that neither match nor overlap reads. Instead it translates a layout problem into a path problem (Pevzner et al. 2001).

[111] At http://www.ncbi.nlm.nih.gov/genbank/genomesubmit.html. Manual curation is critical because automated assembly can operate with too stringent a cut-off scores, ignoring overlaps that are too small for the algorithm to take into consideration (Meyers et al. 2004).

[112] Iron bacteria such as *Leptospirillum* have been known since the early 19th century to play a significant role in the geological processes of the oxidation of iron (Emerson et al. 2010). Its occurrence in acid mining systems is of interest to scientific discovery because *Leptospirillum* is

highly acidic biofilms floating on the surface of waters flowing from mines are highly toxic to the environment (Gray 1997), the study of *Leptospirillum* is critical for evaluating the impact of mining and predict its environmental consequences. Like the iron bacteria *in vivo* pouring out from the depths of mines, the *Leptospirillum in silico* and on screen comes to light from hidden depths, buried among much else. Yet, as soon as it makes an appearance, the curator's mouseclick dis-assembles the bacterium into scaffolds and, subsequently, these scaffolds into base data.

We have moved on to the next record. "This", she says pointing at the most recent submission, "is *Pediculus humanus*, it's human body louse and JCVI [J. Craig Venter Institute] has submitted the sequence for this." (GB10) Her explanation continues: "They sent me an email and they said they've loaded this file and it has the annotation. This record is already out. Let's see." We move from the curator's Sequin program to the web browser and the Entrez gateway (see chapter 4). She types "pediculus humanus" into the search box, and we instantly find ourselves looking at the master record, which like all WGS master records, features some publications, contigs and scaffolds. The current submission is in fact an update for some scaffolds that lack annotation. Now the submitters "want to add genes and CDSs [coding regions] and stuff." (GB10) She moves from the master record view in the web browser back to her desktop where instead of the familiar Sequin she brings up a less graphical interface.

> GB10: I have this directory which is just for Pediculus and we're on version 5 now. Because that's how many rounds of files I've gotten from them.
>
> TN: I see you're working on the command line?
>
> GB10: Yeah, Pediculus is so huge, there's no way… you just can't do anything else with it. How many files are there? [scrolls] So, this thing has 510 annotated scaffolds and each one is in its own file. All together it has over 1,000 scaffolds but they're only annotating 510 of them. And there's tens of thousands of contigs that these scaffolds are made up from. And so when I get a file that comes in as a tar-zipped [compressed file format] file I have to uncompress it and open it up and what I got from them is this. Each one of these is a directory and inside... It represents

---

recognised as a fundamental element in industrial extraction and conversion of metals. At the same time, *Leptospirillum* is considered to control production of acid mine drainage.

> one scaffold and inside the directory is a table file because that's how their software spits it out. And I take all their table files and... Where do I want to go? … [scrolls and types] And I put them all together in one directory. [types] I call it "tables" ...and from that one directory I can run my command-line stuff. GB10

A cascade of folders unravels while her narration mirrors the relentless multiplication of directories and files and folders. It seems that in assembling a genome there is neither time nor space for full stops or aesthetics. Instead, the repetition of "and" turns sentences into a torrent. It is hard to keep up, watching her actions on screen and following her narration of it. Things on screen appear and disappear at the same speed. Each uttering of "and" accompanied by a click of the mouse revealing yet another dialogue box or file. Hers is a different urgency than the one I had observed during triage. In this instance, the urgency is not caused by the amount of submissions pending review but by the complexity of the louse and its project. I am not sure what follows what in her frenzied clicking, typing and scrolling but I am fascinated by the amount of room claimed by a usually diminutive creature such as the louse. Here, the louse is indeed "huge".

Dealing with genomes requires special processing paths, more specifically, "the special scripty way" (GB10). This generates files – table files, input files, validation files – whose arrangements and interactions constitute vital activities in the making of the louse genome. The error log, a summary created by executing the validation file, shows a "suspicious frame".[113] Most errors, she says, can be ignored but others cannot. "Like I want to know what this suspicious frame was, then I'd have to go and *find* it. So I would grep."[114] She locates the annotation and the actual feature which contains the error. Pausing to look at the feature she laconically concludes, "it just means the splicing isn't quite right." (GB10)[115]

[113] A reading frame refers to the way of reading DNA: It is a contiguous and non-overlapping set of nucleotide codons (triplets) and with double-stranded DNA, there are 6 possible frames. There are multiple errors associated with reading frames, from so-called frameshifts to erroneous stop codons (premature truncations) within a frame.
[114] Grep is a command-line text search utility that allows the curator to search for the frame within the submitted files.
[115] Splicing refers to a molecular interaction where introns are removed and exons joined for the production of proteins. It is a key interaction for functional regulation.

Watching her curate this way certainly provides for a physical journey through the scaffolds, folders and alignments of the louse genome. What strikes me when talking with curators and observing them is how seamlessly the narrative of their work and the stories behind the on-screen data weave in and out of the digital, in and out of macro and micro worlds, oscillating between the record and the organism. The film critic Laura Marks (2002) writes of "haptic visuality" in describing a way "to explore the relationships of present, absent and remembered bodies" (Ahmed & Stacey 2001, p.6). This, I suggest, can be used to understand the way in which the curator's gaze and actions coherently grasp a multitude of ontologically very different entities, some of which are right here (the JCVI's email), some of which were never there and are long gone (the louse that had provided the DNA in the first place) while others (contigs, gaps and scaffolds) have a less determinable position. Also, we can see the many kinds of activities the curatorial process demands of the curator: running scripts, finding errors, generating files, locating and clearing routes, assembling genomes. And the many kinds of entities aside from files that curators handle: contigs, scaffolds, emails, frames, human lice, directories, scripts. Reaching in and out of the screen, curation here is most definitely an immersive activity, not the detached and automated routine which "data work" might suggest.

In following WGS curation, it is indeed difficult to clearly demarcate the *manual* and the *automated* parts of this process. This distinction comes to bear on the more general perception of the biosciences. With the proliferation of bioinformatic tools and techniques, organisational and epistemic stratifications within bioscientific research were seen to increase. Science studies have demonstrated the emerging boundaries between scientists and (information system) developers (Star & Ruhleder 1996) and between theoretical biologists and biological scientists working in wet labs (Fujimura & Fortun 1996). In these studies, the sequence database exacerbates the division of labour in biology while also precipitating larger structural changes in the world beyond: how we think about disease, medical institutions, health research, the nature of public and private funding. But curatorial practice at GenBank emerges as an entanglement of human, manual and automated elements. The next section will add to this entanglement by attend to the seemingly "technical" work carried out by developers at GenBank.

Maintenance work: Plumbing and traffic

The sourball of every revolution: after the revolution, who's going to pick up the garbage on Monday morning.[116]

Mierle Laderman Ukeles (1969)

Expressions such as "data dump", "data flood", "data stream", "deluge of data", "data torrent" (St. Pierre & McQuilton 2009) and "data revolution" (Lord & MacDonald 2003) hold a visceral imaginary of data not only as potential danger but waste. Such terms suggest that the abundance of data can quickly turn into a burden requiring containment and management. "Who's going to pick up the garbage after the revolution?" asked the artist Mierle Laderman Ukeles (b. 1939) in her 1969 manifesto which summarises the ethically urgent and politically relevant rationale of her oeuvre: the recognition and representation of the myriad of silent, humble and invisible labours that keep cities alive.[117] In Ukele's practice the city emerges as an ecology, a dense and vibrant network of trucks, rats, pipes and sanitation workers. As an artist she plays with the notions of visibility and invisibility: Cladding a garbage truck in mirrors, passers-by would see their reflections on the sides of the container which ferried their waste away from sight. Given the spectre of data as waste, the databases offer a suitable milieu for reflection upon similar issues. The 'omics revolution, the impact of high-throughput genomic technologies, would certainly not have been possible without the continued efforts of curators, developers and engineers entrusted with sorting, cleaning and accommodating the proliferating data dumps (Field et al. 2009).

The WGS curator notes that master records make "it easier to *maintain* the data" (GB10, emphasis added). With regards to the kinds of work carried out in EMBL-Bank and GenBank, *maintenance* emerges as a central practice. Maintenance, like care, implies a corollary of steady and modest activities which usually happen away from our field of vision, beneath surfaces and less tangible thresholds of perception, oftentimes away from

[116] From "Manifesto for Maintenance Art 1969! Proposal for an exhibition 'CARE'". Available at www.feldmangallery.com/media/pdfs/Ukeles_MANIFESTO.pdf. Last accessed: 8 March 2011.

[117] For 30 years, from 1969-1999, Ukeles was the (unsalaried) artist-in-residence at the New York City Department for Sanitation, a position she had created for herself. She refers to herself as a "maintenance artist" and her work is concerned with rendering visible the mundane yet vital activities such as cleaning that keep infrastructures like cities or museums working.

formal recognition. This is to say that maintenance work often is *invisible* work, work that is not seen to be productive or innovative. Importantly, maintenance intimates physical labour. Yet, while the account of WGS data provisions feature decidedly hands-on aspects, they also stress productive and innovative moments: Assembling and aligning contigs, constructing scaffolds and, effectively, building whole genomes around gaps give the impression of an intricate apparatus of maintenance. Notions such as "data pipelines" and "high-throughput" data heighten the sense that data, much like other raw materials, are entangled within multiple visible and invisible infrastructures. The phrase "blasting a sequence against the database" also evokes a distinctly mechanical process while participants' continuous reference to the databases' "production activities" emphasises the industrial nature of the enterprise. Even curators refer to their work as foremost a "production job" (GB16).[118]

There is indeed a great emphasis on production activities. For engineers and developers this means primarily the production of files as well as other deliverables (such as enhanced search facilities) to the curators and the scientific community. The delivering of "sequence data products that go out to the public every day" (GB14) means that the system and workflows have to reliably facilitate a continuous flow. The biggest product is the entire release, that is, the total data volume of the databases. This means "pulling things together from a variety of different types of systems and packaging it and spitting it out the door" (GB14). GenBank releases are made available once a month: Release 184 containing over 140 million sequences was made available 19 July 2011, Release 185, containing over 207 million sequences followed on 18 August 2011. The entire release, made up of the totality of flat files, represents a considerable part of the sequence universe in a nutshell and is obtainable for anyone independent of retrieval tools like Entrez. Users obtaining such releases then run their own analysis programs to process data on their own sites. These could be pharmaceutical companies or academic institutions but also other government institutions. Once again, the narrative evokes a decidedly material

[118] Describing her work to me one curator says: "Some of it takes a little bit of literature research. Not too much because, again, this is a *production job*." (GB16, my emphasis)

world: “packaging” different things that are subsequently shot through a pipeline that ultimately spits them out as “one of my job is getting the product out of the door” (GB14).

The scientist in charge of GenBank’s infrastructure variously refers to himself as an “unsticker”, “a master of the practical kluge”, and, most aptly, a “genetic information plumber” (GB11). Calling himself a plumber, he paints a picture of sequence data as a discharge or residual that continuously veers between excess and impasse. Given the prospect of surges and deluges, plumbing would suggest itself as a useful analogy in managing such data flows. The spectre of surges returns when, in answer to my question about future challenges brought about by next-generation sequencing technologies, he retorts “how do you drink from a firehose?”. Aquatic metaphors are never far when discussing the role of data in genomic sciences.[119] Although post-genomic research continues its data output, the difference between “data dumping” and “data mining” has become much more pronounced (Huttenhower & Hofmann 2010). Naturally, the databases are crucial mediators in this transformation from dump to resource. The work of curators, developers and engineers at EMBL-Bank and GenBank represents an integral part of bestowing intelligibility upon this “dump” to facilitate any meaningful mining. Plumbing, as Liss observes in her text on Ukeles’ work, can also be seen as a form of maintenance labour, which is “fundamentally about nurturing and maintaining natural and psychic life systems in all their detritus” (Liss 2009, p.44). The valorisation of data waste is certainly advanced by ever more sophisticated statistical models and resulting sets of algorithms (see, for example, tools such as Velvet, AllPaths, ABySS) but it is equally sustained by the “nurturing and maintaining” and other supposedly mundane practices that keep the databases and their contents alive and, as another curator put it, “vibrant” (GB7).

Unlike (most) curators, the section chief of infrastructure and genetic information plumber has his own office. I meet him there and am taken by the amount of stuff covering every square inch. There are stacks of paper, boxes, preserving jars and plants. I spot a microwave underneath his desk and a set of strangely shaped Tupperware containers.

[119] Though more recently it appears that they, wary perhaps of the careless promises that had accompanied the early days, have adopted a more cautious handling of number (see for example the *1,000 Genome Project*).

These, I soon find out, are his invention: Based on a trapezoidal lid that could be inserted into the container, they can be nested and stored with greater ease than the traditional rectangular boxes. Another patent he holds was for an *in vitro* mutagenesis technique. "If I have something unique to contribute", he tells me, "it's how to do something from scratch where nothing exists." Before joining the NCBI he had worked at Los Alamos in the Theoretical Biology department, the birthplace of GenBank and, more generally, computational biology (see chapter 1). Like others at GenBank, he is part of the first generation of scientists *doing* bioinformatics and much of this pioneering spirit still defines the technical solutions and systems that rely on kernels devised more than 20 years ago. Back then there were no ready technical provisions. This is why GB11 and his contemporaries had to build codes, softwares and machines concurrently with asking questions about gene expression or the speciation of T4-infected *E. coli*. For his PhD, GB11 had turned his department's word processor, which was the only "computer" in the department and was exclusively reserved for use by typists, into a "mini computer" to produce the analyses required for his thesis.

In fact, most of the underlying technical system at GenBank is somewhat accumulative, bearing traces or in some cases thoroughly relying on "kludgy" solutions that had been built many years ago. Discussing the Smart tool, the software that facilitates the handling of submissions and accompanying correspondence, one of the developers light-heartedly admits that none knows what "Smart", an acronym, stands for anymore. Talking about one of the early tools he had developed for the curation group, GB11 describes it as follows:

> GB11: Ugly, *ugly* solution but it was required for productivity.
>
> TN: When you say "ugly", do you mean somewhat "clunky"?
>
> GB11: Yes, I would say clunky. At least a little clunky. I'd say kludgyness was more from the maintainability and the artificial and the sense of plumbing of the mismatched pipes I had to put together to get things to flow. So I guess it was clunky underneath the *hood* but the indexers didn't mind too much.

Peering under the hood of GenBank reveals a sight/site that is not unlike the one found

just below the surface of most modern cities: an assemblage of "mismatched", that is, historically and functionally divergent, infrastructural pieces such as pipes, tunnels, walls, cables and reservoirs. It appears that as with all infrastructure, database systems too can manifest their history. Recounting the development of the first editing tool for curators, GB11 tells me that they "were using a SYBASE product that the SYBASE consultant couldn't figure out how to make into an editor" (GB11). SYBASE was one of the first high-performance relational database management systems for online applications, that is, ordering and making available data across computers on a network. It was first released in 1987 and, for the first time, offered a client/server architecture rather than storing all data in a central mainframe computer. Since the early 1990s, the SYBASE system has provided the database infrastructure to the Human Genome Project and the National Center for Genome Resources. However, it required customisation as well as creative plumbing given that "sequences aren't something that one can handle in a relational database".[120] "So I had to", GB11 tells me, "think outside the box".

GB11's reference to "kludgyness" certainly suggests that within the tangle of pipes, one can discern instances of individuals' skills and styles. Enquiring about idiosyncrasies in the programming of GenBank systems and applications, one programmer replies, "everyone has different tastes and different approaches and different styles." (GB15) This means that "you can actually recognise people from their writing styles." Different parts of the system bear traces of different people – some of whom might be long gone.

Developers are located on the same floor as curators, some in open plan offices, others (the more senior developers) in separate office. Exchange between curators and developers is facilitated by regular meetings but oftentimes happens on an *ad hoc* basis where curators, encountering a specific issue, would call on one of the developers. One programmer tells me about the development of new functions:

[120] As we will see in the next chapter, the database is made up of flat files. Here, the data type (the attribute like "Source" or "Features") is part of the data. Flat files also contain many fields with a very large amount of duplicate data. In addition, for flat files, the order of records matters though records can be hierarchical and contain multiple sub-records. Importantly, flat file databases are designed around a single table, hence changing a record will not have any bearing on any other record.

> We want to do it really fast, really quick so users can take advantage of the new functions. We have a short cycle when we get a request to do such a new additional function, then we develop it and release a test version for them and the cycle repeats. We don't spend half a year preparing a new version like commercial software would do. (...) We update our versions on a weekly basis. Even more often. (GB15)

"Here", he says, "we're constantly in [the] developing cycle." (GB15) Like triage and WGS curation, programming too is accompanied by a sense of urgency. While this highlights the scope for kludgy intervention, it also opens another avenue to explore the inventive potential carried by maintenance work. From GenBank's programming perspective, development and maintenance work happen concurrently. Here, routine can quickly turn into innovation.

Routing traffic and making the database forget

Responsible for ensuring the continuous stream of data through and out the various systems, the dataflow team coordinates the traffic between sites (e.g. researchers' laboratories, databases, users' terminals) as well as between conditions – from data to records to packages and products. Here plumbing is not "just" metaphor but programme. I have joined a dataflow programmer (GB14) who is about to tackle an issue from his queue, the list of requests and queries received from curators, developers and other programmers at NCBI. The one we're looking at illustrates as he calls it a "nuts and bolts issue", a "showstopper for getting sequence database loaded into" the database (GB14). It is a request put forward by RefSeq curators, the group that handles reference versions of archival GenBank records. The showstopper issue GB14 is currently tackling requires him to generate an exception, or put differently, to break a rule. Specifically, the RefSeq group has asked for a connection to be severed between a RefSeq record and an archival GenBank record. Here, as GB14 comments, "our system says 'no, you can't do that'". In order to work around this categorical *No*, he will "have to go in and do a series of update statements to the databases". We look at the screen where we find ourselves "on the level of the database server".

> So here I'm interacting with the SQL [a standard language for accessing and managing data in relational database management systems] server to ask essentially "show me one of the example records from the RefSeq group to which they assigned an original accession and they have a related identifier that ultimately comes from the archival GenBank record. Show me what that looks like." (GB14)

As he is asking the database to "show him" things, addressing his screen as if it were some kind of oracle (which, incidentally, is the name of a popular database management system) he writes the Sequel commands which query the server. Once again, I am looking at a command-line interface similar to the one encountered during WGS curation. This time, however, the entity that is being worked on is the system itself. Instead of assembling or transforming the database record in such a way that it would fit the system, it is the system which is being manipulated to accommodate a particular RefSeq record. I use the word "system" here with hesitation because it would suggest a stable, static and coherent technical infrastructure which, in this moment, is not what we are engaging with. What does this say about the system's strive for stability and closure – a central dynamic in explaining how large technical systems work and endure? Continuous routine interventions, innovations and improvisations are required not just to "keep it going" but are, as we will see in chapter 7, solicited from the users in engaging with the database. Stability and durability emerge as outcomes of many precarious practices, of "mismatched pipes" and variant visions.

Having brought up the RefSeq records in question, the dataflow programmer resumes:

> "What am I gonna have to do?" I'm gonna have to identify these secondary accessions, basically kill them so they can be re-used in the context of a different RefSeq accession. I'm essentially creating a little list for this exceptional case. I wanna run a little Sequel against the database see what the associated identifier is, tell the database to forget about it: "This wasn't really here, you didn't see that pairing before." GB14

He improvises a ruse that tricks the database into “forgetting”. The terms of engagement here do not fit with the command line interface that the programmer is working on. He is,

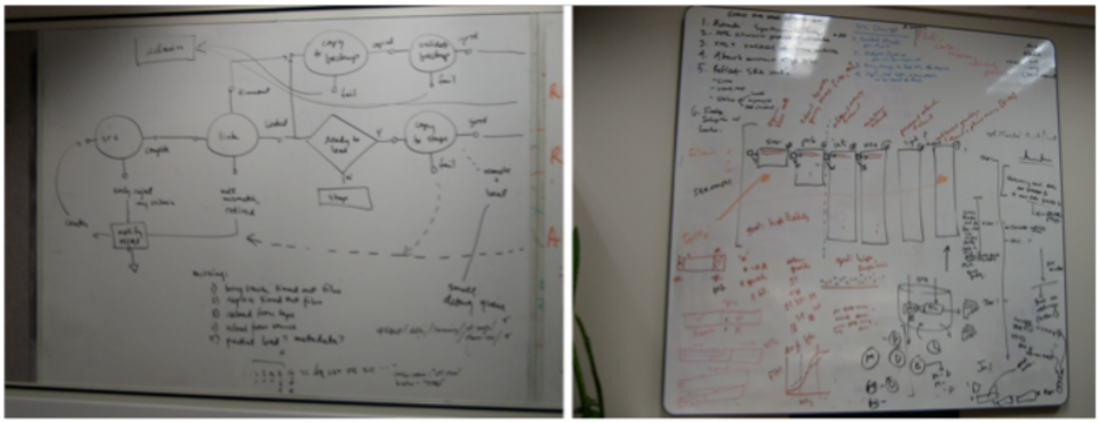

Figure 4: Flow and structure charts at GenBank

of course, typing Unix commands but the narration of his activities betray a different order and, in turn, a different kind of system. Depending on the kind of work required, the vision of the database system changes. Dealing with an exception like the RefSeq record prompts a figuration of the database as a sentient if somewhat naive entity that is responsive to tricks and negotiations. In managing data flows, however, another figuration takes centre stage, that of “traffic”.

A programmers reflects on the data pipelines passing through GenBank:

> It's like we have different highways for all different data flows. It's not that we put all the small cars and motorcycles and big trucks on the same road. We have a separate road for them. (GB15)

He tells me this as we gaze at his whiteboard. Most of the cubicles at GenBank feature whiteboards and so do the offices at EMBL-Bank. They are clearly used a lot, bearing diagrams, writings, symbols and less discernable entities. I notice these boards because gazing at computer screens all day, they strike me as somewhat out of sorts. Whiteboards conspicuously illustrate the experimental side of curation, attesting to the fact that processing of data, ensuring the “data flow” requires thinking, testing, problem solving, designing, communication and learning. I make a habit of photographing the whiteboards in offices and corridors, all the while asking my respondents to explain what they depict. To me they offer a novel vision of the databases whose actual physical location as a

technical system in the shape of servers continues to elude my respondents and me. Together with a database developer I am studying his whiteboard (Figure 4). It features what he calls a "structural chart", a chart that shows the components of the Smart system. Gazing at this chart, he muses:

> You want the system to do certain things...well, I always think of big systems especially like a living thing. I give it something, it's like input... I feed it something and it digests it and makes certain good things and there is also some garbage it produces. Or some temporary files that need to be cleaned or deleted. For me every program is a small living organism and big system is a collection of small living organisms in a symbiotic way…in symbiosis. They help each other do different things. GB15

This returns us to the idea of infrastructure as life systems that emerges in Ukeles' work. Listening to GB15's explanation it becomes clear that the chart we are looking at can, at times, not be an abstraction. It does not perform the disembodied and representative iteration of a phenomenon. This is concerned with the representation of relations between elements rather than a representation of an "outside" world. This re-works the relationship between technical network, environment, curator, model organism and bioinformational artefact. "Without vision", one engineer at EMBL-Bank notes, "it [the database] would be nothing" (EB1). Here, vision is a means of enrolling and shaping capacities to manage the present while also being orientated towards the future.

Looking around GB11's office, the thought occurred to me that this seemingly haphazard collection of items and their innocuous arrangement might be the closest I have come to a representation of the underlying database system. Rather than the diagrams of flow charts and structural charts I had spotted on whiteboards in the offices of engineers and developers, GB11's workshop landscape seemed a much more accurate model for the way in which the database system actually *looked like*. Databases have histories although admittedly their traces might be more difficult to perceive. I ask him about the future challenges and he tells me that he is excited about "getting his hands dirty" in the efforts to connect genotypic with phenotypic data. There is a pioneering spirit and sense of adventure and enthusiastic readiness in his accounts.

Partial visions, cosmic landscapes

> A man cannot afford to be a naturalist, to look at Nature directly, but only with the side of his eye. He must look through and beyond her. To look at her (…) turns the man of science into stone. I feel that I am dissipated by so many observations.
>
> Thoreau 2009, p.186

I take Thoreau's caution not to mean that nature exerts a quasi-divine awe but that we have to adopt situated and partial vision for comprehending the resonant fields of more-than-human life (Manning 2010). This appears a self-defeating stance because once we situate our vision amongst the many devices and narratives which shape our sensory awareness of nature, "nature" as such disappears from view altogether. There is, however, no loss implied in that. On the contrary, Thoreau's feeling "dissipated" has a decidedly exalted ring to it. Relinquishing the apical position, here the viewer is never separate from the world.[121] Curation, I would argue, constitutes a practice that encompasses multiple kinds of vision that equally dissipate objects, subjects and environments.

At first sight, the differences between this and Grinnellian curation appear obvious: In addition to the lack of specimens, curators do not venture out into "the wild" much, their work being based almost exclusively in offices. Yet looking closer, parallels emerge – not just when relating the required personal dispositions mentioned by Grinnell but to the more formal concerns and kinds of work being undertaken. Particularly his concern to capture and preserve information around the specimen, claiming that it is this information that might surpass the specimen in importance, appears prescient from today's perspective. The information assembled around sequence data, its annotation, does indeed ensure its longevity and value within the sequence universe and beyond.[122]

Curation at EMBL-Bank and GenBank, as I have detailed in this chapter, involves a corollary of activities, from the moment a submission is received to the moment it becomes part of the monthly release. It is a heterogeneous process that involves many different practices, objects and knowledges. Though biocuration *per se* constitutes a novel field of scientific activity, it can be seen in relation to traditions of curating museum

[121] Such an interpretation could perhaps also account for Thoreau's profuse dislike of his surveying duties.
[122] The work of annotation is discussed in detail in the next chapter.

specimens that developed in the late 19th century. Contextualising biocuration through the work of Grinnell allows for continuity, that is, for understanding biocuration within a history of collecting and looking after natural objects. The Grinnellian method was concerned with assembling an environment for the specimen that would ensure its longevity as well as its relevance not just to its "natural" associates but also to new ones (such as ecologists and systematists but also museum visitors). This kind of curation still very much defines the work carried out at EMBL-Bank and GenBank. Grinnell's description of curation suggests that the processes of recording the specimen and its environment and of writing labels and accounts contribute to the process of actually preserving the wildness. This might at first sound incongruous: Does not the labelling, classification and identification of natural specimen contribute to their taming, to reigning in this wildness? We encounter a similar argument in accounts that conflate the rise of data in the biosciences with a purging of vitality and a "flattened" epistemology (N. Rose 2007). Yet, examining the work of biocurators suggests that the tropes of data and information are anything but flat. It points to a kind of "imaginative domestication" (Cosgrove, 2008) that does not necessarily contain and restrain "wilderness" but transplants it.

Curating encompasses multiple visions. Here, "looking" and "seeing" do not constitute hands-off, detached activities but very much involve haptic immersion in virtual, molecular, organisational, textual and organismal environments. "Looking after" the sequence universe comprises many other senses and sensors as it becomes materialised in different assemblages involving technologies, people, objects and their affective and affiliative associations. At the same time, the curatorial activities themselves materialise things, that is to say, curation deploys performative visions that assembles entities such as "contigs". The sequence universe can afford visions that are more partial than the normative visual orderings usually associated with geneticization and more generally bioinformatic technologies. As motorcycles and trucks, WGS data, lice, correspondence, sludge and coding regions flow through mismatched pipes they leave, in the wake of their traffic, far fewer categorical distinctions between matters. Mackenzie (2003b) argues that "[i]magining effectively connects bodies" while "it leaves that

connection in question" (2003b, p.369). This describes the work of vision, the reflective and performative role that imagination plays in the everyday routine interactions with technical infrastructure. Imagination is linked to intelligibility, establishing a tangible purchase on matters that might otherwise remain incomprehensible.

Lastly, curation betrays a false opposition between "manual" and "automated" annotation. Ewan Birney, founder of the Ensembl genome browser, remarked on different annotation methods: "An aside. I hate using the words "automated" and "manual" for these two processes." (2008) Stressing the importance of human intervention in the gene build (essential part of automated tools) as well as the use of bioinformatic tools in manual curation, Birney in actual fact argues that annotation is a *socio-technical process*. Hence, arguments that privilege either one of the constituents – the technology or the curator – fail to understand the basic nature of annotation, namely its dependence on human-more-than-human-machine interaction.

Chapter 6. Between dung cannons and the deep blue sea: reading the record and assembling a bioinformational artefact

In this chapter I turn my attention to the content of the databases, the database record. Each record documents a nucleotide sequence, which is provided in full at the bottom of each record. It is preceded by a host of additional information that describes and qualifies the sequence. This complimentary data described source organism, author details and publication references while also identifying relevant features on the sequence such as genes and proteins. This chapter examines two specific database records: one derived from EMBL-Bank documenting a sequence from the fungus *Pilobolus crystallinus*, the other derived from GenBank and recording an uncultured bacterium discovered in the course of Craig Venter's Global Ocean Sampling (GOS) project. In *reading* the records, I examine the different data elements and use them as routes for exploring the records' constituent elements, most of which lay beyond their immediate confines and point to disparate times and spaces. Doing so, I assemble the records' material, discursive and political lives. Here, the record emerges as a bioinformational artefact that comes into being through a *cumulative relationality*.

Introduction

When historian of science Charles Weiner, in an interview with legendary physicist Richard Feynman, intimated that Feynman's notes represent "a record of the day-to-day work", Feynman adamantly refuted this by responding: "No, it's not a *record*, not really. It's *working*. You have to work on paper, and this is the paper. Okay?" (Gleick 1994, p.409) Feynman's indignant correction suggests that above serving any representational or documentary function, his notes *are his practice*, they *are Feynman's work* and should therefore not be seen as attendant to or separate from his thinking. Similarly, the database records contained within EMBL-Bank and GenBank are more than dormant traces of

sequencing efforts, more than representations of particular strings of nucleic acid molecules. They are, like Feynman contends, *working* and they do so continuously: Queried by search functions and processed by algorithms, they make connections across databases, molecular pathways and species while oftentimes causing troubles (see chapter 7) or surprising encounters (see chapter 4). As the previous chapter has shown, database records themselves demand *work*. Scientific efforts continue to be directed at sequence data even after it has fulfilled its task for whoever generated it. Curators check and investigate it, technical staff build appropriate habitats for it and institutions like the NCBI and the EBI invest in ways to make it more useable, sustainable and intelligible. Thus, database records constantly engage people, networks, institutions, hypotheses, viruses and other more-than-humans.

In April 2011, EMBL-Bank contained 199,575,971, GenBank 132,015,054 records. Each record accounts for a single contiguous DNA or RNA sequence and contains contextual information about this sequence called "annotation". Records here are digital files, so-called "flat files" (see Figures 5 and 6), which can be accessed in different "views" (see below). Flat files are unstructured text documents – they are devoid of any structural mark-up (as opposed to the Word file to which I am writing this chapter, which has headings, different font sizes and styles for specific elements of text).[123] These flat files form the most basic building block on which all other structures of the sequence universe rests. The flat file contains no inherent information about the data and requires a degree of expertise to interpret it. Data inside the flat file are organised by appending textual markers to data elements: "DE" (for EMBL-Bank records) or "Definition" (for GenBank

[123] Flat files provide for very simple data handling while also allowing it to be easily processed by bespoke computer programs.[123] It was recognised early on that sequence data did not lend itself to any of the established database structures (relational, hierarchical or network). As Greg Hamm, the first staff member at EMBL Data Library, described, sequence data was "different" and did not fit the "table view of the world" (García-Sancho 2011, p.89): "The aim of the operator with these records was not only to establish connections between sequence entries, but also to find patterns in the strings and to attribute to them certain features; e.g., the presence of a gene within the sequence." (ibid.)

```
ID   FJ536284; SV 1; linear; mRNA; STD; FUN; 1869 BP.
XX
AC   FJ536284;
XX
DT   20-JAN-2009 (Rel. 99, Created)
DT   27-SEP-2009 (Rel. 102, Last updated, Version 2)
XX
DE   Pilobolus crystallinus putative blue-light photoreceptor PCMADA1 mRNA,
DE   complete cds.
XX
KW   .
XX
OS   Pilobolus crystallinus
OC   Eukaryota; Fungi; Fungi incertae sedis; Basal fungal lineages;
OC   Mucoromycotina; Mucorales; Pilobolaceae; Pilobolus.
XX
RN   [1]
RP   1-1869
RX   AGRICOLA; IND44262955.
RX   DOI; 10.1007/s10267-009-0496-y.
RA   Kubo H.;
RT   "Isolation of madA homologs in Pilobolus crystallinus";
RL   Mycoscience 50(5):400-406(2009).
XX
RN   [2]
RP   1-1869
RA   Kubo H.;
RT   ;
RL   Submitted (10-DEC-2008) to the EMBL/GenBank/DDBJ databases.
RL   Biology, Shinshu University, Asahi 3-1-1, Matsumoto 390-8621, Japan
XX
DR   StrainInfo; 483610; 0.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..1869
FT                   /organism="Pilobolus crystallinus"
FT                   /strain="NBRC 8561"
FT                   /mol_type="mRNA"
FT                   /db_xref="taxon:369761"
FT   CDS             1..1869
FT                   /codon_start=1
FT                   /product="putative blue-light photoreceptor PCMADA1"
FT                   /db_xref="GOA:B8YIE3"
FT                   /db_xref="InterPro:IPR000014"
FT                   /db_xref="InterPro:IPR000679"
FT                   /db_xref="InterPro:IPR000700"
FT                   /db_xref="InterPro:IPR001610"
FT                   /db_xref="InterPro:IPR013088"
FT                   /db_xref="InterPro:IPR013655"
FT                   /db_xref="UniProtKB/TrEMBL:B8YIE3"
FT                   /protein_id="ACL81171.1"
FT                   /translation="MTGTSHMDQLMVHEPSAIELSPVDEAAGSGEPLTGVYSSSGFDMV
FT                   GVLSRLVNRPHPQINLGPIDMSCSFLVTDARQYDCPIVYCSPTFEHLTGYHANEIVGRN
FT                   CRFLQAPDGQVTCGSRRTYTDNQAVFHLKAQMLQNKEHQASIINYRKGGQPFVNLITVI
FT                   PITNDNNEVAFFVGLQVDLVEQPNAILEKMKDGTYIVNYQQMNIPPYIPGSSFSSEPVD
FT                   DYFRELPTNTPACSTLASPEILELVSCAGDNEQQLQQEWNKLLLDQSEDFIHVLSLKGF
FT                   FLYSSRSSSHLLEHDPEELVGHPLSSICHPSDIVPVMREVKEAASHPDRVVNLIYRVRR
FT                   KYSGYMWMECQGKIHVDQSKGRKCLILAGRERPMYELVRKEIVQAGEITQGPEFWTKAT
FT                   LSGLFLHVTPSSEEVVGSTADMLEGSTIYQYVGDNNVRDITRALELVKEGRIVNLHHTM
FT                   QNNKGDYIPVFSTFYPGDVSFGVGRPSFALIQIRSKESSQPTSDVLYEAMSDVPSENDE
FT                   NIFAELETVRGTSWQYELHQLQLANRKLKEQLESLNNPKRRKQKKKKMADTVDMPKMCA
FT                   QCQRVDSPEWRKGPNGPKELCNACGLRYAKSLANLLKHKSTDSVLK"
XX
```

Figure 5: Flat file for EMBL-Bank record FJ536284 Pilobolus crystallinus putative blue-light photoreceptor PCMADA1 mRNA, complete cds

Nucleotide

Alphabet of Life

Display Settings: GenBank

Uncultured bacterium clone 6C233420 16S ribosomal RNA gene, partial sequence

GenBank: EU805409.1

FASTA Graphics

Go to:

```
LOCUS       EU805409               1491 bp    DNA     linear   ENV 10-JUL-2009
DEFINITION  Uncultured bacterium clone 6C233420 16S ribosomal RNA gene, partial
            sequence.
ACCESSION   EU805409
VERSION     EU805409.1  GI:190710334
DBLINK      Project: 39291
KEYWORDS    ENV.
SOURCE      uncultured bacterium
  ORGANISM  uncultured bacterium
            Bacteria; environmental samples.
REFERENCE   1  (bases 1 to 1491)
  AUTHORS   Shaw,A.K., Halpern,A.L., Beeson,K., Tran,B., Venter,J.C. and
            Martiny,J.B.
  TITLE     It's all relative: ranking the diversity of aquatic bacterial
            communities
  JOURNAL   Environ. Microbiol. 10 (9), 2200-2210 (2008)
   PUBMED   18637951
REFERENCE   2  (bases 1 to 1491)
  AUTHORS   Shaw,A.K., Halpern,A.L., Beeson,K., Tran,B., Venter,J.C. and
            Martiny,J.B.H.
  TITLE     Direct Submission
  JOURNAL   Submitted (06-JUN-2008) J. Craig Venter Institute, 9704 Medical
            Center Drive, Rockville, MD 20850, USA
COMMENT     ##Metadata-START##
            depth                                 :: 2431 meters
            salinity                              :: 32.3 ppt
            temperature                           :: 29.3 C
            sampling site monthly chlorophyll level :: 0.33 mg/kL
            sampling site yearly chlorophyll level  :: 0.28 +/- 0.02 mg/kL
            hi_filter_size                        :: 0.8 microns
            lo_filter_size                        :: 0.1 microns
            ##Metadata-END##
FEATURES             Location/Qualifiers
     source          1..1491
                     /organism="uncultured bacterium"
                     /mol_type="genomic DNA"
                     /isolation_source="250 miles from Panama City"
                     /db_xref="taxon:77133"
                     /clone="6C233420"
                     /environmental_sample
                     /lat_lon="6.493 N 82.904 W"
                     /collection_date="20-Jan-2004"
     rRNA            <1..>1491
                     /product="16S ribosomal RNA"
                     /note="derived by combining overlapping reads
                     ti:2066191492 and ti:2066191847 from site GS22"
ORIGIN
        1 agagtttgat catggctcag gacgaacgct ggcggtgcgc cttatgcatg caagtcgagc
       61 gaagcttttc agtagtttac tacagaaaat gactgagcgg cgaacgggtg agtaacgcgt
      121 gagcaaccta ccttagttac tgggatagcc cgaggaaact cggattaata ccggatattc
      181 ttatttaatc acatgatttt ttaaggaaag gtcagccgaa ctaagatggg ctcgcgttct
      241 atcagctagt tggtagggta atggcctacc aaggctacga cggatagctg gtctgagagg
      301 acgatcagcc acactgggac tgagacacgg cccagactcc tacgggaggc tgcagcaggg
      361 aatattgcgc aatgagcgaa agcttgacgc agcgacaccg cgtgtgggat gacggatcta
      421 ggtttgtaaa ccactttcag gagggaagaa aatgacggta cctccacaag aagccccggc
      481 caactacgtg ccagcagccg cggtaatacg taggggcga gcgttgtccg gatttattgg
      541 gcgtaaagag ctcgtaggcg gttcaacaag tcggtcgtga aagttcaggg ctcaaccctg
      601 aaatgtcgat cgatactgtt gtgactagga tacggtagag gtgagtggaa ttccgagtgt
      661 agcggtgaaa tgcgtagata ttcggaggaa caccaattgc gaaggcagct cactgggccg
      721 ctatcgacgc tgaggagcga aagctagggg agcaaacagg attagatacc ctggtagtcc
      781 tagctgtaaa cgatggatac tagacgtagg aattggatta acgatttctg tgtcgtagct
      841 aacgcgttaa gtatcccgcc tggggagtac ggtcgcaaga ctaaaactca aaggaattga
      901 cgggaccccg cacaagcggc ggagcatgcg gcttaattcg atgatacccg tagaacctta
      961 cctggacttg acatataggg aaaagttata gaaatataat gtgcattagc gccctataca
     1021 ggtggtgcat ggctgtcgtc agctcgtgtc gtgagatgtt gggttaagtc ccgcaacgag
     1081 cgcaacccct gtcctatgtt gccagcaagt aatgttgggg actcatagga gactgccggt
     1141 gataaaccgg aggaaggtgg ggacgacgtc aagtcatcat gccccttatg tccagggctg
     1201 cacgcatgct acaatggcaa gtacaacgag tcgcaatacc gcgaggtgga gcaaatctct
     1261 taaagcttgt ctcagttcgg ataggagtct gcaactcgac tccttgaagt tggagtcgct
     1321 agtaatcgca aatcagcaaa gttgcggtga atacgttctc ggggtttgta cacaccgccc
     1381 gtcaagtcac ggaagtcggc aatacccgaa gccagtggtc caaccctttt gggaggaagc
     1441 tgtcgaaggt agggtcggta actgggacta agtcataaca aggtaaccgt a
//
```

Figure 6: Flat file for GenBank record Uncultured bacterium clone 6C233420 16S ribosomal RNA gene, partial sequence

records) suggest that any data following this prefix *describes* the record. There are only very slight variations in the record template between EMBL-Bank and GenBank. GenBank, for example, has a "miscellaneous" data element (the "Misc" feature that had annoyed the triage curator in the previous chapter) which EMBL-Bank lacks, though it does provide space for miscellaneous features elsewhere. Each record tells of a DNA or RNA sequence: how it came about, who and what was involved in its generation and what its function or purpose is. But this information is not easily ascertained from a first glance. Looking at a flat file record, only a few bits of information appear to be intelligible to the untrained eye. Especially in the *Features* section (prefaced FT in the EMBL-Bank record), numbers and acronyms prevail. Aside from including the actual sequence, the string of amino acids (A, C, G, T), the record carries information about its source organism, its "author" (whoever submitted it), its appearances in other media (journal articles or other databases) and its "features". The *Features* section is where any particular "regions" (genes and proteins) contained on the sequence are noted and described: the biological function they fulfil, the interactions with molecules they engage in, whether they result from a recombination of different sequences or exhibit variations.

Non-commensurate readings

At first sight, not much of the *working* life of the record is visible. Particularly, in flat file view, the records appear decidedly lifeless, or in any case "flat" all around. How to overcome the records' eponymous flatness and account for their innate working? How to account for the records' various capacities to intervene, perform, represent, document and engage? We know from much work in science studies that a "simple pattern of rows and columns" (B. Latour 1987, p.237) encloses and obscures conventions and struggles so as to efficiently move and combine with the world and meet demands construed elsewhere (Bowker & Star 2000). This is certainly true for EMBL-Bank and GenBank records, which enfold many standards and dissensions: "Records", the NCBI Handbook tersely states, "can contradict each other." (Karsch-Mizrachi 2007) Whereas Feynman's description would invite a decidedly *praxiographic* orientation (Mol 2002) in studying the workings of database records, the NCBI's acknowledgement opens a seemingly antithetical position:

Suggesting that records stand in relation to one another, forming a body coherent enough to impart certain qualities on its constituents, it draws attention to the records' *intertextual* qualities which are not necessarily formed elsewhere (in extraneous practices, for example). Here, two records might contradict each other while they could still accurately reflect biology. In contrast to the normalising and deflationary functions usually ascribed to databases and their data, this chapter argues that the database record can be an inventive, sometimes unruly, entity that mediates between insides and outside, natures and cultures, materiality and virtuality, representations and interventions, texts and bodies. It therefore offers itself as an apposite moment for observing "the inextricable entanglement of material, biocultural, and symbolic forces" that distinguish the nexus between biology and (information) technology (Smelik & Lykke 2008, pp.xxiii–xxiv). More specifically, the database record enacts different semiotic and material figurations of the sequence and its universe. In order to account for the record's vibrant capacities, this chapter activates the two sample records shown in Figures 5 and 6 as quasi-ethnographic sites. Just as chapter 4 has followed an ethnographic journey through the sequence universe, this chapter proposes a similar methodological orientation by analytically exploring the record as a bioinformational artefact. More precisely, it attempts an exegetical *reading* of them as *dialogic texts* (Bakhtin 1981). This is of course not how these records are, or indeed, should be, read. In fact, once deposited, they are very unlikely to find readers like me because it is mostly algorithms that "read" database records. What I will do in the following paragraphs, however, is read the record in order to extract wider stories and illustrate, performatively so, how records mediate between apparently irreconcilable ontic states and epistemic figurations.

This is not a dialectical attempt to contrast and reconcile text and phenomena but to engage with *texts as phenomena*.[124] In doing so, I am not subjecting the database records to a strictly semiotic analysis. Instead, I illustrate how records are in continuous correspondence with other texts but also, as I suggest, *other entities* such as oceans (see below). Like the readers-*cum*-protagonists in Italo Calvino's *If on a winter's night a*

[124] This can be seen as one of the inventive aspects of the *nouveau roman* but is also a characteristic of writers such as James Joyce, Flann O'Brien, Italo Calvino, George Perec, Jorge Louis Borges and Gertrude Stein.

traveller (1981), the process of my reading performatively renders relations and occasions. Through an exegetical immersion in the record, the chapter assembles a number of situations in which data and information refuse to level with the record's appearance. A text is a versatile device in sounding out *presences* and *absences*. Much like the *Sorcerer II* expedition that yielded the uncultured bacterium documented in the GenBank record (see below), reading throws out probing lines, prospects landscapes and explores the deep ends. The kinds of reading outlined in this chapter correspond to what I call *non-commensurate reading*, that is, a way of reading that generates residuals, leaving – so to say – a lot to hope for, infer from and associate with. This is perhaps the kind of reading which Rheinberger's concept of *experimental écriture* would invite – a writing that encompasses texts, textures and contexts (Rheinberger 1997). Here, the database record proliferates relations across very different entities, gives rise to a menagerie of unlikely neighbours and yet still manages to *make sense* and effect representations that matter. The efficacy behind making sense and being effectual "spokespersons" of objects is customarily tied to immutability. What kind of effectiveness or kind of resonance can we conceive through objects when we loosen the strings around immutability and tether some (partial) connections to vibrancy instead?

A prologue for the records: presence, absence and invention

The EMBL-Bank record relates a sequence derived from *Pilobolus crystallinus*, also known as dung cannon, a coprophilous fungus that lives on animal dung. In contrast, the GenBank record documents a new, yet to be identified bacterium that was sequenced as part of the large ongoing Global Ocean Sampling (GOS) project carried out by the J. Craig Venter Institute (JCVI). The former was chosen because I encountered the dung cannon in a talk[125] expounding the wonders of our "mushroom planet" and because fungi made several appearances in the course of my research: They prompted a controversy involving GenBank, which is discussed in the next chapter. They were also repeatedly mentioned in my interviews with curators and staff at the databases to illustrate biological complexity

[125] Delivered by Kathie T. Hodge, leader of Hodge Labs at Cornell University and renowned mycologist. See http://www.cornell.edu/video/index.cfm?VideoID=559. Last accessed: 3 December 2009.

and uncertainty. And lastly, the beginning of my fieldwork at GenBank was accompanied by food poising, which can be caused by fungi (mycotoxins).[126] Conversely, I chose the GenBank record because the GOS project is considered the biggest single deposit (or "data dump" as my respondents called it) into GenBank.

How to get to these records? Browsing to the EMBL-Bank website I locate the search function. EMBL-Bank lets me search "All Databases", a first indicator that I am about to step inside a larger discovery environment, and I enter "Pilobolus crystallinus". This generates results in 4 different categories: nucleotide sequences (EMBL-Bank), protein sequences (UniProtKB), literature (Medline) and ontologies (Taxonomy). As I am interested in EMBL-Bank, I concentrate on the first item listed in the first category: "FJ536284 Pilobolus crystallinus putative blue-light photoreceptor PCMADA1 mRNA, complete cds." The accession number is hyperlinked to the record, and the description followed by a choice of views: ENA, EMBL format, SRS, and EMBL-SVA.[127]

Starting to read

I choose "EMBL-Bank format" knowing that this will take me to the flat file. Whereas the ENA view (see below) formats the data in a visual way and offers ways of interactions such as hyperlinks, the "EMBL-Bank format" (Figure 5) does away with colours, interactive graphics and visual layout elements. Instead, it shows the sober "flat file", a line-by-line, black-on-white text. There are no hyperlinks or buttons to click, all that can be done with it is to *read* it. So that is what I start doing.

In the beginning I encounter numbers and abbreviations. Apart from dates, the numbers and abbreviations require some basic familiarity with the databases to decipher (though what they actually *mean* requires an altogether different level of knowledge). There is the accession number (FJ536284 and EU805409) that is assigned to each record by curators once it has been established that the submission meets the minimum criteria

[126] Also, genomic and post-genomic research is ruled by yeast – it was the first organism to be sequenced and its traces are everywhere. For example, the two-letter code used to abbreviate modified bases in the record's *Features* section has been adopted from a 1982 paper by Sprinzl and Gauss which had applied the numbering of tRNA used for yeast to all published tRNA sequences.
[127] SRS is a life science data integration tool that provides a uniform interface to different data sources. SVA, the Sequence Version Archive, contains all entries which have ever appeared that enables researchers to see what a particular record looked like on a specific day – a kind of time machine for the universe.

(see chapter 5). It also tells me the molecule type, mRNA (messenger RNA) for the fungus and DNA for the bacterium. The expression "linear" refers to the plasmid DNA used to clone the sequence in the laboratory. In addition, the data class in the EMBL-Bank record is given as standard (STD), which indicates the methodology by which the sequence data has been generated.[128] Furthermore, the introductory section indicates the length of the sequences (1,869 and 1,491 base pairs, respectively) and its taxonomic division (FUN and ENV). In order to decode the taxonomic divisions, I consult the EMBL-Bank's user manual provided in the Documentation section of the EMBL website and written to assist both, submitters and researchers. The manual, aside from FUN (fungus) and ENV (environmental), offers the following list of possible divisions: human, invertebrate, other mammal, other vertebrate, *Mus musculus* (house mouse), plant, prokaryote, other rodent, synthetic, transgenic (organism with genes inserted from another species), unclassified, viral. This, I surmise, reads more like the division of animals in the Chinese Encyclopaedia quoted by Borges (1964).[129]

The description line of the record offers some more details about the source and function of the sequence. The EMBL-Bank record's description states that the sequence is derived from *Pilobolus crystallinus*. This fungus resides in different environments during its lifespan: in animal intestines, in meadows and grassland and on animal dung. Most of its life is spent on dung from where it also releases it spores at very high speeds, hence the colloquial term "dung cannon". It is not just the speed of the spores' release that has demanded scientists' attention but also its aim: *P. crystallinus* directs its volleys at the sun, specifically in the mornings and evenings. This ensures a trajectory that will land the spores in surrounding grassland rather than return them onto the dung. Such precision is necessary because its spores require animal intestines for their development, and deposing them on grassland increases the likelihood that they will indeed be ingested by a suitable host, a cow perhaps. It is this very ability of tracking the sun that the part "putative blue-light photoreceptor" refers to. The use of the word "putative" after the

[128] Other examples include WGS, EST for expressed sequence tag or GSS for genome survey sequence.
[129] Made famous by its inclusion in Foucault's *The Order of Things* (Foucault 1970), it lists animals "that belong to the Emperor" to "embalmed ones", "fabulous ones", "stray dogs" and "those that from a long way off look like flies".

organism name suggests that this record represents an *inference* about the sequence, that is, there is no direct experimental evidence to confirm the product. The product in this case is given as "blue-light photoreceptor", indicating that this sequence is associated with phototropism, a tropic response in plants that most commonly makes them bend towards light. "PCMADA1" is the name given to the product that this sequence *is believed to* produce while "complete cds" specifies that this record's sequence contains a complete coding region.[130]

The succeeding section, prefaced by "OS" (for "organism"), designates the scientific name of the organism from which the sequence was obtained, using the Latin genus and species names in the standard binominal nomenclature. This is followed by details of the organism's taxonomic classification ("OC"), based on a taxonomic tree model and listed in top-down structure, meaning that the general (highest) order comes first. The EMBL-Bank record begins with the superkingdom (or domain) of eukaryotes and includes the order (*mucorales* or pin molds), the family (*pilobolaceae*) and the genus (*pilobolus*). The taxonomic classification given in the record also contains "Fungi incertae sedis", which literally means "fungi of uncertain placement". As is discussed in the next chapter, fungi are especially adept at defying order and taxonomic classification. Here we can see how this capacity has been accommodated in the taxonomic tree where "incertae sedis" does assume a very certain position and functions as a formal category of classification. *Incertae sedis* is not limited to fungi. Instead, it delineates a taxonomic group whose wider relationships have yet to be defined or are unknown. It can therefore designate uncertainty on two levels: It either refers to an ambiguity of membership of the higher taxon or, as in this case, to an uncertain phylogenetic position in relation to other members of the taxon. In comparison, the GenBank record contains very sparse details about taxonomic classification, merely stating "Bacteria; environmental samples".

[130] The designation "madA" refers to a collection of genes (named madA to madJ) that are associated with phototropism. Phototropism was studied by Max Delbrück in the 1960s on the basis of *Phycomyces* mutants where this ability was inhibited. In honour of Delbrück, these mutants were named *mad* mutants, hence the designation of *mad* genes (Idnurm et al. 2006).

Bioinformational artefact: presence in absentia

I not yet reached the complicated part of the record (the *Features* section) and already I am faced with incongruities. There is a conjectural association between a fungus, light-sensing capacities and a gene. There is no information on the methods and techniques employed to obtain the sequence. There are lists, classification systems even, which conform entities of ontologically and epistemologically very different natures. And organisms appear in uncertain places. The clean appearance of the flat file betrays a conglomeration of rather disordered or unsorted things, giving way to multiplicity, historicity as well as subjective interpretation. It appears that like in Flann O'Brien's *At-Swim-Two-Birds* (1939), the characters populating the record are bent on scorning the script and spill over the edges of the page.[131] Yet, the discordant array of entities still manages to do its *working*. *Incertae sedis* is not only a viable *category* but indeed a very necessary condition that needs to be actively preserved – much like many of the organisms that are designated by it. In response to a motion put forward to the South American Classification Committee of the American Ornithologists' Union that suggested the elimination of *Incertae sedis*, one ornithologist responded:

> Incertae Sedis allows us the necessary option to indicate a reasonable level of uncertainty. It further provides the useful educational function of alerting everyone to the need for further research and that much research is still needed to make certain that higher-level taxa represent monophyletic groups.[132]

This then describes one way in which indeterminacies work. How else can we account for the records' working?

How not to know

Both records carry indeterminacies in their titles: Whereas the EMBL-Bank record makes a "putative" inference about the light-sensing pathway in *P. crystallinus*, the GenBank record documents an "uncultured" bacterium with no species name and no further details

[131] Chapter 7 shows how bad data, just like badly written figures from Irish mythology, can come to life and haunt their authors (and readers).
[132] For the (rejected) proposal (#359) put forward by Manuel Nores in 2008 and the responses, see http://www.museum.lsu.edu/~Remsen/SACCprop359.html. Last accessed: 10 February 2011.

about its taxonomic status. "Uncultured" here means that the sequence was obtained from an organism that cannot be grown in the laboratory. In fact, more than 99% of bacterial species remain uncultured (D'Onofrio et al. 2010).[133] To some extent, this uncertainty can be interpreted as absence: absence of experimental confirmation for the EMBL-Bank record and absence of a "cultivable representative" for the GenBank record. The mRNA sequence of *P. crystallinus* contains a stretch that is significantly similar to the sequence of *madA* genes, a stretch that the submitter has termed "PCMADA1". As the *madA* gene has previously been identified as the blue-light receptor in algal fungi of the same class, the submitter assumed it had the same function in *P. crystallinus*. Though the specific nature of the region in relation to *P. crystallinus*' tropic response remains *putative*. In this case, the record mediates relations between the fungus *in silico*, *in vitro* and *in vivo* which suggest that accuracy here is not a matter of fidelity to nature or approximation to any other reference value "out there". Rather than recording or documenting an occurrence, the record makes room for presences as well as absences. In doing so, it is representative for the ways of knowing in and through the sequence universe as the next paragraph will detail.

Reflecting on the landmark that was the first complete manual annotation of what is believed to be the full set of human proteins (proteome)[134], Amos Bairoch of the Swiss Institute of Bioinformatics argued that this effort revealed great gaps in our knowledge of proteins. What they do, where they do it and how still remain largely unanswered. But, according to Bairoch,

> [t]his gap resides not only in the available experimental information, but also in the way this information has been stored, which is far from being sufficient to help researchers *making sense* of what all these human proteins do in our bodies! (Bairoch 2010)

Knowing or not knowing proteins is a matter of the presence or conversely absence of information as well as of the way in which such information *is existent*: how it is

[133] When it comes to bacteria, "most Divisions – the largest taxonomic units – do not have a single cultivable representative, and we know of their existence only from 16S rDNA isolated directly from the environment." (K. Lewis 2010).
[134] This was achieved in September 2008 by UniProt/Swiss-Prot, a protein database with which EMBL-Bank and GenBank are integrated. Part of the UniProt/Swiss-Prot team shares a building with EMBL-Bank on the WTG campus.

assembled, where it resides, in what shape or form it does so. Much like the proteins themselves, database records, once released into the sequence universe *do* things which might confound. Here, the database record, far from an immutable and mobile yet docile inscription becomes a vibrant, potentially active, constituent. Bairoch's articulates uncertainty in relation to the characterisation status of human proteins by means of a pie chart that comprises the following segments: maybe, potentially, putative, expected, probably, and hopefully.[135] Although there is no indication in Bairoch's presentation of whether this list of adjectives corresponds to a hierarchy of indeterminacy, where determinacy increases from "hopefully" to "maybe", it bears testimony to the considerable amount of apprehension, doubt and belief invested in knowing proteins and their functions: Almost half of protein characterisation is deemed "hopefully". This shows not just the extent to which protein characterisation remains indeterminate but also the complex way in which this lack of knowledge itself becomes classified and made known. This highly artful register of absence also betrays life forms, such as the uncultured bacterium, that inhabit the spectrum between *Maybe* and *Hopefully*. Bairoch's presentation casts the unknown and absences as resources and evidences the extent to which the *in silico* discovery environment enfolds inventive ways of being and becoming.

The two records analysed in this chapter are representative of two distinct ways in which the sequence universe has complicated the indexing of the natural world. The use of *Incertae sedis* points to the wider shift in naming and classifying organisms occasioned by phylogenetics, that is the determination of evolutionary relationships on the basis of DNA/RNA sequence (rather than morphology). Conversely, the detailed explorations of the molecular structures and pathways have unravelled radically novel landscapes. Both contribute to the emergence of strange neighbours and unexpected kith and kin. The next section will examine some of these unusual relations.

Links in the sequence universe: accumulating relations

I return to my initial search for "Pilobolus crystallinus" and once again access the first record "FJ536284.1" but this time through the ENA view. The record opens up in a clean

[135] An approximate distribution: Maybe (~10%), Potentially (~13%), Putative (~7%), Expected (~10%), Probably (~20%), Hopefully (~40%).

and well-ordered web page, subdivided into 7 sections by pale purple bars (Figure 7). Given the frequency of hyperlinked terms, from every stage of the lineage to all the protein matches in InterPro (see below), this is not a space for dwelling but for taking off. Chapter 4 has demonstrated the sequence universe to afford different kinds of travels and travellers. The present case might reveal some ways in which such travels are facilitated. Will following the hyperlinked terms get me further *into* the record, deeper down into "Pilobolus crystallinus putative blue-light photoreceptor PCMADA1 mRNA, complete cds", or will it take me away, both literally and metaphorically? It appears that, if I decide to linger in this view despite

Figure 7: ENA view for the EMBL-Bank record

the visual clues to do otherwise, I should follow one of the prompts for interaction, such as specify the "Visible feature range" or "Find similar sequences". That the record is dotted with springboards and exits in the form of hyperlinks is incidentally very fitting for *Pilobolus crystallinus* given its capacity for accurate discharge.

The liberal dispersal of hyperlinks in the record that point to other constellations in the sequence universe is indicative of the importance of relationality: Like the fungus *in vivo*, the database record cannot exist by itself and on its own. In other words, it requires a level of convivality in order to make sense and be useful. To explore the make-up of this convivial relationality further, I choose the StrainInfo link which is given under the *Navigation* section for the record. This does indeed take me away as it lands me on a site outside the EBI's web domain (straininfo.net). As the name suggests, the website offers information on the particular strain of *P. crystallinus* that the record refers to. Including the strain information can be significant because strains, as genetic variants of microorganisms, can have different phenotypic expressions. The StrainInfo page (it tells me it is "Beta", meaning in testing) displays what it terms the "Strain Passport" for "NBRC 8561 *Pilobolus crystallinus*". Even in the sequence universe, it appears that one requires travel documents.

Subdivided into four boxed sections, the strain passport offers overview, history, sequences and publications. A number of unfamiliar logos placed at the bottom of the page alert me to the producers of the StrainInfo site. Clicking on each logo, signs reveal signified by launching their respective homepages: the Belgian Science Policy; the Laboratory of Microbiology at the University of Gent; the KERMIT Research Unit Knowledge-based Systems, again based at the University of Gent; and the University of Gent. It is perhaps not surprising that an entity issuing "passports" originates from Belgium, the seat of the European Commission, associated with bureaucracy and its endless production of documents but also with the making of a somewhat common world.

You say passport, I say potato

Passports for humans are issued by national governments and contain name, date of birth, sex and place of birth. Similarly, the strain passport for *P. crystallinus* features its name,

NBRC 8561. It also presents an "availability map" showing that the strain NBRC 8561 is available from Japan but that other strain numbers can also be obtained from Taiwan. NBRC stands for the NITE (National Institute of Technology and Evaluation) Biological Resource Centre in Chiba, Japan. They, like other biological resource centres around the world, produce and sell a variety of biomedical products including bacteria cultures, human cDNA (complimentary DNA) clones and microbial strains.

Instead of a date and place of birth, the strain passport includes a history or "histri" visualised by means of a tree diagram. Histri here refers to a strain's exchange history – from their first isolation to their deposit in a biological resource centre. This is tracked and compiled because "[e]ach transfer imposes a risk of contamination or human mistake, and thus of possibly confusing scientific results. In order to check the validity of strains, we need to reconstruct their exchange history ('Histri')".[136] In the case of NBRC 8561, the original isolate is identified as "FN-2" but it tells me: "This Histri was built automatically but not manually verified. As a consequence, the Histri can be incomplete or can contain errors." Behind the veneer of official documents and strain numbers, which assemble to vouch for *P. crystallinus*' identity, lie cascading contingencies: false histories, contamination, human error, confusion and artefacts. What appears more troubling yet, is that NBRC 8561's original isolate FN-2 does not appear to be hyperlinked. After the abundance of links propping up the EMBL-Bank record in the ENA view, an entity like FN-2, a dead-end, seems suspicious. As there is no more information to be gathered on its passport, I follow the fungus' paper trail and pay a visit to its last known address.

This takes me, via the StrainInfo website, to the NBRC. There, I locate NBRC 8561 in the catalogue, which details the 16,209 microbial strains (2,775 yeasts, 7,843 fungi, 160 archaea, 5,063 bacteria, 308 algae, 60 bacteriophages) that are preserved in the NBRC and made available for distribution.[137] The catalogue lists another common name for the fungus, Wiggers Tode.[138] Aside from expanding on the history of the organism's documentation and its concurrent accession into scientific purview, the catalogue also

[136] See http://www.straininfo.net/docs/histri. Last accessed: 17 April 2011.
[137] See http://www.nbrc.nite.go.jp/e/catalog-e.html. Last accessed: 17 April 2011.
[138] This refers to Friedrich Heinrich Wiggers, a German botanist and doctor, who in the late 18th century had first named and published this taxon (under its basionym, the first name ever given, of *Hydrogera crystalline*).

supplies details on the gestation of this particular strain. It tells me that this strain of *P. crystallinus* was grown at a temperature of 24 degrees Celsius, using rehydration fluid containing peptone, yeast, Epsom salt and distilled water. Furthermore, the medium used to grow it on is identified as potato sucrose agar (a Petri dish with potato sucrose as a nutrient) for which the NBRC catalogue provides the following instructions:

> Wash potatoes with tap water, peel and cut into 1 cm cubes. Rinse with tap water quickly and boil 200 g of potato cubes with 1 L of distilled water for 20 min. Mash and squeeze through a muslin bag. Add sucrose and stir till dissolved. Adjust pH. Add agar. Make up to 1 L. Autoclave to sterilize.[139]

This is not so much the "place of birth" than details about the birth itself. For the identification of biological materials it is not only the who and what but the where and how that is of import. Again, relationality and conviviality are stressed: Pilobolus needs to be close to potatoes, cotton and sugar in its gestation which, as the instructions relay, also involves the care and maintenance of human hands. Despite the apparent ontological distance between *P. crystallinus in vitro* and the database record they undergo similar cultivation and husbandry processes (see chapter 5 for caring in relation to database records). Furthermore, the strain passport reveals some of the commercial relations that exist through the sequence universe, namely the global trade in specimen and strains for scientific research.

References: holding together a conditional universe

After this excursion, I return to the EMBL-Bank record in flat file view. The section that follows the source organism information is *Reference* section. Here, details are given about the location of the sequence in other forms and media, such as in science literature or indeed other bioinformational resources. For *P. crystallinus* we can discern a cross-reference location (Agricola) for the sequence as well as its identification number in this location (IND44262955). Agricola is the US Department of Agriculture's National Agricultural Library, the world's largest agricultural information collection. It looks as if

[139] http://www.nbrc.nite.go.jp/NBRC2/NBRCMediumDetailServlet?NO=1. Last accessed: 17 April 2011.

our fungus might have some bearing on the world of agriculture. On the next line, the record provides the unique identification number for the literature reference, its digital object identifier (DOI). In addition, the record lists the reference author as "Kubo H.", the reference article's title ("Isolation of madA homologs in Pilobolus crystallinus") as well as the publication details, stating that the article was published in *Mycoscience*, volume 50, number 5. A second reference reveals when the sequence was submitted and the institutional affiliation of the submitter, which, in this case, is the biology department at Shinshu University in Matsumoto, Japan.

In moving to the *Features* section, which describes any higher order genetic roles and functions, further references are revealed. After some introductory information that re-iterates much of what has gone before (molecule type and organism information), the feature starts in earnest. It is first named as a coding sequence (CDS), which indicates that this feature represents a sequence of nucleotides that corresponds to a sequence of amino acids in a protein. This means that the sequence submitted is part of a region that codes for a protein and, therefore, contains a gene. The start codon is given as 1, meaning that the translation commences with the chain initiation codon ATG (also read as AUG).[140]

The following lines list a number of further cross-references. The first, GOA:B8YIE3, points to the gene ontology annotation database UniProtKB-GOA. This database maintains what is called an "ontology" of terms related to genes, genetic processes and functions. Given that automated text processing is a key function for *in silico* discovery, a consistent nomenclature – of gene and protein names but also molecular functions and processes – is extremely important. These are maintained by gene and protein ontologies. However, there are no universal guidelines and rules and no definitive authority though the Gene Ontology (GO) project aims to alleviate this by standardising "the representation of gene and gene product attributes across species and databases". Gene ontology here describes functions, processes and components using a controlled vocabulary of terms that are related in such a way as to facilitate computing on the basis of annotations within and across species. For example, searching the GO database (which can

[140] The start codon depends on the organism – although ATG is so far the most common chain initiation, alternative codons are possible. The codon sets the reading frame. It describes the set of three nucleotides in an mRNA molecule with which the ribosome begins protein synthesis.

be done via the EBI interface) for the term "apoptosis", a form of cell death, the GO database yields, among other information: a unique GO ID for the term, a definition, a list of "child terms" (such as "inflammatory cell apoptosis"), and a set of synonyms. The latter are graded, from "exact" (apoptotic cell death") and "narrow" ("apoptotic program") to "related" ("signalling [initiator] caspase activity").

Where GO annotation has been experimentally confirmed and manually curated, it serves as a basis to predict functions of related but yet uncharacterised gene products. InterPro, which the record lists in the following six cross-references, serves as an integrated database of protein "signatures", which describe *predictive* models for representing protein domains, families and functional sites. InterPro, currently produced by 17 different organisations including the Sanger Centre, provides a taxonomy for protein sequences, grouped in families that are putatively linked to the occurrence of functional domains and other important sites. This facilitates the classification and automatic annotation of proteins and genomes: Comparing an unknown sequence against the InterPro database yields signatures linked to homologue sequences. By checking the IDs against the database, a set of additional detail about the sequence is revealed, in particular, domains and their predicted functions.[141] The last cross-reference, UniProtKB/TrEMBL, refers to the un-reviewed and automatically annotated section of the UniProt Knowledgebase, the central collection and access point for functional information on proteins, built and maintained by EBI, the Swiss Institute of Bioinformatics and Georgetown University.

The *Features* section endows the record with very real capacities: transducting signals, binding DNA, regulating transductions and folding proteins. At the same time, it points to a very rich and highly nuanced vocabulary for describing genetic processes on a molecular level. It also reveals the extent to which a function like *P. crystallinus*' phototropic response is assembled by linking multiple processes. The precision with which GO relays molecular going-ons would also suggest that *how* such processes are

[141] IPR000014 links the record to the PAS domains that are involved in signal transduction. IPR000679 and IPR013088 link to a Zinc finger domain, a small protein motif that contains multiple finger-like protrusions that are involved in DNA binding. IPR000700 refers to a domain involved in two-component signal transduction while IPR001610 is associated with the regulation of transcription. IPR013655 relates to a domain concerned with protein folding.

described becomes crucially important to its function. Description here is performative in that it actively contributes to assembling a sequence's functions while enabling other sequences to be associated with (similar) functions. Generally, these references and cross-references, like the hyperlinks in the ENA view, point to the multitude of other resources that make up the sequence universe. Resources such as InterPro and UniProt, though ostensibly coherent entities, are often an aggregation of many different groups, consortia, organisations, methodologies, types of data, degrees of diffusion (for example, "exact ", "narrow" and "related") and levels of validity (for example, experimentally confirmed, manually curated, automatically predicted).

Amidst these references, the record emerges as a heterogeneous assembly. It is an *account* of itself that includes (links) to the conditions of its own emergence. In *reading* the record, the entities it thus assembles get a chance to narrate themselves. Here, the record reveals itself to be a story of relations amongst many constituents – some of them very present, like the nucleotide sequence which is given in full in the last section of the record. Others, like the dung cannon, are not so present though some of its forms, such as its taxonomic lineage, are rendered. Other others (*pace* John Law), such as the algal fungi that underwrite the putative association between PCMADA1 and phototropism, only appear when following the references (and clues) which the record encompasses. Reading the record undoes its initial flat coherence and posits it as a product of a *cumulative relationality*.[142]

Like in Porter's journey recounted in chapter 4, links are followed and connections are made. The record makes itself intelligible to others yet on terms which are not its own (J. Butler 2005). Here, disparate data are assembled yet instead of narrowing in scope, complexity or meaning, as is so often imagined for inscriptions (B. Latour 1990), they spread and swell in every sense. Given the many kinds of certainties, presences and absences, the sequence universe seems to be upheld be a vast hypothetical scaffold of references, functions, processes, hopes and pathways. Rather than forming a space of universal knowing, it favours "partial intelligibility" (Fisher 1998, p.138) where it is wonder and not recognition that makes us "see a question" (ibid., 83). The following

[142] Thanks to Mike Michael for this expression.

section, focused on the GenBank record, will expand on the record's dialogical import while analysing the likely presences (and absences) that underpin such wondrous exploration and probing.

Hopeful presences and uncultured encounters

> Inspired by 19th Century sea voyages like Darwin's on the H.M.S. Beagle and Captain George Nares on the H.M.S. Challenger, The Sorcerer II circumnavigated the globe for more than two years, covering a staggering 32,000 nautical miles, visiting 23 different countries and island groups on four continents.[143]

This is how the J. Craig Venter Institute's (JCVI) website describes its Global Ocean Sampling (GOS) project which, since 2003, has been circumventing the globe with a research vessel collecting ocean samples. The pilot study in 2003 took place in the Saragossa Sea near Bermuda and recorded about 1,800 different prokaryotic species (J. C. Venter et al. 2004). Following on from this, the JCVI embarked on the *Sorcerer II* Circumnavigation (2004-6) and a second expedition (2007-8) that focused on more diverse and extreme environments, such as deep-sea vents and Antarctica. The projects were sponsored by the Department of Energy, the Gordon and Betty Moore Foundation and the Discovery Channel, which produced a documentary on the *Sorcerer II* expedition.

GOS represents an instance of environmental sampling where DNA fragments are extracted from mixed microbial assemblages in their natural environment such as lake water, acid mine drainage, farm soil, whale fall or human gut. Metagenomic analyses, despite being based on analyses of such assemblages, allow for the isolation of complete genomes from otherwise uncultivable species contained within environmental samples.[144] Therefore, metagenomics constitute a form of "culture-independent" analysis and are particularly prominent in the study of microbial biodiversity as most of this remains uncultured, that is, resistant to isolation and cultivation in laboratory settings.

In the course of GOS, water samples of between 200 and 400 litres are taken at approximately every 200 miles. These are then processed through a sequence of

[143] From the GOS project description at http://www.jcvi.org/cms/research/projects/gos/past-voyages/#c1302. Last accessed: 30 March 2011.
[144] Only an extremely small percentage of cells (0.1-1%) derived from environmental samples can actually be cultivated on a synthetic medium.

progressively smaller filters (20 to 0.1 micrometers) in order to capture differently sized organisms, from plankton to eukaryote viruses. Once extracted, the filters and the microorganisms contained within them are frozen and sent back to the JCVI laboratories in Rockville, MD or San Diego, CA where they are sequenced using shotgun sequencing.

For the INSDC, the deposit of data GOS represented the largest dataset ever to be put in the public domain. The GenBank record in Figure 6 is one of the 7.7 million sequences that form part of the GOS submission. In order to get to one of these sequence records deposited in GenBank, I search for the GOS project in Entrez. From there, I am taken to the BioProject database. There, I locate the master record for the expedition which, in a stack of tables, records key project data and identifiers and links to 9 sub-projects. I can ascertain that GOS has deposited over 3 million nucleotide records and over 6 million protein records and that there are a dozen publications associated with the project. Choosing the Metagenome sub-project, I am taken to a similar BioProject page, entitled "Global Ocean Sampling Expedition Metagenome: Metagenomic analysis of marine microbes isolated during the Global Ocean Sampling Expedition". Once again, tables list project data and links to publications. But there is additional information: A narrative describes the purpose of GOS whose broad objective "is to assess the genetic diversity in marine microbial communities and understand their role in fundamental processes in nature".[145] The project "attributes" list scope ("Environment"), material ("Genome"), capture ("Whole") and method type ("Sequencing"). And like organisms, projects too come with a lineage. This one reads: "unclassified sequences; metagenomes; ecological metagenomes; marine metagenome".

Excess in absence

There is an enormous, excessive amount of data accessible through this page. It seems that this space is a continuation of the ocean by other means. This is certainly congruent with how curators were describing metagenome submissions:

> They were trying to fit the metagenomic submissions into the same genome submissions that we're using to process WGS and complete genomes. Well, a lot of those started off with "I know

[145] See http://www.ncbi.nlm.nih.gov/bioproject/13694. Last accessed 30 March 2011.

> what the organism is". So there's been a lot of work on updating the Genome project database to accommodate this metagenome data. There's really a *lot*… there'll be 10 different projects under Yellowstone hot springs. That was on that hot spring or that one was on a different day... cause they are all different microenvironments. Metagenomes are gonna be a challenge. GB16

Not only have these submissions exceeded the data templates that had been in place until then, but they have also created multiples of their originating environments (10 different Yellostones). A further complication arises from researchers not knowing "what organism they [the sequences] come from" (GB18). Thus, environmental sampling is very much a shot(gun) in the dark. One way to identify things in this impenetrably rich assemblage of microbes, sludge and data is by means of targeted gene surveys. In looking for one of the organisms "discovered" by GOS, I therefore turn to the GOS sub-project "Targeted Gene Survey from Global Ocean Sampling (GOS) Expedition". This takes me to a list of 6,413 nucleotide sequence entries, specifically, records for 16S ribosomal RNA sequences associated with GOS. The 16S rRNA gene has become the key region for genus and species identification for bacteria and archaea. It forms a key target locus for metagenomic analysis in gene targeted surveys.[146] Via sequencing all 16S ribosomal RNA present in an environmental sample, phenotypic characterisation is determined and known as well as previously unknown species can emerge. The GOS project, too, used this technique to estimate the microbial (phylogenetic) diversity and create taxonomic inventories of marine microbial populations. It has subsequently opened up a vast population of previously unknown microorganisms (Hugenholtz & Tyson 2008; Pignatelli et al. 2008).

Meaningful metagenomic analysis, beyond phylogenetic identification, "relies heavily on the accurate knowledge of the universe of proteins stored in the databases." (Pignatelli et al. 2008) In the course of the initial analyses of the *Sorcerer II* samples, not many reads could be assembled into scaffolds – the microbial diversity was just too large and, hence, no sequence similarity could be established (D. B. Rusch et al. 2007). Instead, the research team used completed (and draft) microbial genomes that had already been

[146] This is because these genes are "conserved across vast taxonomic distances (…), yet show some sequence variation between closely related species" (McEntyre 2004).

deposited in public databases as reference and used less conservative search parameters to discover even small similarities to GOS sequences.

The first bacteria in the list is entitled "Uncultured bacterium clone 6C233420 16S ribosomal RNA gene, partial sequence". Using the taxonomy browser, the full lineage for the taxon reads: "Cellular organisms – Bacteria – environmental samples – uncultured bacteria". We can therefore surmise that this record attests to the existence of a particular uncultured bacterium by providing the sequence of its 16S rRNA, which, as it has no other match in the sequence universe, is recognised as a new species.

Compared to the EMBL-Bank record, the GenBank record relates (to) a very different entity. The EMBL-Bank record, as discussed above, evidences a putative association between a sequence derived from *P. crystallinus* and a biological process – a process which contributes to a tropic response to a light source. It therefore mediates *in vivo* observable occurrences: a dung cannon and the directional and timed release of spores toward the sun. There, the organism – derived from an identified biological resource centre – resembles a technical object as the epistemic emphasis lies on the PCMAD gene and its relation to phototropic mechanisms. The dung cannon acts as an almost incidental showcase for a particular permutation of this process, which can be found in many other filamentous fungi. The GenBank record, on the other hand, documents a less concrete, manifest or, indeed, *real* entity. Yet, its focus, its entire purpose, is to attest to this entity's existence. To be sure, unlike *P. crystallinus*, grown in Japan in a Petri dish containing potato sucrose, the bacterium here comes straight from its "natural" environment. As discussed below, however, this "natural" environment is in fact a carefully assembled scenario involving state-of-the-art instruments, national governments, the Pacific Ocean, Craig Venter and other actants. In metagenomic analysis the organism is (long) gone as identification happens on the basis of DNA clones of fragments. At the same time, it operates on an immense environmental scale, sequencing, in the case of GOS, entire oceans. As much as it facilitates formal identification, environmental sampling raises the spectre of different kinds of presences and absences.

Given the bacterium's fickle existence, the GenBank record does indeed record and document the organism's presence and, in all likelihood, represents the only instance

within literature (scientific and otherwise) that does so. This is because the record in the database is more than often the closest that scientists will come to organisms derived from environmental samples. Given that this uncultured bacterium only existed as part of an environmental sample, does it in fact exist on its own, as an uncultured but individual bacterium? As there is no specimen as such, what does the GenBank record relate to? If the database record is in some ways its only presence, what kind of entity is uncultured bacterium clone 6C233420? The following will examine some "known" variables provided by the record before detailing the ghostly and monstrous presences enacted by metagenomic analysis.

From deep sea to flat file

While it is fraught with indeterminacies in some parts, in others, the GenBank record offers very detailed data, such as in the *Comment* section. This details the conditions pertaining to the organism's collection. Like in the BRC catalogue, we encounter elements that, in concert, offer a vivid picture of the organism's origin. In this case, the data give measurements for six variables: depth, salinity, temperature, sampling site monthly chlorophyll level, sampling site yearly chlorophyll level, hi_filter_size and lo_filter_size. Whereas the last two relate to the instrument utilised in obtaining the sample, the former circumscribe a snapshot of the environment that gave rise to the sample and our uncultured bacterium. These quantities offer a helpful matrix for conjuring up the world out and down there. At 2,431 meters, the environment falls at the border of the bathyal (between 200 and 2,000 meters) and abyssal (between 2,000 and 6,000 meters) zone. Here, the ocean is pitch black and organisms feed off the detritus that falls from the zones above ("marine snow"). The only mammalian visitor, apart from Craig Venter, is the sperm whale who hunts for giant squids. The chlorophyll level refers to the occurrence of a pigment that allows plants to convert sunlight into energy – in the ocean this is used as a measure to indicate the abundance of algae, which in turn determines a given water's trophic state (the quantity of nutrients). The highest levels of chlorophyll (up to 60 mg/kL or mg/m^3) are found in cold polar waters, revealing that the levels measured at the sampling site are very low (congruent with the fact that at 2,000 meters depth no sunlight

penetrates). The level of salinity is given at 32.2 parts per thousands (ppt), which is close to the average salinity of sea water at about 35 ppt. In the *Features* section of the record, some more information is revealed about the uncultured bacterium's provenance. Its isolation source is identified as "250 miles from Panama City", which puts it somewhere just outside the Gulf of Panama. In comparison to the other measures provided, this stands out as rather indeterminate. But importantly, 250 miles makes for a "high sea" spot, just outside the 200-mile zone (exclusive economic zone) that designates territorial waters. To a project like GOS, this distance therefore represents an escape from dealing with access and benefit sharing agreements and policies (and diplomatic embarrassment). All these data help recreate and, importantly for the enthusiastic "reader" of the record, *imagine* the habitat of the bacterium and the moment of its capture. The bacterium might be uncultured in the world "up there" but it thrives very well in the dark depths of the Pacific.

In imagining the bacterium's surroundings on the basis of the *Comment* section, the temperature reading of "29.3 C" appears odd. A temperature of 29.3 degree Celsius seems spurious at a depth of 2,431 meters which is usually associated with much colder temperatures. Like Porter's journey recounted in chapter 4, an incongruous bit of data causes a brief pause and room for speculations. Three possible explanations for this outlier come to mind: The unit is wrong and it actually refers to 29.3 degree *Fahrenheit*, at -1.5 degrees Celsius consistent with the range of temperature at that level. Or, the sample originates from near a hydrothermal vent. This would cause the surrounding waters to be much warmer than this depth usually permits. Or, the temperature relates the *surface water temperature* which again would be consistent with the average temperature range for this area. It is unclear whether the temperature recorded does in fact point to an inconsistency. But either way, it does indicate a limit to intelligibility or perhaps, more accurately, a different scope for complexity.

How does the exactitude of collection measures relate to the indeterminacy of the collectable, the bacterium? In other words, can we use one scale for comparing presences and absences contained by the record? The Belgian passport, for example, performs a degree of certainty in relation to the source organism's provenance and birth conditions. Would experimentally confirming the role of PCMADA1 make for a qualitatively similar

certainty in relation to the organism's functioning? The notation of belief rather than of result is, in most scientific accounts, highly differentiated: Results may be presented in the form of quantities whereas belief most commonly finds expression in narrative accounts which interpret these quantities (even then it is mostly conveyed as *confidence* rather than belief). In the case of the EMBL-Bank record, this differentiation no longer fulfils the function of division between what is known and what is not (quite) known (yet). In fact, the record inverts this function in allowing what might usually count as a gap or a deficiency (the "putative") to act as a connection between entities and different states of knowing.

Making sea monsters

There is something monstrous about metagenomics. The excess of data and the multiplication of sites make for ever more abysmal gaps. This is accompanied by an indiscriminate coming alive: Suddenly sludge and the deepest and darkest sea are teaming with very vibrant and defiant matter. And there is something particularly monstrous about the uncultured bacterium. For one, it emerged from depths, which in our imaginary usually hold quite fearsome creatures. It also resists cultivation in any medium accessible to us and so escapes even the most sophisticated of human interventions. At the same time, Venter's expedition remains uncomfortably close, in both action and metaphor (see the quote describing the mission that introduces this section), to the often monstrous imperial expeditions that have paved the way for colonial horrors to come.

The collection of agreements with national governments regulating the various exploits from the expeditions are archived on the *Sorcerer II* project website. They attest to the easy slippage between making microbes present and the rendering of (novel) spaces of global governance. Here, biopiracy and bio-prospecting, in particular, raised concerns from governments whose waters were entered in order to collect samples as well as from groups engaged with the ethical and environmental issues of biodiversity, such as the Action Group on Erosion, Technology, and Concentration.[147] Like the 19th

[147] The vessel was detained in French Polynesia following biopiracy concerns (Nicholls 2007) while civil society organisations from Chile, Mexico and Costa Rica had already voiced concerns over the JCVI expeditions at the first Americas Social Forum in Ecuador in July 2004. For a detailed account of the controversies around *Sorcerer II* expedition see Rimmer (2009).

century sea voyages it compares itself to, GOS explores and claims not mere "natural" phenomena but is equally engaged in navigating and mapping these into political domains. In the figure of exploration and, more accurately, prospecting, another kind of viability emerges in connection with uncertainty. In colonial literature of the 18th and 19 century, from *Robinson Crusoe* to *Heart of Darkness*, the "unknown" has featured prominently as a device for projecting and installing imperialist fantasies and demands unto the bodies and lands of others. By evoking such narratives, the GOS project continues this tradition, casting the oceans as mysterious and rich environments that new sequence technologies and paradigms (metagenomics) can finally render comprehensible. And much like the imperialist explorations and exploitations, the GOS project too charts a particularly *viable* territory. The unknown of the deep sea is mined for innovation and capital. Here, environmental sampling indeed effects materiality – where there was once water, we now have a bacterium and the record to prove it.

The exploration and study of ecosystems has always relied on counting and measuring as a strategy for rendering the world "visible" and less monstrous. But the worlds of bacteria and many others escape such typing and counting on account of their morphology, habitat or behaviour. Their appearances are quick to change. Bairoch's evocation of the manifold presences that remain unseen and unknown amongst and within us is not without fright. What Law calls "manifest absences" are, in our Euro-American metaphysics (to continue in Law's parlance) most readily associated with ghosts and creatures of the supernatural kind. Many of these manifest absences can be detected in the oceans and it is through database records like the ones discussed above that these manifest absences are materialised. Yet, the subsequent en-culturing and "en-souling" of objects produces a differentiation of presence and absence that far exceeds Law's spectrum and does not necessarily assuage the fright. There is, it seems, a particular position the unknown retains in relation to vibrancy and vitality, a relation that is affective, affiliative as well as pragmatic.[148] Similarly, both Thoreau and Haraway summon

[148] Bairoch expresses the consequences of not knowing by evoking the unknown and autonomous antics that proteins engage in within bodies. This brings to mind Bennett's description of rubbish as "an accumulating pile of lively and potentially dangerous matter" that prompts her to ask: "What difference would it make to public health if eating was understood as an encounter between

"the monster" to describe a world that is hybrid, foreign, unknown yet thoroughly vitalised: Thoreau likens drinking from a stream, which teems with unseen lives, to "[suckling] monsters" (2009, p.67), while Haraway poses the entire world as a monster and us as residing within its belly. Here, the fright is generative as *not knowing* becomes a prerequisite for tuning in to the world. Bairoch's "gap" turns out to be not so much the promise (Haraway 1992) but the *premise* of monsters.

Vibrant workings

In reading the records, this chapter brought to life the data embedded in the flat files populating the databases. Such exploration of their vibrancy also goes some way to highlighting the work they do for scientific discovery. The records themselves are traces of work but they also *work* – much like Feynman's notes. While the processes which have led to their creation, the sampling, shotgunning, tagging, sequencing, assembling and annotating have (long) concluded, they are far from inert remnants. The next chapter will unravel how some database records are being enrolled in heated debates around accuracy and representation. For now, I wish to suggest that the records can relate most vibrant matters in their performance of indeterminacy. In the work of Bennett and fellow "new materialists" (Coole 2010), vibrancy emerges as a "capacity of things" (Bennett 2010, p.viii) that exceeds mere obduracy to human interventions.

For Barad (2007), the capacity of things is evidenced by "agential cuts" through which subjects and objects become performatively delineated and causal structures arise. Barad argues that an agential cut "enacts a resolution within the phenomenon of the inherent ontological (and semantic) indeterminacy" (2007, p.334). Yet, not every time "a body is found [in the archive], (…) a subject can be recovered" (Arondekar 2009, p.3). This is to say that indeterminacy is not always resolved. Sometimes, the agential cut, the making and committing to paper and database of the sequence and associated data, proliferates indeterminacy. At other times, it is this very indeterminacy which agential cuts want to retain. Venter's release of GOS data, though ostensibly serving the public good in making it publicly available, could also serve other interests: "Dumping" data is, at

---

various and variegated bodies, *some of them mine, most of them not, and none of which always gets the upper hand*?" (Bennett 2010, p.viii, emphasis added)

times, as useful as withholding it entirely.[149] Equally, intra-actions assembled around the agential cut, such as preparing, guessing or worrying, can themselves enact cuts which may revert resolutions.

EMBL-Bank and GenBank provide sites for repeated encounters and continuous integration of different kinds of information and data elements. Reading the records has allowed me tease out the diversity of data contained within the databases. Data here take many different forms, each of which affording different kinds of relations, demanding different sorts of attention. The records here are constituted by "ontological indeterminacy" (Parry & Gere 2006, p.139): Whereas databases are usually thought to curb disparities and incongruent entities, here the records enact multiple indeterminacies. They are not fixed entities: Similar to laboratory objects, they are continuously stabilised, de-stabilised and re-stabilised in relation to the ever-changing sequence universe. Their epistemic content is tied to fickle arrangements – homology searches, "extreme" environments, assemblies and predictions. Only through relations does the database record make sense. It *works* because it is able to relate. The *working* of database records is not so much monstrous as it is ghostly. Invested with a strange vibrancy, they make connections and in their moves through various visualisation tools, algorithms and search results, they are made to reason for likely presences *and* absences (Hinchliffe & Whatmore 2006). Here, indeterminacy cannot be tempered by calculation or more (better) information. Likely presences and absences are important strategies in accounting for entities that routinely escape logging efforts or whose habitats do not facilitate continuous, accurate or meaningful recording practices. It is equally important for issues such as biodiversity and, on a more philosophical note, for articulating and managing the kinds of expectations generated around genomic research. Ambiguous ontologies are not antithetical to *results* but point to the exquisite ways developed by genomic sciences to work with different kinds of not knowing and absences.

For Parry and Gere (2006) "ontological indeterminacy" makes biotechnological artefacts hard to grasp. These artefacts are at once digital, informational, technological,

[149] Famously, Venter had refused to deposit the sequences that originated from the HGP in GenBank – something that still irked many of my respondents.

and biological while "[t]heir identities are not fixed, but made and re-made" (Parry & Gere 2006, p.141). Moving through different spaces and times, from the Pacific Ocean to sequence machines and bioinformational resources, a trajectory is assembled that "imparts heterogeneity to the information itself." (Bowker and Star 1999: 290) The document as artefact reveals material practices that draw attention to specific instances of their entanglements with cultural convention, symbolic traditions or epistemic things and communities (Riles 2006, p.7). Doing so renders tangible not only various contexts of production but establishes the record as a continuous effort of multiple practices, each of which, more or less, actively interprets, construes or indeed enacts the artefact (differently).

Chapter 7. *To GenBank with love*: how to address a sequence database

This chapter takes as its starting point an open letter that was written by a group of mycologists and published in *Science* magazine in 2008. The letter faulted GenBank's accuracy in relation to fungal sequence records and proposed the institution of a wiki model for GenBank records. In the following, I present and discuss the letter and the responses it garnered from GenBank and the wider scientific community. The latter's reactions to the letter are gathered from scientific journals and blog entries. The chapter analyses the letter and the ensuing discussions *as a controversy*. In doing so, it introduces and discusses the controversy's key actors and issues and their respective affordances and entanglements. The first section examines the role of affects in the controversy by means of tracing the concern, anxiety and frustration expressed by various constituents. It suggests that the open letter can be an effective political device in rendering affected entities and publics alike. This is followed by discussions of "wikification", annotation and accuracy that highlight the ambiguities of the terms and practices called on by signatories, researchers and GenBank. In the final part, the chapter examines the organism whose sequences have caused the debate, the fungus. Drawing on various fungal instantiations and representations in science and literature, fungi reveal themselves to be very adept at causing a stir and eliciting affective attachments.

Introduction

In March 2008 a letter (the Bidartondo letter) appeared in *Science* magazine that drew attention to inaccuracies affecting certain GenBank sequence records, particularly associated with sequences originating from fungi.[150] It called for a new approach to

[150] The letter (Bidartondo 2008) and an accompanying editorial (Pennisi 2008) both appeared in *Science* magazine on 21 March 2008.

annotation, which was perceived to be the most error-prone section of the records. The letter read as follows:

> GenBank, the public repository for nucleotide and protein sequence sequences, is a critical resource for molecular biology, evolutionary biology, and ecology. While some attention has been drawn to sequence errors (1), common annotation errors also reduce the value of this database. In fact, for organisms such as fungi, which are notoriously difficult to identify, up to 20% of DNA sequence have erroneous lineage designations in GenBank (2). Gene function annotation in protein sequence databases is similarly error-prone (3,4). Because identity and function of new sequences are often determined by bioinformatic analyses, both types of errors are propagated into new accessions, leading to long-term degradation of the quality of the database.
>
> Currently, primary sequence data are annotated by the authors of those data, and can only be reannotated by the same authors. This is inefficient and unsustainable over the long term as authors eventually leave the field. Although it is possible to link third-party databases to GenBank records, this is a short-term solution that has little guarantee of permanence. Similarly, the current third-party annotation option in GenBank (TPA) complicates rather than solves the problem by creating an identical record with a new annotation, while leaving the original record unflagged and unlinked to the new record.
>
> Since the origin of public zoological and botanical specimen collections, an open system of cumulative annotation has evolved, whereby the original name is retained, but additional opinion is directly appended and used for filing and retrieval. This was needed as new specimens and analyses allowed for reevaluation of older specimen and the original depositors became unavailable. The time has come for the public sequence database to incorporate a community-curated, cumulative annotation process that allows third parties to improve the annotations of sequences when warranted by published peer-reviewed analyses (5). (Bidartondo 2008, 1616)

These nearly 300 words draw attention to how GenBank has failed to accurately account for fungi (as well as other unnamed organisms), how such mistakes reproduce themselves across a network of resources, and how GenBank has so far addressed this problem. Most importantly, they propose redress in the form of a new way of annotating that, perhaps

surprisingly (or maybe not, see chapter 5), harks back to the traditions of natural history. It was printed under the heading "Preserving accuracy in GenBank" as the first letter on p. 1616 in the letter section of *Science* 319, published 12 March 2008. The signature appears as "M. I. Bidartondo *et al.* Imperial College London and Royal Botanical Gardens, Kew TW9 3DS, UK" while footnote 5 explains that "[t]he names of all 256 authors can be found in the Supporting Online Material" followed by the hyperlink. In what follows, I shall provide some context to the controversy before moving onto a closer reading of the letter itself and the reactions it spurred.

The issue, inaccurate annotation, had been smouldering for a number of years. Discontent over the quality of information stored in GenBank (and other data archives) had variously been voiced in research articles, opinion pieces and more informal settings (see below). I learnt of the letter's existence at a data sharing workshop in June 2008 through a young bioinformatician who – in responding to my surprise about the perceived lack of data quality in the world's biggest repository for nucleotide sequence –urged me to look into a recent "open letter about mushrooms" published in *Science*.[151] During the workshop, one frustrated delegate had proclaimed that 90% of published peer-reviewed micro-array data is "poor". The quality of annotation is also seen to quickly spill beyond databases: For example, Parker et al. note that the "accumulated semantic ambiguity" besetting naming conventions used in annotation (such as gene ontology) presents a problem not just for "researchers but clinicians, manufacturers, patent attorneys, and other who use biological data in their routine work" (Garrity et al. 2010).

Through the Bidartondo letter the issue received a certain momentum, which was amplified by *Science*'s editorial decision to accompany the letter with a double-spread news item, entitled "Proposal to '*Wikify*' *GenBank* Meets Stiff Resistance". *Science* staff writer Elizabeth Pennisi duly gathered a somewhat irate response from David Lipman, director of the NCBI, as well as comments from the scientific community while elaborating on "the standoff over the quality of GenBank's data" (2008, p.1598). The editorial item brought together reactions to the Bidartondo letter from James Hanken, director of the

[151] "Data Sharing in the Biosciences: a Sociological Perspective", 26 June 2008, National e-Science Centre, Edinburgh.

Museum of Comparative Zoology at Harvard University, who contended that "the problem extends far beyond fungi"; Thomas Kuyper, a mycologist at Wageningen University in the Netherlands who deemed "error propagation [is] all too likely"; Stephen O'Brien, comparative genomics researcher at the National Cancer Institute; Steve Salzberg, bioinformatician at the University of Maryland; and Carol Bult, a geneticist at the Jackson Laboratory. Apart from Lipman, who stressed GenBank's status as an "archive" and thought that the proposal would lead to "chaos", all responses sympathised with the Bidartondo letter's complaint about bad or insufficient provisions in relation to publicly available biological data.

The letter found an echo amongst the *Escherichia coli* (*E. coli*) community, which in response published its own open letter (J. C. Hu et al. 2008). The Bidartondo letter was also discussed in the science blogosphere, including Sandra Porter's blog we encountered in chapter 4, which played host to numerous discussions around the issues raised by the letter (Porter 2008b; Porter 2008a; Eddy 2009; Ebert 2008; Lathe 2008).[152] Thomas Bruns, Professor in the Department of Plant & Microbial Biology at Berkeley, referred to the "open letter to *Science*" in a presentation at the Fungal Environmental and Informatics Network of the Ecological Society of America meeting (2008). Bruns presented on the importance of identifying environmental (metagenomic) sequence, much of which remains unidentified (appearing in GenBank as "uncultured soil fungus") or worse still, misidentified. The issue even prompted Francis Collins, director of the National Human Genome Research Institute, to comment on what he scornfully termed the "news and views coming from *Science* magazine" in his talk at GenBank's 25th anniversary conference (Collins 2008). The *Science* editorial concluded with a gloomy quote by Salzberg: "'I think

[152] The Hu letter offers a more reconciliatory tone. It recognises the shortcomings of the GenBank model, most notably that "individual curators cannot fully encompass the collective expertise of the larger scientific community". And whereas it highlights a number of community-organised wiki-based annotation resources, it also questions the efficacy of radically altering the GenBank approach per se. Rather than further faulting the latter, Hu et al. appeal to both protagonists, the mycologists and GenBank, to work on a collaborative solution based on community-developed tools and GenBank-support of these tools. Appropriately, the Hu letter was simultaneously published on the EcoliWiki, a "Wikipedia for *E. coli*" that facilitates community annotation relating to non-pathogenic *E. coli*. See http://ecoliwiki.net/colipedia/index.php/Letter_to_Science_about_wikifying_genome_information.

it will be solved eventually,' he says. 'But it's not clear how it will be solved.'" (Pennisi 2008, p.1599)

Controversy

In short, the letter caused something of a controversy. Controversies make situations legible by allowing the tracing, mapping and description of its constituents intramurally while accommodating the spectrum of the possible, that is, of emerging externalities. Science has betrayed many of its less visible practices and biases through studies of its controversies (Marres 2007; Whatmore 2009). Examining the actors within the controversy reveals the tenuous lines drawn and subsequently betrays a less clearly defined oppositions between chaos and archive, and, importantly, open and closed. Scientific controversies can be taken as a *genre* of conflictual encounter where genre is "a loose affectual contract that predicts the form that an aesthetic transaction will take" (Berlant 2008, p.847). Figuring this controversy as a genre allows for two lines of enquiry, which this chapter seeks to converge: Firstly, controversies unravel "possible connections between problems" (Callon et al. 2009, p.28) thereby teasing out a much less determined problem-space. Secondly, a genre, as Berlant points out, makes room for decidedly non-representational, that is, affective, claims. Controversy relies on overflows (Callon 1998), that is, on *more than* just the initial actors and (perspectives on) facts: voices become amplified or muted, issues gain urgency, commitments turn fevered, publics are roused and representations are made.

Yet this "more", this adjunctive layer of intensities, is difficult to account for. A recent paper described the onset of controversy as follows: "When scientists and other interested parties challenge contrarian science, the first sparks of controversy appear. (…) I use the term [impedance] to allude to electrical resistance (…) suggesting (….) the 'heating up' of controversy (…)." (Delborne 2008, p.513) One way to account for the heat is by examining the capacity of issues and objects to *affect* and be *affected*. For the purpose of this analysis, I suggest that this capacity is integral to overflows. Put differently, in the following I suggest that in controversies, affects render overflows and *vice versa*. What emerged in 2008 was primarily an intramural scientific controversy lacking the

ingredients that would otherwise ensure a satisfyingly loud public outcry with its attendant political interference: It didn't concern *human* DNA, the main opponent wasn't a Nobel prize winner or otherwise recognisable scientific eminence, and the issue ("accuracy") was dull, to say the least. In effect, GenBank came under criticism from a group of fungal researchers for its inability to remedy inaccurate entries and outdated information. In the subsequent sections, I will describe how the letter, wikification, accuracy, GenBank, fungi and mycologists have found various ways to enrol affect in order to make and stake their claims in the controversy caused by the Bidartondo letter.

Gaps, anxiety and annotation

A letter is a means to bridge a gap, usually of geographical nature. It manifests a "multivalent negotiation of human separation" (Decker 1998, p.10). Therefore, gaps and absence form constitutive conditions of epistolary practices. Crossing a distance in space, it overcomes, literally and symbolically, this separation. In doing so, it attests to a relation, sometimes only to break it. Paradoxically, the letter also asserts this separation, its presence emblematising remoteness and disparity as it delivers and documents messages that cannot be said or that will not be heard otherwise. This suggests that the object of the letter performatively enacts the very distance it seeks to overcome. The Bidartondo letter shows, most noticeably, a disparity between sender and addressee – on one hand a group of concerned mycologists and, on the other, the world's largest bioinformational resource. It raises the question of how to effectively bring forward a claim or contestation when the addressee is a distributed technological infrastructure such as GenBank. Thus, it brings into relief a key concern in relation to information infrastructures, namely how to challenge "the ways in which software and its attendant categories become 'frozen policy'" (Bowker & Star 2000, p.157).

A letter might be an odd choice for such a challenge. It is, however, not without precedence in the sciences where letters have frequently been used to overcome different kinds of distances. Historically, letters were instrumental in bringing together communities such as astronomers (Eisenstein 1979) or naturalists to fill the gaps in local observations and furnish universally-valid perspectives. Arguably the most famous, if not

most prolific, writer of letters in this vein was Charles Darwin whose global web of correspondents included scientists, cattle breeders, amateur naturalists, explorers, and government officials (Browne 2002). Another kind of gap continuous to be bridged by so-called "Letters to the Editor" which, since the 1950s, fashion the majority of contents in dedicated journals such as *Physical Review Letters*, *Biology Letters*, or *Physics Letters*. These letters appear as letters only in name, mostly taking the form of short, peer-reviewed papers. However, despite foregoing the letter's formalities, their purpose of rapidly disseminating new research findings resonates with the timely conveyance of messages. Accordingly, these letters overcome a temporal expanse that is caused by the lengthy procedures of conventional scientific publication.

A frightful gap

While it performatively illustrates the distance between sender and addressee, the Bidartondo letter's content revolves around a seemingly widening gap – that between data and its description (annotation). As chapter 6 has shown, this description, which includes organism name but also functional identification, establishes the basis for the database record's relationality and, thus, intelligibility. In specifically addressing errors in fungal sequence, the signatories' grievance relates to the accurate representation of their organism, fungi, in GenBank. Therefore, it appears that their concern over annotation points to another gap – between what is *in-here* (GenBank) and that which is *out-there* (fungi). The reason for this gap, the controversy suggests, is *inaccurate* annotation. Annotation, as mentioned previously, refers to the assemblage and attaching of biological knowledge to raw nucleotide sequence and represents "a major challenge facing bioinformatics today" (C. E. Jones et al. 2007). Beginning in the late 1990s, annotation itself has become the object of much research and discussion, as have "models of data storage and distribution that support a continuous stream of end-user submissions, frequent updates, integrated search across databases, and access to data formats (preferably community standards) that are amenable to computational analyses." (Pico et al. 2008, p.e184) Prior to the Bidartondo letter, a number of articles and opinion pieces in scientific journals and blogs had already drawn attention to the fact that GenBank and

other public databases contain annotation errors (Brenner 1999; Bork 2000; Stein 2001; Devos & Valencia 2001; C. E. Jones et al. 2007; Salzberg 2007).

The discussions around annotation, though unlike the open letter carried out on more formal terms, have entailed some expressive rhetoric, highlighting the investment of scientists in the issue of data quality. In his comprehensive and much cited critique, Salzberg (2007) emphasises the potential for inaccuracy by outlining scenarios based on misleading GenBank records in an opinion piece in the journal *Genome Biology*, a key publication in the field.[153] The text commences ominously:

> So you think that gene you just retrieved from GenBank is correct? Are you certain? If it is a eukaryotic gene, and especially if it is from an unfinished genome, there is a pretty good chance that the amino acid sequence is wrong. And depending on when the genome was sequenced and annotated, there is a chance that the description of its function is wrong too. (2007, p.107)

Salzberg evokes different levels of error that researchers can expect from GenBank, affecting both data and annotation. Given the diversity of the latter, illustrated in the previous chapter, types of annotation error range from spelling mistakes (like the "phoshoprotein" in chapter 4), wrong organism designation (like "Angrem52 *Homo sapiens*" in chapter 4) and misnaming (either wrong or inconsistent) to erroneous functional translations and genetic product misidentifications. Directly addressing the reader *qua* researcher, Salzberg sketches out a foreboding sequence of events that must ring frightful to readers of all disciplines. *Science*, too, commences its editorial, which accompanied the Bidartondo letter, in a style similar to Salzberg's foreboding prelude:

> When Thomas Bruns turns to GenBank, the U.S. public archive of sequence data, to identify a fungus based on its DNA sequence, he does so with some trepidation. (Pennisi 2008, p.1598)

[153] Salzberg lists three main domains affected by erroneous annotation: 1) gene models, which may be wrong because of missing genes; 2) gene names, which may either be wrong or inconsistent due to the constant improvement of our knowledge of genes; and 3) false positives, an outcome of the inclusion of predictions in the gene list in cases where the prediction generated by a gene-finding programme does not match any previously known protein.

Here, scientists approach GenBank with "trepidation", hesitant and fearful that any retrieved data can set their research on a potentially spurious path. Painting a similar scenario, this time highlighting irritation, Lathe writes:

> In fact, in searching databases like GenBank, I find this one of the more frustrating aspects, finding a sequence or data and then spending an inordinate amount of time confirming that sequence (or not doing it and getting misleading data). (Lathe 2008)

The affective repertoire that has been mobilised to articulate the issue of accurate data ranges from frustration to irritation and fear. Given such established tone, an open letter certainly appears an apt vehicle for voicing concerns over misannotation. Arriving into a sufficiently charged community, it momentarily intensified the debate. Nevertheless, the Bidartondo letter's wording is scarcely rousing. The letter's first paragraph sets the scene but also serves as an indictment of GenBank and a prelude for the signatories' proposal. Rather than making explicit reference to "accuracy", Bidartondo et al. take issue with "common annotation errors". Specifically, they fault lineage designation, that is, the identification of the DNA sequence in relation to originating organism (its location on the phylogenetic tree), as well as gene function. The latter error, however, is not so much attributed to GenBank than to protein and other "higher level" sequence databases that variously feed from and to GenBank. In accounting for the different kinds of errors (sequence error, common annotation error, functional annotation error), the letter's narrative mirrors the trajectory of annotation, each erroneous step harbouring potentially more severe consequences than the next as it makes its way upwards from nucleotide sequence to cell and metabolism. The references provided throughout the letter embed it in an established argument while making their claim part of a wider discussion across scientific disciplines (ecology, evolution, genetics, bioinformatics, medicine). This suggests that the concern for accurate annotation does not remain limited to fungi alone and has repercussions for a range of organisms. In fact, couching the term "error" within a decidedly viral metaphoric ("propagation") evokes the image of a pandemic from which no organism and no scientists is safe. While such rhetoric sets up the need for urgent, if necessary drastic, redress it also allows readers (with a stake in "molecular biology,

evolutionary biology, and ecology") to join in the fear. The following section will examine the affordances of the open letter in making an issue and raising a public.

The open letter: making an issue

The open letter holds a particular efficacy in relation to this controversy and its issues. Once again, this can be imagined in relation to the capacity of letters to cross gaps. This time, however, this crossing is achieved through the inversion that is the essence of the *open letter*: It turns a private correspondence into a public statement. The sacrosanct confidentiality of the letter, which the open letter instrumentalises, was an achievement of the French Revolution (M. Lyons 1999). And it was there that 100 years later, the open letter reached its apical moment in the form of Emile Zola's *J'accuse* (1898).[154] While the open letter's potencies therefore might reasonably be associated with more conventional political arenas, it has also experienced uptake in the sciences. Here, too, it has been deployed as an effective device for turning a scientific concern into a public issue. Perhaps the most illustrious example of this is Niels Bohr's letter to the United Nations published in *Science* on 7 July 1950 in which he urged for more openness in relation to nuclear development and appealed for non-proliferation as well as peaceful applications for nuclear energy.[155] In the field of genetics, the "Berg letter", published in *Science* magazine in 1974, marked the beginning of a serious public engagement with the control of recombinant DNA (genetic engineering).[156]

These instances illustrate how the letter's capacity to invert is connected to its performative potential. Like any successful open letter, it became a topic *in and of itself*

[154] In 1898 Emile Zola published his open letter *J'accuse*, addressed to President Felix Faure. The letter listed the injustices that had accompanied the trial of Alfred Dreyfus, a Jewish artillery officer falsely accused of conspiring with the Germans in the Franco-Prussian War. It was published 4 years after Dreyfus' conviction, when the affair had lost some of its initial ardour. However, Zola's letter effectively reignited the issue, re-turning the matter into a heated affair and ultimately leading to the exoneration of Dreyfus and the prosecution of his accusers.

[155] "Open Letter to the United Nations", *Science* 112, 7 July 1950.

[156] The Berg letter, "Potential Biohazards of Recombinant DNA Molecules", was published in *Science* 185, number 4148, on 26 July 1974. It led to the Asilomar conference in 1975, which was a key moment in defining regulatory science policy concerning bioscientific research (Jasanoff 2007). According to the National Library of Science, its publication led to a voluntary universal moratorium by researchers on experiments with recombinant DNA (rDNA), the first consensual global self-censorship in the history of science. See http://profiles.nlm.nih.gov/CD/Views/Exhibit/narrative/dna.html. Last accessed: 3 September 2010.

thereby making room for concerns that are not strictly of purely scientific import.[157] The open letter functions as a coordinating device that explicitly seeks to solicit sentiments and assemble a public by mobilising interest and construing "affectedness", in short, by making an affective claim in public. Proust's *A la Recherche* depicts with sociological exactitude the manner in which Zola's letter rearranged the spaces and actors of French society.[158] Suddenly, members of Parisian society had to take sides in public, were forced to distance themselves from friends and families, bring into being new salons, and make allegiances no longer based on ancestors (or property, as in the case of Proust himself) but on their support (or lack thereof) for Dreyfus.[159] In this case, certain gaps were closed when, for example, the Dreyfus affair is discussed across the various social strata – amongst the servants as well as the recruits in Doncières and the guests at Madame de Villeparisis' salon. This not only illustrates the kind of intervention staged by the open letter but also attests to a curious characteristic: Though it eschews the intimacy afforded by privacy, the open letter raises a different kind of intimacy by eliciting affective attachments and detachments in public.

A public comes into being in relation to issues (Marres 2005) as well as in relation to (affective) imaginaries (Berlant 2008; 2004). Here, texts are crucial mediators, crossing distances in search for a public and turning strangers into a social entity (Anderson 1983; Michael Warner 2002) while affording intimate fantasies. The open letter's primary objective then is to materialise an issue in public so as to effect/affect *a* public. Certainly, the open letter makes no secret of raising an affective register. It is always already imbued with affect as it customarily emerges from and marks a moment of distress, of outrage, anger or indignation over a situation that predates the letter itself: the injustice of

[157] To be precise, the letter eschews the formal conventions usually associated with *bona fide* open letters: It omits a specific addressee and shuns personal pronouns. Yet, the scientific community and media instantly regarded it – and responded to it – as an "open letter". Burns (Burns 2008), Hu et al. (2008) and numerous science blogs (Lathe 2008; Porter 2008a; Ebert 2008; Pasigraphy 2008) refer to it as an "open letter". Members of the *Escherichia coli* (*E. Coli*) community, "stimulated by an open letter from a large group of fungal researchers about using wiki models for improving GenBank" wrote their own open letter in response, entitled "The Emerging Worlds of Wiki" (J. C. Hu et al. 2008). This was submitted to *Science*, one month later, on 13 April 2008 (published June 6, 2008).

[158] Discussions of the Dreyfus affair feature in volume 3, *The Guermantes Way* (1920/21).

[159] Zola's letter famously resulted in Zola's own prosecution, which brought to light further evidence that eventually cleared Dreyfus of any wrongdoing. Some suggest that the Dreyfus affair laid the foundation for the topography of Frances' political spectrum, left and right.

indulgences, a wrongful prosecution, the collusion with a fascist regime, or anxiety over nuclear armament and new genetic technology have all sparked passionate public retorts in the form of open letters.[160]

The Bidartondo letter too is an attempt to take a concern, accuracy of data, and install it as an issue. Admittedly, accuracy in annotation, does not by itself provoke quite the visceral imaginary to warrant much public attention. Despite this, it sparked impassioned responses, not least by GenBank director David Lipman and some of my respondents, one of whom referred to the open letter's signatories as "dissenters", a designation customarily reserved for people expressing disagreement or nonconformity in more highly charged environments.

Inversions and issues

The following examines the key terms that have been mobilised in the course of the controversy: wikification, chaos, openness, accuracy, and natural history collecting as historical precedent. In discussing these terms through their entanglements with each other, they reveal meanings and affordances that spill beyond their initial framing thereby assembling a capacity to affect that similarly reaches further than the initial assessment might suggest. Here, inversions, like breakdowns, make present materials that have until then efficiently worked in the background hidden away from scrutiny. Making a private concern into a public issue, the claims voiced in the Bidartondo letter give occasion to further upsets. These concern two contrarian figures, on one hand the database and on the other the wiki. *Prima facie*, the battle line is clearly drawn between, on one hand, scientists and, on the other, GenBank. Accuracy, like reliability and replicability, is an integral, constitutive *quality* of scientific method that rides on a consensual commitment to accuracy as something that can be attained through adhering to certain formal arrangements. Failure to do so is received with disquiet, annoyance and discontent. But as is customary with inversions, a closer inspection of the issues reveals that neither of the

[160] Martin Luther to Pope Leo X in 1520; Emile Zola to President Fauvre in 1898; Thomas Mann to the Dean of the Department of Philosophy at the University of Bonn in 1937; Berg et al. to *Science* magazine in 1975. These and other open letters are brought together in the collection *"Wer schweigt, wird schuldig!"* (Essig & Nickisch 2007).

figures quite sticks to its script, throwing doubt over the neatness that distinctions such as open/closed and normative/affective commonly suggest.

Wikification

In order to avoid the proliferation of mistakes and the subsequent degradation of GenBank, the Bidartondo letter proposes a collaborative annotation process, "an open system of cumulative annotation". This would allow researchers to re-visit records and amend their annotation should the existing one prove insufficient, outdated or inaccurate. The *Science* editorial was quick to label this a demand to effectively "wikify" the database, summarising the proposal as entailing "a community operation, like Wikipedia, in which users themselves update and add information, but not anonymously." (Pennisi 2008, p.1598)

The term "wikification" refers to a structural transformation of an information resource from a closed, static or top-down system into a collaborative, transparent and open-ended process. It derives from "wiki", an open (based on an open source code that can be adapted and edited) software program that supports the collaborative and open-ended creation, editing and interlinking of web pages through a web browser. It also enables transparent documentation of any changes done to a page and facilitates small contributions from a large number of participants.[161] Wikification is therefore commonly seen to foster a timely and distributed production of information by a large number of people – all perceived to be distinctly advantageous qualities that rid knowledge of special interests and allow for the creation of stable, relevant, accessible and accurate data. In that, it is indicative of recent developments in the sciences (not just in the biosciences) that call for more "open" and collaborative practices supported by technologies such as science blogs, open lab books, open access repositories and wikis. Here, the wiki is not just evoked for ensuring quality but the future and viability of research in general as wikis "aim to help biologists turn the data flooding into the large public (…) databases into useful knowledge." (Waldrop 2008, p.22) An editorial in the *BMJ* gave the wiki model an

[161] This has been termed the "long tail" model and has become synonymous with a retail concept popular in e-commerce based on selling a large number of unique items in small quantities (Anderson 2004).

equally warm reception, stating that medical wikis "may be the answer to the world's inequalities of information access in medicine." (Giustini 2006, p.1284)

To make systems more open and collaborative requires group interaction and coordination, individual engagement and other socio-technical arrangements. Consequently, wikification implies not just a "technical" transformation but also a "social" one. Proposing wikification contains the implicit demand for a substantive change in how researchers relate to each other, their community and their research. It brings to the fore capacities usually bracketed from scientific knowledge production for obvious reasons: public quarrels over qualifications, community boundaries and level of participation as well as disagreements over accuracy and relevance of data (bearing in mind that wikis make visible the edits done to a page and the various discussions those may generate) but also, importantly, discussions on how to acknowledge and valorise contributions in the realm of data curation. In a response to the letter, Trey Lathe, a bioinformatician running the OpenHelix blog, picks up on the signatories' concession to traditional academic research validation:

> The wikification of GenBank would be allowing only credentialed editors make changes I suppose, but in reaction to a suggestion for academics with research credentials to enroll [*sic*] as identifiable editors of Wikipedia (in order to increase the accuracy of those articles), one blogger writes: "There's nothing nastier or more tenacious than credentialed scholars squabbling about their area of research." (Lathe 2008)

Drawing on similar discussions around ensuring accuracy in Wikipedia, OpenHelix's concern centres on the difficulty of credentialed researchers reaching agreement over their area of expertise, or, indeed, reaching agreement over their credentials (Michael 2009).

Chaos: the open archive

The proposal to wikify and open GenBank conveys the image of GenBank as a closed, somewhat authoritarian system resistant to change. This further exacerbates the disparity brought about by the disproportion between sender and addressee symbolised by the

open letter. In the *Science* editorial GenBank is represented by David Lipman, director of the NCBI, who is quoted as follows: "'That we would wholesale start changing people's records goes against our idea of an archive', Lipman states and concludes with asserting that any wikification 'would be chaos'." (Pennisi 2008, p.1599)

The "chaos" evoked by Lipman in response to the proposal can take different forms. The letter's proposal effectively turns the GenBank record into an open document to be edited by anyone with a desire to do so. "Chaos" here can appear in the shape of indeterminacy and uncertainty entailed by the record's regression from technical object to epistemic thing. The previous chapter has argued that despite its archival status, the GenBank record, rather than immutable proves to be somewhat unruly and malleable – called up into ever changing constellations and relations. As a digital text in the public domain it can be accessed and reused anytime, as an archival entry within GenBank, however, its existence remains fixed. Wikification suspends the linear temporality associated with the record as archival entry in favour of the immanent future of the record as (potentially eternal) epistemic fragment. It therefore, according to Lipman, unhinges the GenBank record from its determinate temporal trajectory, throwing it back into a process of revision and contestation and indeterminate outcomes.

Lipman does not address accuracy directly but his concern lies with the structural integrity of all records as a whole. Rather than making representations for the demands of fungi or more recognisable creatures, Lipman's response switches perspective and, concurrently, changes scale. In this sense, "chaos" is not so much a response to the issue of data accuracy but an artefact of GenBank's second-order description of its activities, descriptions of its organisation as an *archive* (Strathern 2000). If every group constituted around an organism would make demands of GenBank to accommodate their particular organism's characteristics then this may indeed lead to one kind of chaos. Conversely, changing *archival* records would result in another kind of chaos. In Lipman's plea the database record belongs to history and history is not open for revision. In defending the sequence database as an archive, Lipman turns any alteration of entries into a suspicious act of political manipulation.

For early genetic researchers, open data repositories were regarded the *sine qua non* for the progress of the discipline (Strasser 2011). As far back as the 1960s, researchers were aware that raw data had to be handled in a collaborative and open way – the Protein Data Bank, for example, started making data freely available in 1971. Many regard so-called open data as a driving force and "part of the ethos of genomic research" (Greenbaum et al. 2011). Donald Lindberg, since 1984 director of the National Library of Medicine, remarked that "the tradition of public science pooling the results in a public fashion so that smart people all over the world could interpret them and make discoveries, that really started with GenBank".[162] In the same talk, he averred adamantly that the foundation of GenBank represented a "turning point in access to information". And GenBank does indeed habitually appear as a model for open data in the sciences, such as in another open letter published in *Science* asking to build "a 'GenBank' of the Published Literature" to further research (Roberts et al. 2001).[163]

The wiki model encompasses many of the beliefs voiced by the early genetic scientists involved in the creation of GenBank though they did not have at their disposal a popular, off-the-shelf model for talking about integration, sharing, community participation, public data and access. The early years of EMBL-Bank and GenBank were marked by struggles to make the science itself a community effort, deploying appeals to issues concerning all of humanity (from the Atomic bomb to cancer), building *kludgy* solutions and laboriously reproducing and distributing data.[164] With the progression of bioinformatics and computational biology, *data* in themselves became an efficient foil for enrolling a global community. The data-sharing model for the HGP, for example, not just created a system to facilitate DNA analysis. It also aimed "to form a network of researchers linked by shared materials, create orderly systems of exchange among them, and

[162] Donald Lindberg, *GenBank Celebrates 25 years of Science*, symposium, April 7, 2008, Natcher Conference Center, NIH Campus, Bethesda, MD.
[163] *Science* response to the Roberts letter, entitled "Is a Government Archive the Best Option?" (Editors 2001), betrays a common fear among traditional scientific publishers who fear loss of revenue. It is interesting to note the sly deployment of "government archive". It appears, from their editorial decisions in the present case, that wikis do not harbour the same dangers.
[164] In his talk for GenBank's 25th anniversary symposium (see fn. 162), Roberts recounted how he travelled from conference to conference in the 1970s, presenting and distributing his database of restriction enzymes (REBASE). His secretary had to manually type out lists and send them out. It was so popular that the KGB classified it.

institutionalize rules for ownership of research products emerging from the network." (Hilgartner 2004, p.140)[165] The recent uptake and discussions of the wiki model in bioscience communities is a continuation of the early efforts to collectively develop such open resources and networks. But they can also be seen as a way for dealing with the *aftermath* of these efforts, which continue to produce enormous streams of data.[166] In this sense, chaos, archive and wiki become part of the same history, mutually dependent even.

Labours: Making data open

The reference to historical precedents further emphasises that open data is not only a technical but also a social effort and achievement. The preceding chapters have offered a glimpse into the kinds of work required to maintain a data archive such as GenBank. The responses by GenBank staff to the issue of data quality raised in the Bidartondo letter commonly addressed the issue with reference to these efforts:

> I think it's because of the nature of the database as being a primary data archive. The burden is upon the submitter to submit high-quality sequence and we try, if we find something that we know is… that there are problems with it, that there is vector contamination. The way we go about it is we remove the vector contamination and then we'll email the submitter and tell him we've done it but because the submitter owns the sequence, owns the record. They could actually come back and say "no you're wrong you have to put it in that way". And I'd say in my ten years this has happened maybe twice and in one case we were able to convince the submitter and in the other case we put it out there and added some kind of note, the database staff. (…) So we are trying to address it, it's not easy. (GB1)

GB1's response mirrors Lipman's in that it stresses GenBank's nature as being primarily a data *archive*. But GB1's reply also allows for an understanding of the issue that is more nuanced than the polarised divisions between archive and chaos and their relation to

[165] Hilgartner (Hilgartner 2004) traced the model for the HGP to the cDNA reference libraries, stored in a database and made accessible through the Internet, developed by Günther Zehetner and Hans Lehrach at the Imperial Cancer Research Fund (Zehetner & Lehrach 1994). Lehrach anticipated that sharing the resource of the library would facilitate collaboration across laboratories and in effect establish a "new institution for mapping genomes based on orderly exchanges among a network of laboratories." (Hilgartner 2004, p.137)

[166] See for example Giles' "Key biology databases go wiki" (2007) and Wang's "Gene-function wiki would let biologists pool worldwide resources" (2006) both published in *Nature*.

accuracy. GB1's quote illustrates the work that annotation entails at GenBank: finding and identifying problems, addressing and ameliorating them, corresponding and negotiating with researchers and convincing submitters, which might involve elaborate explanation and justification. Accuracy here emerges as a process of negotiation between submitters and curators but also between curators, submitters and "problems". For the curator, accuracy is not primarily about ensuring a faithful relationship between sequence and organism but about maintaining a dependable relationship between GenBank curators and the submitting scientist who "owns the sequence".

A review of wikis in the biosciences makes a comparable argument by addressing the problem of participation: "Founders enthusiastically put up a lot of information on the site, but the 'community' – either too busy or too secretive to cooperate – never materializes." (Waldrop 2008: 23) Likewise, a researcher suggested that a "culture change" is needed because much of the work required to produce and maintain accurate data cannot "be recorded on your CV".[167] Before the emergence of the wiki, Stein (2001) had identified 4 models of annotation (factory, museum, cottage and party models) that highlight the importance of coordinating scientists, data and work practices.[168] Importantly, Stein's review emphasises the *labour* required by annotation something which the Bidartondo letter, though steeped in community rhetoric, somewhat relegates:

> Currently, primary sequence data are annotated by the authors of those data, and can only be reannotated by the same authors. This is inefficient and unsustainable over the long term as authors eventually leave the field. Although it is possible to link third-party databases to GenBank records, this is a short-term solution that has little guarantee of permanence. Similarly, the current third-party annotation option in GenBank (TPA) complicates rather than solves the problem by creating an identical record with a new annotation, while leaving the original record unflagged and unlinked to the new record. (Bidartondo 2008, p.1616)

[167] Peter Ghazal, Chair of Molecular Genetics and Biomedicine at the University of Edinburgh, at the workshop "Data Sharing in the Biosciences: a Sociological Perspective", National e-Science Centre, Edinburgh, 26 June 2008.

[168] The "factory model", premised on automated annotation pipelines; the "museum model", driven by human expert curation; the "cottage model", a variant on the museum model where curators are recruited from post-graduate fellows and students and work on annotation part-time; and the "party" or "jamboree model", which "puts leading biologists from the community into the same room together with an equal number of bioinformaticians and has them spend a solid block of tie (typically a week) annotating the genome" (Stein 2001, p.501).

Here, the problem for Bidartondo et. al presents itself in GenBank's structural inability to satisfyingly accommodate re-annotation. Mitigating measures such as including links to more topical information or allowing third-party annotation (TPA is a separate database maintained by the NCBI containing annotation) are acknowledged but still perceived to be of a technical nature. Work-related concerns are implied in their suggestion that any edits to a GenBank record have to be "warranted by published peer-reviewed analyses". This proviso is obviously concerned with the problem of quality control: If anyone can edit pages how can we know that this is done with the relevant expertise? For Bidartondo et al. instituting an alternative model still requires recourse to a scientific tradition that the model supposedly seeks to bring to an end. Peer-reviewed analyses and their publication in (recognised) journals are still perceived to be the only reference by which to establish the quality of knowledge (although establishing the quality of data which underlies this knowledge is a different matter). However, unlike the ethos of the wiki, whose quality is assured by something akin to self-policing (contributors continuously updating records without being prompted to do so and without reward), peer-review is still dependent on a culture that favours secrecy and closed circuits but also tangible incentives and returns. But these are also the very features which the wiki model seeks to replace. While the calls to wikify the database might certainly avow the merits of *collective* efforts, the actual materialisation and recognition of *efforts* cannot be taken for granted.

Ostensibly, the controversy presents itself as a clash between GenBank, the monolithic, centralised and closed behemoth resisting change; and the fungi community, which enacts concern for their organism and scientific progress in general, which is young, open to new things and ready to shake up the institutionalised establishment. However, if these models are perceived to be based on espousing certain normative and ethical imperatives, then this opposition is no longer as clear-cut: GenBank, as an open access pioneer, promotes many of the tenets that have now been subsumed within the promises of wikis and "social" software in general. GenBank has in fact championed the cause even before it assumed any coherent form as a "cause" and, as the product of scientists' dedication to the significance of sharing data in public, it is a resource fiercely committed to putting and preserving data in the public domain (which exceeds even the liberties

granted by open access as there can be no copyright or other proprietary claim made over material in the public domain, nor can its reuse be restricted).

Being closer to: affective accuracy

Both parties argue that accuracy is not an intrinsic property of the record as such but an effect of certain socio-material practices. However, they differ in their understanding of accuracy and the kind of practice appropriate for achieving this. Accuracy in the mycologists' proposal for a collective annotation process also turns into a *temporary* achievement, to be revised when and if new, more "accurate" knowledge becomes available. For Lipman, accuracy is an outcome of a cumulative effort whose basis is open data and this is where GenBank's task lies: maintaining the integrity of the archive and sequence universe *in toto*. Contrastingly, for Bidartondo et al. the issue is about accuracy of individual elements. They regard accuracy as an outcome of iterative interventions on the record itself. Accompanying the difference between accumulation and iteration are different spatial imaginations. The Bidartondo letter evokes "public zoological and botanical specimen collections" while Lipman and others provoke the image of the archive. While this makes for different notions of "chaos" it also connects accuracy to locatedness or approximation to a reference value and context. In the case of the Bidartondo letter, this reference value takes the form of the specimen, which returns us to yet another gap.

Error-free identification of functional elements on nucleotide sequence – and its subsequent recording in annotation – depends on experimentation. Currently, only about 1% of such data in public archives is experimentally confirmed (Sjölander et al. 2011). There remains a considerable imbalance between the production of data and its verification and initiatives such as COMBREX, a database launched by the EBI in 2010, have begun to address this. COMBREX operates as a "clearing house" collecting hypothetical entities, such as proteins that were initially inferred through prediction, and distributes them to experimental groups for verification (Roberts et al. 2010). Such confirmation is key to data quality, which suggests that data can be inaccurate in multiple ways: It can be compromised at the point of production (for example, vector

contamination), in the course of accession to the databases (annotation errors) and, once accessed, it often remains unconfirmed and, therefore, potentially erroneous. While the Bidartondo letter ostensibly addresses the issue of annotation, its reference to specimens and functions suggests that their concern extends to this problem of confirmation. Here, the gap identified by Bairoch in chapter 6, indicating our lack of definitive knowledge of proteins, firmly installs itself as a frightful divide between *in vivo* and *in silico*: Bad data is threatening our grasp of reality, or at least in this case, the reality of fungi.

Connections: back to the future

Given the complexities of annotation discussed in the preceding section an obvious question concerns the site of accuracy. If inaccuracy refers to a divergence between an observed, predicted or calculated value and a true value, where, in the annotation process do we locate this true value? Or, asked differently, taking accuracy to mean a truthful representation, what is it that annotation seeks to represent? What makes a representation here more truthful? Completeness? Timeliness? What is more, can accuracy be seen as an intrinsic quality of an object?

Or does it rather describe a capacity? Is it a product or a process?

The Bidartondo letter's evocation of "public zoological and botanical specimen collections" associates accuracy with a closer relationship between researcher and specimen. Collection practices from natural history, therefore, would offer themselves as the most advantageous model for data accuracy by instantiating a closer relationship between specimen and sequence. Orienting their demand in pursuance with the specimen collections of herbaria and museums is not incongruent, as the practice of data curation described in chapter 5 has demonstrated. Neither is it an uncommon claim in the literature on annotation: In his typology of annotation models, Stein (2001) deems the "museum model" (see fn. 168), taken up by model organism databases, the most accurate. Yet, while the proximity between specimen and data might reasonably give rise to more exacting annotation, it is not without difficulties. The following quote by the taxonomist illustrates the problem with the museum model of annotation where accuracy and function diverge:

> You see published in the literature all these codes for museums and herbariums and culture collections and stuff. And people have done this for centuries and no one has ever collected up a database of what they mean. Turns out that there is a good database in the plants but that's it. There's none. And we set about to do one because what we wanted to do is be able to take something (…) and build a hotlink to the specimen (…). This is in Berkeley, California, The Museum of Vertebrate Zoology and they got a link back to our site and this tells you in more detail about where they got the specimen and what they've got … they got some skin and a tissue sample. (GB4)

Like the GenBank record, the specimen does not guarantee direct access to nature. It too needs to be attended to in a fashion that requires specialised knowledge and know-how as well as controlled environments. Like the GenBank record, the stuffed, dried, alcoholised or fossilised specimen requires processing and is a product of sequential arrangement of techniques and technologies in order for it to be eligible for accession into the collection. Similarly, the specimen in a museum collection, rather than a discrete referent, is an arrangement of volatile parts that need to be aligned both internally and externally (see the Grinnellian method described in chapter 5).

As shown in chapter 5, much of the curatorial work undertaken by GenBank can be seen to continue practices that had been honed in the service of botanical and zoological collections. Talking about the efforts undertaken at GenBank to further data accuracy, the taxonomist noted recent developments in relation to the Barcode of Life project, which seeks to establish a global standard for species identification (discussed in chapter 8):

> Ideally, they [the submitters from the Barcode project] are supposed to keep the specimen and deposit it in a museum somewhere so that someone can go back and verify that actually it is from that specimen. (…) (GB4)

From GB4's account it is clear that the specimen and its deposition within a museum constitutes a complimentary yet distinctive practice to submitting sequence to GenBank. Unlike the fate of most sequence data, which is always already set for inclusion in GenBank due to funders' and publishers' requirements, the fate of the specimen remains uncertain as submitters are only "supposed" rather than required to keep and deposit their specimen. Once it has found its way into a museum collection, it then serves the

function of verifying the GenBank entry which it has given rise to. So the sequence data can become once more reunited with its organism, which has been lying dormant "in a museum somewhere" in the search for accuracy. The notion of a "somewhere" that can be returned to and the subsequent evocation of a more localised "here" suggests a very different landscape than the one of described in chapters 4 and 6 where trajectories and connections preceded orientations. The specimen collections referred to in the Bidartondo letter were clearly meant to evoke straightforward coordinates for both specimens and sequences. But for the specimen to serve as a meaningful "true value" in relation to a GenBank record becomes a matter of making and sustaining connections – connections between identificatory codes and museums, between codes and scientific publications, between specimens and their extraction site, between the Museum of Vertebrate Zoology and GenBank, between the practice of collecting and the practice of recording. The resulting pattern brings together the weak and the abstract, the local and the universal, the past and the present.

According to the materials analysed so far, accurate annotation requires the production and maintenance of connections and negotiations between different epistemic practices, objects, scientists, and sites. In this, accuracy has become constructed as an *affected* and *affective* process. While it is seen to be suffering at the hands of GenBank's apparent disregard for the quality of annotation, accuracy also encompasses enough agential capacities to demand attention, cause affront and make others suffer. The next section will examine the controversy's central affected entity, fungi.

Tangled mess

Fungi are the central organism in this controversy and this, I suggest, not necessarily by coincidence. Their allure is further cemented in the *Science* editorial, which is accompanied by a picture of what appears to be an arbuscular mycorrhiza, that is, a fungus in a symbiotic association with tree roots. It is impossible to identify what the photograph exactly depicts: The foreground is taken over by an amorphous composition of what looks like two different textures, one filamentous, the other more solid while the background appears evenly blue-coloured. *Science* chose the somewhat *Lawian* caption

"Tangled mess" for the image.[169] It brings to the fore the organism that has caused the stir in the first place and raises curious questions about the specific affordances posed by fungi in relation to the controversy. In the following, I render some of the capacities of fungi in relation to the controversy with reference to other fungal appearances.

Ionesco's play *Amédée* (1965) offers a striking depiction of a fungus intervening and confounding and, hence, a relevant figure to explicate some fungal capacities:

> Amédée: A mushroom! Well, really! If they're going to start growing in the dining-room! [*He straightens up and inspects the mushroom.*] It's the last straw! ... Poisonous, of course!
>
> [He puts the mushroom down on a corner of the table and gazes at it sourly; he starts pacing about again, becoming more and more agitated, gesticulating and muttering to himself; he glances more frequently towards the door on the left, goes and writes another word, which he crosses out, then sinks into his armchair. He is worn out]

Amédée's frustration and annoyance, present in both words and actions, are directed at the "mushroom" which he cannot make sense of. It has appeared in a place (the dining room) where it shouldn't have appeared. Amédée's inability to comprehend in the face of this arrival causes him agitation. What's worse is that the mushroom is inverting scales, no longer eating something away but making it horrendously large. It anticipates the palpable frustration voiced by the GenBank taxonomist in relation to fungi ("arghhh..the fungi", see below). It also points to fungi's propensity for scaling up (or down) that was suggested by James Hanken when he asserted that the problem of misannotation might extend to "more recognizable creatures" (Pennisi 2008). But it also reflects the confounding nature of fungi themselves: They are now closer related to us humans than to plants, they form the largest living organism in the world, yet only 5% of the world's fungal diversity is formally known.[170]

---

[169] The messiness of reality is the topic of Law's *After Method* where he suggests "simple clear descriptions don't work if what they are describing is not itself very coherent" (2004, 2). It seems to work quite well here though.

[170] The world's largest living organism is the *Armillaria ostoyae* of the honey mushroom genus which covers 880 hectares (8.8 km^2) of the Malheur National Forest, in eastern Oregon. It has been growing for approximately 2,200 years (Filip & Ganio 2004).

Fungi in the form of yeast (especially *Saccharomyces cerevisiae*) have played a central role in genomic research. Despite (or perhaps because) of this importance, they habitually appeared as complex and troublesome entities in interviews with GenBank curators and, especially, taxonomists. Taxonomists are located on the same floor as curators though the offices of the 2 taxonomists I interviewed were more reminiscent of the space occupied by the "information plumber" described in chapter 5: They seemed "lived in", containing lots of books and, in one case, music CDs. There were Apple desktops instead of the PCs that dominated the other offices and unlike curators, taxonomists distributed assignments according to expertise (incidentally, their fungal expert had left a couple of month prior to my arrival). It was there that I encountered the most impassioned views in relation to the controversy. This is perhaps not surprising given that the Bidartondo letter's criticism is specifically directed at "erroneous lineage designations in GenBank" and, therefore, at taxonomy. The taxonomy group at GenBank maintains the Taxonomy database (including a taxonomy browser and additional resources), a sequence-based (phylogenetic) taxonomic resource that contains the names and lineages of all organisms represented in the INSDC with at least one nucleotide or protein sequence.[171] Explaining their work routines, one taxonomist remarked:

> However, sometimes I get a new genus, particularly in the fungi, arghhh the fungi! (…) The fungi are in total ferment (*sic*) and last year in fact a major work was published on fungi that really just completely re-did the traditional classification of fungi which had been recognised again since the molecular revolution, people have recognised that the traditional classification of fungi has nothing to do with their evolutionary relationships. (GB6)

Fungi, according to this GenBank taxonomist, are a very recalcitrant group that invites contestation and agitation, or as the taxonomist puts it, "ferment" (ironically, fermentation of course indicates a biochemical breakdown most often triggered by yeast). The breakdown that GB6 is referring to relates to the naming and classifying of fungi in the organismal order of things. Fungi had traditionally been subsumed under the kingdom of plants (kingdom *plantae*) and were only recognised as a separate kingdom in the late

[171] http://www.ncbi.nlm.nih.gov/taxonomy

1960s (Whittaker 1969). On the basis of phylogenetic studies, fungi continued to defy classification. Molecular analysis proved that fungi are more closely related to animals than to plants and are therefore now grouped together with animals in the monophyletic (sharing a common ancestor) group of opisthokonts (a broad group of eukaryotes that includes animal and fungus kingdoms). It appears then that fungi have a history of escaping order and can evoke, as in the case of GB6, aggravation and frustration.

But more than just eluding the classificatory schema, fungi's ambiguous in-between position has uncomfortable consequences for other species. Guiding me through GenBank's taxonomic trees on her computer screen, GB6 clarifies:

> So here are the three main domains: archaea, bacteria and eukaryota. And then there are all these group collectively called algae, here's fungi/metazoa. All animals are sister-grouped to fungi. Among other reasons this is why it's very hard to treat a fungal infection if you develop a fungal infection, because the problem is that the chemicals that will kill fungi will also kill animal cells. So they tend to be very toxic. (GB6)

Navigating through the complex arboretic structure while clicking on taxa to reveal their organismal units, GB6 demonstrates how the genetic nearness between animals and fungi is represented by means of the taxonomic tree but also provides an account of the effects of this proximity: Remedies against fungal infections will, naturally, also attack its "animal" host tissue. Continuing her exploration of GenBank's phylogenetic tree, GB6 recounts:

> So people generate all these trees but the final step is to actually formalise – formally recognise these using the rules of nomenclature [taps on the book]. Particularly for a group of fungi called the *Basidiomycota* that hasn't been done, so what we have is this huge horrible mess and there are certain...there are these yeast-like forms called the "torula", see, it pops up in all these different parts of the tree [clicks along the tree] and that's because you have something that fits the basic morphological characters of *torula* which is basically a red yeast but you look at the sequences and you realise "Oh! They actually belong...they are completely unrelated, they're in different parts of the tree!" (GB6)

Fungi here are not just obstinate and hard to classify entities but they also have the capacity to surprise and generate a "huge horrible mess" where morphology escapes evolutionary relationships, and phenotype eludes genotype. Translating the closeness from taxonomic tree on screen to referent situations off-screen turns this proximity into a veritable cross-species interference: Fungi depend on other organisms for digestion which is why they often appear in what is called mycorrhizal relations, symbiotic associations between a fungus and another organism – the tangled mess of the *Science* caption.

In the course of the controversy, too, different species become entangled. The *Science* editorial contextualised the issue by quoting James Hanken, director of the Museum of Comparative Zoology at Harvard University, who claimed that "the problem extends far beyond fungi, to much bigger – and [more] recognizable – creatures." (Pennisi 2008, p.1598) Since fungi hardly rouse much public interest, "more recognizable creatures" are enrolled whose genomic misrepresentation could lead to presumably more recognisable consequences. The signatories, too, resort to crossing species-boundaries in bolstering their argument by implying the epistemic alignment of their respective microorganisms with organisms of a higher, more familiar order. They note that "for organisms such as fungi, which are notoriously difficult to identify, up to 20% of DNA sequence have erroneous lineage designations". This indicates that a potentially larger segment of the organismal landscape is indeed affected by such errors. The fungi's mutualistic relationship continues to effect associations *ex vivo*: The wrongly annotated fungal sequence provided the *E. coli* community with a strong enough bond to make representations for bacterial sequence, occasioning their open letter (J. C. Hu et al. 2008).[172]

Fungi habitually confound and escape orders, taxonomic and otherwise. This is

[172] The Hu letter offers a more reconciliatory tone. It recognises the shortcomings of the GenBank model, most notably that "individual curators cannot fully encompass the collective expertise of the larger scientific community". And whereas it highlights a number of community-organised wiki-based annotation resources, it also questions the efficacy of radically altering the GenBank approach *per se*. Rather than further faulting the latter, Hu et al. appeal to both protagonists, the mycologists and GenBank, to work on a collaborative solution based on community-developed tools and GenBank-support of these tools. Appropriately, the Hu letter was simultaneously published on the EcoliWiki, a "Wikipedia for *E. coli*" that facilitates community annotation. At http://ecoliwiki.net/colipedia/index.php/Letter_to_Science_about_wikifying_genome_information. Last accessed: 23 September 2011.

congruent with the difficulty to locate and place fungi in the taxonomic tree, or fungal sequence in the sequence space, with the frustration and incredulity of encountering fungi where they shouldn't be and also with the general propensity of fungi to appear in the strangest places such as the back wings of male ladybugs, cow dung or in absurdist plays and even Nietzsche.[173] But as well as eluding sense and research, fungi have an exceptional propensity for attachments. Given that they digest food externally, their relation to their environment is crucial for their survival. Unlike the life of plants, which in the words of W.H. Auden, "is one continuous solitary meal", the life of (most) fungi is one of a continuous dinner party (1991).

Fungal representations

It is therefore not surprising, given their ability to confound, that the controversy emerged around fungi. It might also suggest that mycologists, used to dealing with defiant creatures, are arguably more versed in making representations for their organism.[174] In relation to the open letter, the signatories were summarily referred to as the "fungi community" by respondents and commentators alike. As such, they are also an "epistemic community", established by means of attachment to a certain knowledge (anchored to fungi) and specific cognitive and procedural orientations to generate and relate to this knowledge.

With the dawn of molecular biology the organism became increasingly atomised, its scope condensed to the smallest of units and its diversity reduced to a small set of model organisms. It is not uncommon in molecular biology to constitute identity around model organisms: the mouse community, the worm community, the *E. coli* community, the

[173] In Nietzsche's *Also sprach Zarathustra*, the fungus is likened to "the small thought" that creeps along unseen, ensnaring bodies, and that wants to be *nowhere* until the body is rotten and everything is consumed by its spores. I would like to thank Isabel Waidner for pointing out the fungus in Nietzsche.

[174] Take for example this excerpt from a memorandum presented on 4 February 2008 to the Lords Select Committee on Science and Technology by the European Mycological Association in which they condemn the "very poor public awareness of the unique status and importance of fungi. Fungi belong in their own separate biological kingdom. They are not animals or plants, and they do not fit into the vague and generalized category of micro-organism. But without them, life as we know it on this planet would not be possible. Despite a campaign by the British Mycological Society, there is, more or less, no teaching about fungi in the national school curriculum. The result is that future voters, politicians, senior civil servants and other decision-makers come out of school with no knowledge that fungi even exist, let alone that they might be important."

Arabidopsis community or the *Drosophila* community. Regardless of the epistemic thing that scientists may investigate (reproduction, infection, cancer, depression etc.), it is the technical object *qua* model organism that can often establish the first order of affiliation.[175] And just as their recalcitrant organism, the signatories too have become "dissenters" (GB6), suggesting a co-constitutive dynamic between scientists and fungi.[176]

What are the representations made by the signatories of the Bidartondo letter? Primarily, they seek to preserve the representational capacities of fungal sequence records in relation to fungi *out there*. Bidartondo et al. regard accuracy as a correspondence between specimen and the GenBank record. But while for them the lack of correspondence is indicative of a widening gap, for others there is little divergence between fungi *in vivo* and fungi *in silico*: They cause the taxonomists trouble in any shape or form. Also, in some cases, where for example environmental or prehistoric samples are sequenced, the sequence is all that remains – the originating fungus has long since disappeared like the uncultured bacterium of chapter 6. Through the techniques and methods applied to organic material, multiples emerge – the fungi, the fungi DNA/RNA, the purified DNA/RNA sample, the vector containing the DNA primers, the millions of copies of the initial DNA fragment (for microarrays), rDNA, or the GenBank record. And a range of object worlds unfolds, each harbouring different affordances, practices, (social) relations and norms.

In this controversy, the signatories rally behind their organism, fungi, in construing it as an affected entity. This role endows fungi with certain agential capacities: Here, they have the ability to thwart research progress when wrongly annotated. They can also elicit attachments across species-boundaries by means of concern, fear, sympathy or

[175] There may be a number of ways to expand on the gravitational pull of model organisms and the resulting constellation of communities: certainly the affordances associated with certain organisms give rise to specific material and semiotic practices that do not translate well across species and delineate boundaries of intelligibility and/or functionality and, thus, community. Similarly, one could also look towards the general dogma of experimental science such as replicability and standardisation indirectly giving rise to entities such as Jackson Laboratories, which is the unrivalled market leader in laboratory mice supplies. Using the same mouse resource surely fosters certain communitarian traits, such as common concerns or language games among researchers around the globe. As do the practices around the care and foster of organisms such as mice, maize and fruit flies. Either way, one can, starting with a model organism, infer a set of practices that become distinguishing features of certain epistemic communities.

[176] One researcher remarked that the mouse community was more "cagey" in relation to sharing data and results.

interest. The most basic function of annotation is to make the sequence intelligible and therefore "attachable" so that it can engage in bioinformatic processes and, in this case, controversy. Hence, annotation affords the sequence to become engaged with others. The controversy can be seen as a figuration that formally establishes and enrols a series of distances and proximities, detachments and attachments. GB6's frustrated references to fungi perhaps betray a more affective relationship with the process of annotation as the GenBank record is viewed in direct relation to the specimen, reverting the scale from molecular back to organism level. Ellis and Waterton (2005) suggest that such an attachment is common in taxonomy, a scientific practice historically entangled with the passions and rituals of amateur naturalists. Proximity becomes a resource in the controversy, serving to enrol the *E.coli* community, which also establishes relations between fungi, bacteria and humans by highlighting other wiki-based efforts (such as Human Proteinpedia, the human protein wiki). At the same time, it is often proximity and attachments that make fungi and fungal sequence so recalcitrant and error-prone in the first place.

Affective gaps

A letter, Simmel (1908) observed, enters a space of many uncertainties once it leaves its writer. As it travels from writer to reader, interception and manipulation are possible. Similarly volatile is its reception by its addressee: Will they read the letter's content the way that the sender had meant it? This "gap between content and effect" constitutes a space of suspense and affect (Massumi 1997, p.218). Likewise, this controversy relates (to) *anxiety* – anxiety over data quality and the effects of bad data but also concern for organisms and their proper representation. The controversy's positions described above comprised in parts very emotive undertones: Salzberg's menacing prelude, Lathe's frustration, the (reported) brusqueness of Lipman's defence but also the form of the open letter itself furnished quite an affective register.

Upsetting the order of open and closed (and public and private) within certain normative arrangements, whether epistolary conventions or scientific communication, certainly lets tempers run high. The normativities underpinning the making and doing of

science do not escape affective attachments but, in fact, require and often solicit them. The Bidartondo letter itself sought to convey an issue that readers should care about. Normativity is also "aspiration", that is, "an affect, a sense of something, organized by but not inhering in its conventional objects." (Berlant 2008, p.266). Equally, the typical tools of scientific enquiry in the biosciences such as the GenBank record are enmeshed with this "sense of something" – accuracy, replicability, trustworthiness, authority, relevance – that is enacted by means of disciplinary techniques but that also finds expression through aspiration-as-affect. In this controversy "accuracy" rather than constitutive of proper scientific method is better understood as an affective mediator for imagining a re-arrangement of relations between people and organisms via contestations over GenBank records. Because of its indeterminate meaning, "accuracy" then is not so much a property of a matter of fact but a legitimate vehicle for collecting and conveying different concerns.

The Bidartondo letter continues to be cited, both in relation to mycological research (Stockinger et al. 2010) and, more generally, genomic resources (Renfro et al. 2011). Complaints about data quality in the fungi community and beyond persist but many have turned into actions. Data derived from mycorrhizal fungi, a recent article complained, particularly suffers from the lack of proper species identification: "While the diversity and geographical distribution of host plants is relatively well-known, the ecology and biogeography of symbiotic fungi remains poorly understood due to their cryptic nature and high costs of identification." (Tedersoo et al. 2011) Here, the more recognisable creatures are doing just fine while the fungi continue to be disregarded. For the authors, however, this serves as a prelude for their solution: Having downloaded data from the INSDC, they then assigned phylogenetic lineage and added metadata. These enhanced records are now available at UNITE, a database set-up in 2011 for the molecular identification of fungi as "a response to the difficulties facing environmental samples of fungi to species level using molecular data and the major international sequence databases."[177] The rise of so-called "wikiomics" (Waldrop 2008) follows a similar route in beginning from the assumption that most data in general public repositories such as

[177] See http://unite.ut.ee/. Last accessed 21 June 2012.

GenBank and EMBL-Bank demands further work. Consequently, the issue is no longer exclusively directed at data quality but at *expectations* of data quality.

By addressing the problem via an open letter, the mycologists made use of a device that transports concern for a particular state of affairs to the attention of a public. In doing so, they sought translate a matter of fact into a matter of concern (Bruno Latour 2004). The open letter here becomes a device that mediates this translation: It solicits an affective space and an affective public. Connections between problems are in parts upheld by recourse to affect – certain things become enrolled within the controversy because they are quite literally moved to do so. Discontent and frustration as well as care move a group of mycologists to write a letter. Similar sentiments prompt declarations of solidarity and comments from other researchers and GenBank users. Irritation and a sense of being misunderstood provoke an annoyed response by GenBank. Recalcitrance and obduracy ensconce the strange creatures of the fungi kingdom and the effusive GenBank record.

More recently, open letters have been used to publicly announce the commencement of large-scale sequencing projects. But they retain some of the qualities mentioned in this chapter. For example, the open letter published in *Genome Biology* to announce the sequencing of three crocodilian genomes also intended to raise a community: "We invite (…) the broader scientific community to access and make use of the draft assembly and raw read data that we have produced." (St John et al. 2012). The German writer Jean Paul called books "letters to friends". For Sloterdijk (1999) this sentiment exemplifies both the triumph and ultimate limitation of humanism: the necessity of a literate, human audience at home in shared texts. In the present case, the open letter serves as a way to enact some unexpected capacities and associations: evoking care, solidarity and concern across species and perhaps make some radically different (fungal, bacterial and otherwise) friends. here, the overcoming of "human separation" that Decker associates with epistolary practice can also overcome the separation between humans and more-than-humans.

Chapter 8. Imagining prepositions for the sequence universe

In this final chapter I revisit the premises established in the previous chapters with a view to articulating an imaginary for the sequence universe. I suggest that the stories and interpretations assembled in this thesis suggest different enactments of *integration*, derived from Barbara McClintock's practice of *integrating*. Here, integration is uncertain and vague, deeply intertwined with situated practices. It thereby conveys the indeterminacies and absences that accompany the orderings effected by the sequence databases. Each chapter describes an enactment of such *doubtful integration* although each instance is followed by a different preposition: Chapter 2 outlines integration *towards* ethnographic and analytical access to the databases. Chapter 3 problematises integrating *with* databases, while the journeys recounted in chapter 4 describe integrations *through* the sequence universe. Chapter 5 shows curator and other staff at the databases integrate *amidst* data streams. The next chapter, which analyses two database records, presents these bioinformational artefacts as texts that by combining manifold entities afford integrations *beyond* the sequence universe. Lastly, the controversy analysed in chapter 7 offers the final prepositional manifestation, integrating *between* radically different ontological entities: letters, fungi and accuracy. In this final chapter, I detail these enactments of vague integrations across the thesis with reference to Michel Serres' philosophy of prepositions. I conclude by suggesting some ways in which this can contribute towards inventive problem-making in the realm of biodiversity.

Introduction

EMBL-Bank and GenBank constitute novel sites for doing science but they also espouse many of the features and forms associated with more traditional settings of scientific practice such as archives, laboratories and museums. This thesis has explored the databases ethnographically as *particulars*, situated in place and time, while also drawing

out figurations towards a database imaginary. They are *in silico* discovery environments where combinations, alignments and playful exploration can lead to unexpected encounters and surprising conclusions. They also constitute places of work that bring together different disciplines – geneticists, molecular biologists, taxonomists, programmers, web developers, librarians and metaphorical plumbers. Many imaginations, visions and affects guide their practices and interactions with data. In turn, these data are revealed to have rather material properties, much like the specimens and artefacts handled by museum curators.

My initial question – What happens at the databases? – has yielded a busy hinterland, containing many different things: viruses, trails, contestations, collaborations, data pipelines, taxonomies, data curators, open-plan offices, portakabins, self-made booklets, programmers, laboratory mix-ups, DNA sequence, research programmes, model organisms, uncultured bacteria, whiteboards and open letters. It has also rendered different journeys, moving from Hinxton to Bethesda, art history to science studies, genomes to proteins, from one species to another while traversing resources, literatures, texts, landscapes and oceans. Moreover, the question unravelled its object, the databases. No discrete object or coherent system, they are entangled amidst a plethora of bioinformational resources and tools that make it difficult, if not meaningless, to draw definitive boundaries. The answers occasioned by the question suggest a curious symmetry: On one hand, making sense of the databases, as a social scientist, entails rendering present and integrating a multitude of relations – in the course of these relational renderings, the sequence universe emerges. On the other, the databases make sense of data flowing in and around them by means of relating it to many different shapes and forms: organisms, scaffolds, the world out-there, biological vision, other data, visualisations.

In sum, these arrangements of relations suggest that making sense *of* and *in* the sequence universe is predicated on enactments of *integration*. In fact, integration of data, resources, standards, algorithms, knowledges, new kinds of data and old paradigms (and

vice versa) stands as a key achievement and remains the biggest challenge.[178] But "integration" here resonates with more, or indeed *less*, than the term might suggest. In chapter 4, Porter's final hypothesis about the appearance of mumps in a yellow fever mosquito involved retrotransposons, the "jumping genes" discovered by Barbara McClintock. In Fox Keller's biography (1983), McClintock, on several occasions, refers to her work as "integration". At the same time she describes trying to understand chromosomes in maize as "integrating". "Integration" here refers to growing and tending to maize *in the field* as well as to observing chromosomes *under a microscope*, discovering retrotransposons on the maize genome and making sense of it all. It is an effect of embodied heterogeneous practices which negotiate different scales (macro and micro), spaces and knowledges. This integration is not necessarily predicated on a prior framework. Neither does it strive for total capture. Viruses, discoveries and ethnographic travellers reveal enactments of integration that differ from the usual practices of integration. Instead of depending on an *a priori* ordering into which entities are integrated into, they allow for integration to be temporary, uneasy and inventive. As such, it is vague, never fully formed, and entailing different ways of relating.

Prepositional relations

The artist Nancy Holt's *Sun Tunnels* (1973-76) consist of four large-scale concrete culverts that lie in Utah's Great Basin Desert like big stranded whales basking in the hot desert sun. There is undoubtedly a purpose to their arrangement, in relation to one another but also in relation to the wide expanses that envelop them. Up close, the *Sun Tunnels* are huge, towering over human and more-than-human visitors. Inside, they reveal a set of small holes whose arrangement relates to the celestial constellations Draco, Perseus, Columba and Capricorn. Like much of land art (or *Earthworks*, as the artist Robert Smithson called it), the *Sun Tunnels* turn the skies and lands into both stage and audience but they have also become part of the topography of the desert like the cliffs, peaks, rocks and limestone formations.

[178] EMBL-Bank puts it as follows: "The need for data integration is now embedded in our culture. In addition to the multi-group collaborations described above, there have been numerous enhancements that make individual resources work more seamlessly with each other." (EBI-EMBL Group 2012, p.5)

There might be nothing farther removed from DNA sequence databases than Holt's *Sun Tunnels* but I suggest that they serve as a model for thinking about the relations that I have rendered around and through the databases in the previous chapters. On one hand, their site and sights are carefully crafted – there is nothing incidental about their alignments with each other, the environment and the stars above.[179] On the other hand, the arrangement is continuously undergoing change, left to the vagaries of weather and climate, to processes of decay and disintegration but also processes of invention: No visitor will ever see the *Sun Tunnels* in quite the same way. As visual reference points, they also afford the observation of change in the surrounding environment – the distinction in scenery depicted in the many photographs of the *Sun Tunnels* is indeed striking. They are large, yet the desert and sky around them are larger and the stellar constellations they relate are even larger.

Precast concrete culverts are mainly used for drainage and, hence, are not usually found lying so ostentatiously above ground. They are part of the invisible infrastructures that keep water, waste and sludge well away from our senses and constructions. And what to make of their presence in a desert, where water is manifestly absent? They are both, document and monument, figure and ground. They work as mediators between spheres and scales without settling *on* or *in* one. In fact, the *Sun Tunnels* offer many more prepositional engagements: One can be in front, amongst, above, inside and outside, or astride. One can walk alongside, through or towards them while looking over and beyond at the clouds and mountain ranges in the distance.

What emerges through the sequence universe is a polyvalent topology that connects digital, imagined, organismal, textual, mineral and historical entities. Like the *Sun Tunnels* and the doubtful guest, these connections and relations are often strange and strangers, and as its odd appearance and positions (inside a tureen) suggest, these relations are best imagined (and made) by way of prepositions (Serres & B. Latour 1995): *inside* the sequence universe, *on* the mumps genome, *through* proteins, *into* VectorBase. It is an incommensurable space of disproportioned scales where the universe meets the

[179] The catalogue *Nancy Holt: Sightlines* (2011) provides a detailed overview of Holt's practice and includes her sketches and notes which accompanied the making of the *Sun Tunnels*.

seed. "[T]he world", Serres writes, "passes from landscape to panorama, from local to universal, rambling changes into method and vice versa" (Serres 2009, p.305). The sequence universe perhaps resembles what Serres' has imagined as the space of angels.

For Serres, prepositions get to the heart of relationality (Serres 1982; 1995a; Serres & B. Latour 1995). Similarly, Michael (2010) argues that prepositions help us get an analytical (and literal) grip on the differential production of spaces and situations of material and semiotic exchanges where bodies and matter intermingle with data. Prepositions make connections but not indiscriminately so. Some, such as "with", "amongst" or "towards", make for potential or indeterminate relations. Others, like "in" or "of" engender more definitive or actual ones. This thesis has described EMBL-Bank and GenBank in terms of materially enacted practices – journeys, curation, database records and controversy. In doing so, it has observed the conventions of science studies of exploring phenomena materially and focusing on the objects that populate the databases: architecture, design, self-made booklets and coding tables, sequence data, genomes, contigs and scaffolds, annotation, and computers. Concurrently, these materials are presented embedded in tasks, conversations, commentary, histories, texts and situations.

The preceding chapters have accumulated many instances of such material relationality, thus once again abiding by the lessons of science studies in *unbracketing practicalities* (Mol 2002). In the course of this thesis, EMBL-Bank and GenBank are rendered as "relational effects" (Law & Mol 1995, p.275). Here, I wish to consider some of the qualities of the relations crafted in the previous chapters. And for this, the *Sun Tunnels* and their prepositional plenitude offer a pertinent departure point. As pointed out above, their concrete presence betrays ambiguous positionings. Existent between scales and modes (change/stasis, absence/presence), they suggest a topology consistent with manifold passages and connections. Objects and shapes appear incidental – what gives cause to wonder and contemplation are the way in which the sequence universe "clots" (Verran 2001) and folds.

This is instructive for thinking about the relation between the databases and the sequence universe but equally apt for representing the relation between omic datascapes

and life matter.[180] Metagenomics and their attendant high-throughput sequencing projects have generated a new data stream that, according to one researcher, "has now become "real" indeed! (…) In a sense, the genomics-bioinformatics nexus has now spilled into the real world." (Ouzounis 2012, p.3) The development of experimental biology and the wet lab had populated biology with strange creatures and novel routines. For science studies, the *realness* occasioned by experimental biology in the laboratory put into question the fundamentals of modern metaphysics – the divide between nature and culture as well as the nature of the "real". The latter came to be seen as an effect of relations: There are no autonomous discrete entities that precede their relations, instead "the relation is the smallest unit of analysis" (Haraway 2003, p.24).

In accordance with the lessons drawn by science studies, I suggest that the dry labs and entities of bioinformational research – collectively referred to in this thesis as the "sequence universe" – continue to entangle and confound divides. While the spill into the "real world" diagnosed by Ouzounis is perhaps more readily understood as a move towards biomedical application, it also points to intensified entanglements between *in silico*, *in vivo* and *in vitro* worlds. The formation of new biological disciplines such as systems biology, has drawn attention the "*movements back and forth across the machine-living organism border*" (Fujimura 2005, p.196). These movements, what Rosengarten (2009) describes as "traffic" between bodies and data, are performatively productive. Bodies are intertwined with information. This argument is gaining more and more traction as the tools and methods around genomic data production, distribution and analyses find wider applications. We are still in the early stages of critically appreciating the way in which data and information work *on* and *with* and *through* our bodies and worlds. Here, a more refined register of prepositions is required to account for spatial and *temporal* (Michael 2010) relationalities. In chapter 1, I suggested that EMBL-Bank and GenBank are concerned with making sense of and amidst proliferating data streams. Like the *Sun Tunnels*, they cut an imposing figure. Yet, in the ethnographic encounter they

[180] Eglash (2011) suggests that the "gene space" of metagenomics gives rise to a figure/ground switch where "species are merely temporary vectors within it". This is certainly one key point of contention in the ardent discussions between systematists and *barcoding* initiatives, that identify species on the basis of conserved regions, such as 16S rRNA described in chapter 6. For a particularly belligerent defence of the former position see (Ebach & Carvalho 2010).

reveal themselves to be much more nimble entities, demonstrating how *databasing* is entangled with uncertain and unsettled (unsettling) activities of sense-making.

Vague integrations

For many scholars the "century of the gene" (Fox Keller 2000) and the concurrent move from field to laboratory, both dry and wet, was distinguished by a symbiosis that very literally conflated language and bodies, namely the concerted union between information science and biology. Here, the genetic database and biobank have emerged as formidable devices for turning bodies into viable data. Technoscientific critiques have cast the database as the new archive, "the register of epistemic arrangements, recording in its proliferating avatars the shifting tenor and debates around the production and ethics of knowledge." (Arondekar 2009, p.2) But this "database logic" forgoes the database as a "figure seen twice" – concept and lifeworld, as explanatory tool and as phenomenon to be examined. This is not to say that the database logic is not a valiant construction but that, to to be an effectual critical notion, it needs to be seen alongside concrete databases realities, not all of which correspond to this, or indeed *a*, logic. The use of databases to collect, preserve, document, manage, project and combine bodies and information demands close scrutiny. In this case, it is instructive to observe that "normativity is not to be expressed in terms of governance by rules or de facto regularities" but "instead involves a complex pattern of interrelations among performances through time." (Rouse 2007, p.8) Thus, in order to make present the normative effects of specific database projects it would appear prudent to account for the local orderings, situated practices and generally messy *hinterland* that maintains and distributes the database and its products (Law 2004).

This thesis concerned itself with drawing some of the contours and contents which constitute these lifeworlds at EMBL-Bank and GenBank and the sequence universe. The previous chapters have traced different "moments of assemblage" (Verran 2009, p.170), relational unfoldings of objects and subjects, matter and matters. This, as I explicated in chapter 1, was done with the aim of circumscribing a "database imaginary" to complement the database logic. In that I have remained faithful to the critical programme and performative proclivities of science studies, which take technologies to be sets of

contingently stabilised heterogeneous materials and social relations. Accordingly, database conventions – the database space, the data input and accession, the database record and data accuracy and quality – were refracted through material-semiotic practices and situations: ethnographic and scientific discovery journeys, data curation, the reading of database records, and epistolary contestations over proper organismal representation.

The four empirical chapters that form the body of this thesis have worked toward "infrastructural inversions" that make visible some of this hinterland (Bowker & Star 2000). In setting up these inversions, I introduced the figure of the laboratory in chapter 2. This has served not just as the archetypal site for doing science but also for *studying* science, giving rise to the genre of laboratory studies. The laboratory, as both thing in the world and theoretical register, offered an appropriate interlocutor for unsettling the database logic. It also served to illustrate how EMBL-Bank and GenBank could be considered as *sites* for doing research, or rather, as sites which afford novel ways of experimentation and research – for both the social and the biosciences. Lastly, interpellating the laboratory worked towards making the databases ethnographically accessible.

Retaining the explorative methods of laboratory studies, chapter 4 mapped the sequence universe via different journeys. Manifold landscapes emerged in relation to travels and travellers as well as inhabitants and guests, doubtful and otherwise. The sequence universe does not offer a coherent panorama that can be taken in and processed according to spatial and representational conventions. Moving through genomes, campuses and algorithmic alignments served as a foil for introducing some of the many bioinformational resources and tools that in concert constitute the sequence universe. On one hand, it afforded certain conventional movements such as "following", "trailing" or "moving along". On the other, these movements connected scales and worlds in quasi-viral patterns as the journeys continuously unfolded new landscapes.

Viruses jump and re-configure scales and relentlessly rework what it means to be inside and outside. Some traces, like spelling mistakes or wrongly designated proteins, remain put (suggesting that the messy practices and ontologies of the laboratory continue their lives in the sequence universe). Yet, as further data is added and more tools

developed, the connections between these traces and others are forever reworked. Visualisation tools such as genome browsers make present additional, often excessive, content and open transversal routes along genes, proteins and taxonomies. In addition, devices such as handmade booklets and coding tables afford more local orientations that connect the sequence universe to affective and intimate spaces (such as concern over workloads, individual visions, ethnographic wonder).

These spaces find further exploration in chapter 5 as it examined the work of curators and developers at the sequence databases. Much care and oversight go into the building and maintaining of the databases, data, tools and other bioinformational resources. The work routines described encompassed reading and re-reading sequence submissions, consulting relevant literature, matching and verifying submissions but also plumbing and having vision. They deploy a range of bioinformational tools, consult a vast breadth of literature and continuously discuss and exchange with colleagues and submitters while also talking to the sequences themselves. Staff at the databases write guides and handbooks for submitters and users and publish scientific papers. And they constantly collaborate amongst each other – curators, developers, programmers, taxonomists – on how to make data more intelligible, more versatile and more easily findable and (re)useable. The primary apprehension or disposition of the curators' work can be understood as a "feeling for the sequence". Integration here is done with much care, continuously negotiating demands from within and beyond the sequence universe. An important moment in curation is "figuring out what it is and where it should be" (GB12). Resolving the ontological status of something goes hand in hand with finding a place for it and if the current provisions do not suffice then we "make room and end up with new databases". The conversations and interviews with curators and developers detailed in chapter 5 revealed how looking after biological data entangles molecular, biological, technical, affective and semantic concerns. Despite taking place in front of computer screens, curation exhibits direct and sensual contacts with things, whether they be submitted sequences (the "guys"), human body lice or SQL servers.

The database record, described in chapter 6, is one of the products of the integration work carried out by database staff. This chapter brought into relief the

indeterminacy and uncertainty, the different shades of "not knowing" or rather "almost knowing" that form an integral part of sense-making efforts. It explored the ways in which the database record and its assemblage facilitate the recording of a scale of absences and guesses without voiding its accuracy, viability and usefulness. This ranges from the (non)identification of organisms to the increase in environmental sampling, which sequences the unknown. Another integration is enacted on the database records themselves where very different kinds of information are brought together and through which manifold connections emerge. In this context, the actual nucleotide sequence documented by the record is just one part of a multi-layered arrangement of, at times, incongruous entities.

The term bioinformational artefact relays the informational and material character of these records while also drawing attention to the fact that their value as a resource relies on the combination of different elements and practices. This *cumulative relationality* of database records imparts novel affordances to the entities documented by the record. Once a sequence has been established as a record in EMBL-Bank or GenBank, sequence derived from, for example, yeast has an easier time in becoming useful for understanding human colon cancer. At the same time, the sequence forfeits its own species body as well as (traces of) its context of production because the record can only contain so much detail regarding how and where precisely the original sequence had been derived. Yet, escaping the vagaries and limits of its species body and *in vivo* processes, the sequence record accedes into the equally entangled ecology of the sequence universe.

Chapter 7 made present a specific instance of entanglement in the form of a controversy about accuracy in GenBank. In addition to unravelling "possible connections between problems" (Callon et al. 2009: 28), this chapter demonstrated how the controversy turned connection into problems. The open letter opened a space of dissent that brought into being a community of mycologists as dissenters. Leaving behind its "message", the letter became an "issue" in itself, enrolling in its wake a public – human and more-than-human. The signatories did not just speak for themselves as scientists – they made representations for their respective organisms, fungi, bacteria and other organisms. The controversy witnessed a proliferation of fungi: *in vivo, in vitro, in silico*, the fungus as

metaphor, as GenBank record, as absent (in popular knowledge) or as tangled mess. The issues of misannotation and accuracy revealed a concern for accounting for (some) relations between this "fungi multiple".

The open letter that had commenced this wrangle presented a curious political device for making representations for organisms. Calling for a return to traditional annotation practices established in natural history, it itself evoked a bygone era where such letters could cause political upset. In this case, the letter succeeded in forging proximities, for example, between researchers and their organism and between one organism (fungi) and another (*E. coli*). While bridging gaps of geographical, ontological and temporal nature, however, it also brought into relief antagonistic positions: the database-as-archive versus the mycologist community, for example, or, taxonomy versus wikis. Accuracy proved an ambiguous object as it struggled to cohere in a world without reference value but nevertheless managed to raise an affective register. The affected entity, fungal sequence, exhibited similarly order-defying traits yet elicited representations and solidarity not only from mycologists but the *E.coli* community. Finally, the controversy also betrayed multiple kinds of disarrangement: The messes mentioned by GB6 and the *Science* caption ("tangled mess") refer to a different order than the "chaos" evoked by GenBank director Lipman. Rather than the man-made chaos of wikification, these messes emerged as an integral characteristic of the organism.

Methods for meetings within and through the sequence universe

The sequence universe suggests that "[i]ntelligibility is not a human-based affair" (Barad 2007, pp.436, fn. 76). Instead, making sense amidst data entangles many different entities. This "forces a new relationship between the natural sciences and the social sciences" which demands of us to "develop scientific thinking at the intersection of different domains." (Smelik & Lykke 2008, p.xiv) Science studies have assembled a repertoire of techniques to overcome invisibility of infrastructure, pry open black boxes and disassemble modern dichotomies: following actors, rendering interpretative frames, looking for breakdowns, controversies and inversions, and creating "thick descriptions". More recently, scholars have drawn together methods more adept at traversing vague

wholes and remaining committed to the idea that reality is never "stable, determinate, and therefore knowable and predictable." (Law 2004, p.144) Indeterminate and unpredictable phenomena like arthrosclerosis (Mol 2002), sick-building syndrome (Murphy 2006), antiretroviral drugs (Rosengarten 2009), the horse-shoe crab (Gisler & Michael 2011), redstarts (Hinchliffe & Whatmore 2006) or future users (Wilkie & Michael 2009) require inventive methods that perform attachments as well as detachments. The biosciences offer particularly fertile grounds for such "tangled objects" (B. Latour 2004, p.22), engaging questions that are at once epistemological and ontological. They require of the researcher to take an *atheoretical* and *anti-reductionist* stance that allows ample room for relationalities to emerge and that uses terms "tolerant enough in their reference to bridge the divides between the various phenomena in which local communities of researchers may be interested."(Dupré 2004, p.334)

In chapter 3 I have described my methods as diffractive (Haraway 1992) and committed myself to inventive problem-making. Here, I wish to specify three attendant methods that have emerged in the course of writing this thesis: angelic figures, wonder, and non-commensurate reading. Chapter 4 took the convention of the ethnographic journey – field notes, field diaries and observations – and paired it with a discovery journey through the sequence universe recounted by a biologist. In projecting the figure of the "doubtful guest" across them, the journeys converged in mapping a speculative space of encounters. The doubtful guest has very much served as an angelic figure (Serres 1995a), productive in making connections and delivering messages. There are other kindred figures that make appearances throughout this thesis – the *Sun Tunnels*, Joseph Grinnell, the open letter and, most prominently, the sequence universe. Though it is very much the object of study, it is equally an explanatory and performative figure not unlike the doubtful guest. These angelic figures help scale and partially integrate worlds.

The appearance of angelic figures often occasions wonder. They defy expectations by delivering novel perspectives, connections and messages. Wonder is a useful disposition in encounters with objects. "Wonder", Fisher writes, "drives and sustains the defective rationality that gives us intelligibility under conditions where we will not even know that we have reached certain knowledge when and if we have." (1998, p.9) Though

it does not preclude criticality, wonder is foremost about encounters amidst messy and vague assemblages where things have not yet settled in-here or out-there. If interest is the condition for experimental engagement "in the *aesthetic*, *affective*, and *ethological* sense" (Stengers 2000, p.92), then wonder is the condition for interest. Wonder encompasses both attachment and detachment – engaging with an object may require dismissing some of its features, removing oneself from a particular locale or community or, indeed, detaching oneself from interests (impartiality). "Why are you looking at us?" many of my respondents asked me, incredulous as to how their work and workplace could in any way be interesting for an outsider, let alone relevant for sociological analysis. My answer to this question changed, recursively – an adaption that re-flected my respondents' sense of wonder through my own vagueness.

Approaching the database record as a bioinformational artefact in chapter 6 invites both exegesis and ethnographic explorations. If we consider the database record as bioinformational artefact, we endow it with corporeality, history and culture but we also make it resonant with concerns formulated elsewhere, in narratives, for example, of disciplines such as anthropology, archaeology or art history. Yet, as a "written" instance it occupies a distinct place that is amenable to certain questions (What language does it use?) but resistant to others (How does it smell?). In engaging with database records, the EMBL-Bank and GenBank flat files, I employed an approach of *non-commensurate reading* of the records. This explores records as texts and, in turn, these texts as phenomena. Narrative, as Porter's discovery journey demonstrates, is certainly one way to move from data to meaning such as protein function. The post-HGP challenge of making sense of all the data has therefore been identified with a "narrative turn" (Holmberg 2005). While this would suggest semantic analysis to take centre stage in sociological work about the practices and promises of "omics", it also invites more liberal, imaginative entanglements with different kinds of texts. Given the disproportionate and disproportional connections enacted through the sequence universe, it is only prudent to do the same in our study of these connections. Non-commensurate reading is performative, it makes present actors, objects, relations, histories. In this sense it is both grandiose and modest, much like the *Sun Tunnels*: Furnishing accounts by means of extravagant associations (to stellar

constellations for example), it does not seek the "bliss of organic symbiosis" (Haraway 1991, p.187). Instead, non-commensurate reading leaves residuals and loose ends, giving space to loss and recovery, fact and fiction, presence and absence, information gain and information loss.

Database imaginaries for biodiverse worlds

The entities and environment made present in this study of EMBL-Bank and GenBank encompass humans and more-than-humans, technologies and techniques, matters and matter. Contestation, arbitration, discussion, and uncertainty are key elements in the making of the databases. And rather than purging indeterminacy, much of their work actually tries to preserve the ambiguities that scientists encounter in experimental settings. Even a cursory search for "accuracy" and "GenBank" will reveal countless research papers and articles that list errors and imprecision while simultaneously offering methods and/or tools for redress. Similarly, the rise of community-led resources such as wikis that seek to build and enhance the primary data provided by the databases suggests that they are but one arbiter in a sequence that sees data continuously worked on and through. The experimental and inventive nature of collecting and "databasing" (Bowker 2006) is particularly prominent in the practices of environmental sampling and metagenomics, as I have illustrated in chapter 6. Here, radically novel matter is made present. And it is the techniques described in chapter 6 of recording species via sequencing conserved regions (in vertebrates it is the mitochondrial cytochrome c oxidase 1 or "CO1" gene) that forms the basis for the barcoding initiatives that are creating reference libraries of species identifiers, most famously the Barcode of Life Database (BoLD).[181]

Biodiversity, as a project, is very much predicated on making and recording presences – of species and habitats – and ordering them in lists which are preserved in databases, most prominently, in BoLD. It also requires the integration of different kinds of

[181] The INSDC is a partner in the Barcode of Life project and has agreed to apply the BOLD data standards. This comprises tagging any BoL submissions with the keyword "Barcode" and ensuring the completeness and accuracy of a number of data elements. See the "Data Standards for BARCODE Records in INSDC (BRIs)" of the Database Working Group part of the Consortium for the Barcode of Life at http://www.barcodeoflife.org/.

data: ecological, biosystematic, taxonomic, social and political (Bowker 2006, p.189).[182] This poses two interrelated questions: Whose presence is (not) being recorded and how will this recording be (made) relevant for biodiversity? In light of the recent formal constitution of the Inter-governmental Panel on Biodiversity and Ecosystem Services (IPBES), these are extremely pressing concerns.[183] There is much reason to fear that the IPBES might go down the same path as the IPCC, the Inter-governmental Panel on Climate Change, which remains by any accounts a spectacular failure. This failure can be attributed to its success in reducing climate change to one variable, global surface temperature, and one commodity, carbon. Consequently, it leaves no room for imaging climate change differently – imagining it, for example, through local crops or social justice. Aside from rendering absent key issues and constituencies this also feeds the affective gap between the scale of disasters and the range of feelings and cognitive competencies required to account (let alone address) such disasters. How can biodiversity escape the same fate, particularly when all indications point towards one coding region and absolute definitions of "preservation", "loss" and "ecosystem"?[184] Given the entanglements with data and databases effected by current biodiversity initiatives it is critical "to create flexible databases that are as rich ontologically as the social and natural worlds they map, and that might really help us gain long-term purchase on questions of planetary management." (Bowker 2006, 121)

What would such flexible databases entail and what kind of management might they permit? These are questions worth further study. They involve both practical and conceptual issues, some of which this thesis has touched upon. Concrete lessons to be drawn from the observations contained within the present study pertain to the importance of annotation and their dialogic and paratextual import, the sustainable long-term maintenance of bioinformational resources and the practice of biocuration for which there exists little training and even less recognition. Attendant concerns relate to forms of collaborations such as wikis or consortia, and the standards and protocols for such

[182] I wish to thank the participants of the "Biodiversity Knowledge Politics" workshop held at the Centre for Research in the Arts, Social Sciences and Humanities, University of Cambridge on 11 and 12 May, 2012, for their important insights on these topics.
[183] This was established by 90 governments on 21 April 2012 in Panama City.
[184] See the IPBES glossary at http://www.ipbes.net/about-ipbes/resources.html.

collective work around data production and sharing. Similarly, the scientific merit system based on the publication of research papers needs to be more responsive in relation to such contributions. There is also much scope for new scientific careers and business models as well as the re-appraisal of techniques and skills such as librarianship.

Regarding more conceptual issues, one key consideration concerns the expectations assembled around data – its lifecycle, its accuracy and general capacities. As demonstrated in the course of the present research, to engage with data is to "swim in an ocean of materiality" (Ingold 2011, p.24). While it diminishes certain contingencies, it proliferates others. This precludes any naïve or unequivocal relation between data and their referents such as coding regions or source organisms. Furthermore, if messy entanglements are constitutive features of organisms then surely, data and annotation need to reflect this. But how to accommodate both accuracy and mess, how to make room for indeterminacy and other absences while retaining credibility? What to do when the specimens are long gone, irretrievable lost amidst museum collections or broken hyperlinks? Sometimes, they are barely living up (or down) to the role of specimen. They are grown in biological resource centres or inhabit a realm so other that their existence is predicated on what is absent rather than tangible collectable evidence. What is the "empiricity" (Helmreich 2011) of the uncultured bacterium or any other similarly elusive entity recorded by a database record?

Flexible databases would suggest making present the ongoing, at times incongruous, socio-material practices that differentially enact these databases. It would therefore be conducive to craft database imaginaries that can mediate between these socio-material practices and curb any hope for definitive numbers or single solutions.[185] Such databases would afford encounters with naturecultures that are not unlike the encounters with doubtful guests or *Sun Tunnels* where the prepositional plenitude gives occasion to relationalities that connect stars, viruses, concrete, desert, wonder, horizon, gaps, scaffolds, and art historians.

[185] Hope, as Latour (B. Latour 2011) pointed out, does not help us relate to the cosmos, let alone make a difference that matters.

Bibliography

Aas, K.F., 2004. From Narrative to Database Technological Change and Penal Culture. *Punishment & Society*, 6(4), pp.379–393.

Agar, M., 1980. *The professional stranger: an informal introduction to ethnography*, New York: Academic Press.

Ahmed, S. & Stacey, J., 2001. *Thinking through the skin*, New York: Routledge.

Allen, G.E., 1975. *Life science in the twentieth century*, New York: Wiley.

Anderson, B.R.O., 1983. *Imagined communities: reflections on the origin and spread of nationalism*, London: Verso.

Aporta, C., 2004. Routes, trails and tracks: Trail breaking among the Inuit of Igloolik. *Études Inuit Studies*, 28(2), pp.9–38.

Appleby, J.O., Hunt, L. & Jacob, M.C., 1994. *Telling the truth about history*, New York: Norton.

Arakawa, K. et al., 2009. Genome Projector: zoomable genome map with multiple views. *BMC Bioinformatics*, 10, p.31.

Arondekar, A.R., 2009. *For the record: on sexuality and the colonial archive in India*, Durham: Duke University Press.

Ashburner, Michael et al., 2000. Gene Ontology: tool for the unification of biology. *Nature genetics*, 25(1), pp.25–29.

Atkinson, P. et al. eds., 2010. *Handbook of ethnography* Repr., Los Angeles [u.a.]: SAGE.

Auden, W.H., 1991. Tonight at Seven-Thirty. In E. Mendelson, ed. *W.H. Auden: Collected poems*. London; Boston: Faber and Faber.

Bairoch, A., 2010. Bioinformatics for Human Proteomics: Current State and Future Status. In *Nature Precedings*. Biocuration 2010. Tokyo, Japan. Available at: http://precedings.nature.com/documents/5050/version/1 [Accessed March 23, 2012].

Bairoch, A. et al., 2004. Swiss-Prot: Juggling Between evolution and stability. *Briefings in Bioinformatics*, 5(1), pp.39–55.

Bakhtin, M., 1981. *The dialogic imagination: four essays*, Austin: University of Texas Press.

Barad, K., 2007. *Meeting the universe halfway : quantum physics and the entanglement of matter and meaning*, Durham: Duke University Press.

Barnes, B. & Dupré, J., 2008. *Genomes and what to make of them*, Chicago: University of Chicago Press.

Barth, J., 2009. *Taste, ethics, and the market in Guatemalan coffee: an ethnographic study*. Ph.D. Oxford: Oxford University.

Bastow, R. & Leonelli, S., 2010. Sustainable digital infrastructure. *EMBO reports*, 11(10), pp.730–734.

Bateman, A., 2010. Curators of the world unite: The International Society of Biocuration. *Bioinformatics*, 26(8), p.991.

Baxevanis, A.D. & Ouellette, B.F.F., 2001. *Bioinformatics: a practical guide to the analysis of genes and proteins*, New York: Wiley-Interscience.

Beaulieu, A., 2004. From brainbank to database: the informational turn in the study of the brain. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 35(2), pp.367–390.

Beaulieu, A., 2010. Research Note: From co-location to co-presence: Shifts in the use of ethnography for the study of knowledge. *Social Studies of Science*, 40(3), pp.453–470.

Beisel, U., 2011. *Who bites back first? Malaria control in Ghana and the politics of co-existence*. Ph.D. Milton Keynes: The Open University.

Bennett, J., 2010. *Vibrant matter: a political ecology of things*, Durham: Duke University Press.

Benson, D.A. et al., 2010. GenBank. *Nucleic Acids Research*, 39(Database), pp.D32–D37.

Berger, J., 1972. *Ways of seeing*, London: Penguin Books.

Berlant, L.G., 2004. Affirmative Culture. *Critical Inquiry*, 30(2), pp.445–451.

Berlant, L.G., 2008. *The female complaint: the unfinished business of sentimentality in American culture*, Durham: Duke University Press.

Bidartondo, M.I., 2008. Preserving Accuracy in GenBank. *Science*, 319(5870), p.1616a–1616a.

Birke, L.I.A., Arluke, A. & Michael, M., 2007. *The sacrifice: how scientific experiments transform animals and people*, West Lafayette: Purdue University Press.

Birney, Ewan, 2008. The gene love-in. *Ensembl Blog*. Available at: http://www.ensembl.info/blog/2008/04/11/the-gene-love-in/.

Birnholtz, J.P. & Bietz, M.J., 2003. Data at work: supporting sharing in science and engineering. In *Proceedings of the 2003 international ACM SIGGROUP conference on Supporting group work*. GROUP '03. New York, NY: ACM, pp. 339–348. Available at: http://doi.acm.org/10.1145/958160.958215 [Accessed June 6, 2012].

Bolt Rasmussen, M., 2009. The Politics of interventionist art: The Situationist International, Artist Placement Group, and Art Workers' Coalition. *Rethinking Marxism*, 21(1), pp.34–49.

De Bont, R., 2009. Between the Laboratory and the Deep Blue Sea Space Issues in the Marine Stations of Naples and Wimereux. *Social Studies of Science*, 39(2), pp.199–227.

Borges, J., 1964. *Other inquisitions, 1937-1952.*, Austin: University of Texas Press.

Bork, P., 2000. Powers and Pitfalls in Sequence Analysis: The 70% Hurdle. *Genome Research*, 10(4), pp.398–400.

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

Bowker, G.C., 2006. *Memory practices in the sciences*, Cambridge, MA: MIT Press.

Bowker, G.C. & Star, S.L., 2000. *Sorting things out: classification and its consequences*, Cambridge, MA: MIT Press.

Brackney, D.E. et al., 2011. West Nile Virus Genetic Diversity is Maintained during Transmission by Culex pipiens quinquefasciatus Mosquitoes. , 6(9).

Brenner, S.E., 1999. Errors in genome annotation. *Trends in Genetics*, 15(4), pp.132–133.

Brouwer, J., Mulder, A. & Charlton, S. eds., 2003. *Information is alive: art and theory on archiving and retrieving data*, Rotterdam; New York: V2/NAi Publishers.

Browne, J.E., 2002. *Charles Darwin: The Power of Place*, London: Jonathan Cape.

Bruns, T., 2008. Classifying environmental sequences - why its necessary and what needs to be done. In FESIN. University of Tennessee.

Burri, R.V. & Dumit, Joseph eds., 2007. *Biomedicine as culture: instrumental practices, technoscientific knowledge, and new modes of life*, New York: Routledge.

Busby, H. & Martin, P., 2006. Biobanks, national identity and imagined communities: The case of UK biobank. *Science as Culture*, 15(3), pp.237–251.

Butler, J., 1993. *Bodies that matter: on the discursive limits of "sex,"* New York: Routledge.

Butler, J., 2005. *Giving an account of oneself*, New York: Fordham University Press.

Callon, M., 1998. *The laws of the markets*, Oxford; Malden, MA: Blackwell Publishers and Sociological Review.

Callon, M., Lascoumes, P. & Barthe, Y., 2009. *Acting in an uncertain world: an essay on technical democracy*, Cambridge, MA: MIT Press.

Calvert, J., 2007. Patenting Genomic Objects: Genes, Genomes, Function and Information. *Science as Culture*, 16(2), pp.207–223.

Calvino, I., 1968. *Cosmicomics*, New York: Harcourt, Brace & World.

Calvino, I., 1981. *If on a winter's night a traveler*, New York: Harcourt Brace Jovanovich.

Carson, C., 2007. Writing, Writing, Writing: The Natural History Field Journal as a Literary Text. *The Townsend Center Newsletter*, (February), pp.6–9.

Cartwright, L., 1998. A Cultural Anatomy of the Visible HUman Project. In P. A. Treichler, L. Cartwright, & C. Penley, eds. *The visible woman: imaging technologies, gender, and science*. New York: New York University Press, pp. 21–43.

Casiday, R.E., 2007. Children's health and the social theory of risk: Insights from the British measles, mumps and rubella (MMR) controversy. *Social Science & Medicine*, 65(5), pp.1059–1070.

Chow-White, P.A. & García-Sancho, M., 2012. Bidirectional shaping and spaces of convergence interactions between biology and computing from the first DNA sequencers to global genome databases. *Science, Technology & Human Values*, 37(1), pp.124–164.

Cochrane, Guy et al., 2010. The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Research*, 39(Database), pp.D15–D18.

Collins, F.S., 2008. Notes from the frontlines of the genomics revolution. In 25 Years of GenBank Symposium. NIH, Bethesda, MD. Available at: http://www.ncbi.nlm.nih.gov/About/genbank25/data/day2/collins/index.html.

Commins, J., Toft, C. & Fares, M.A., 2009. Computational Biology Methods and Their Application to the Comparative Genomics of Endocellular Symbiotic Bacteria of Insects. *Biological Procedures Online*, 11(1), pp.52–78.

Coole, D.H., 2010. The Inertia of Matter and the Generativity of Flesh. In D. H. Coole & S. Frost, eds. *New materialisms: ontology, agency, and politics*. Durham: Duke University Press, pp. 92–114.

Cosgrove, D.E., 2008. *Geography and vision: seeing, imagining and representing the world*, London: I.B. Tauris.

Critchley, S., Marchart, O. & Laclau, E. eds., 2006. Glimping the future. In *Laclau : a critical reader*. London: Routledge, pp. 279–327.

Czarniawska, B., 2004. On time, space, and action nets. *Organization*, 11(6), pp.773–791.

D'Onofrio, A. et al., 2010. Siderophores from Neighboring Organisms Promote the Growth of Uncultured Bacteria. *Chemistry & Biology*, 17(3), pp.254–264.

Daston, L. & Galison, P., 2007. *Objectivity*, Cambridge, MA: MIT Press.

Davies, G., 2003. A geography of monsters? *Geoforum*, 34(4), pp.409–412.

Davies, G., 2010. Where do experiments end? *Geoforum*, 41(5), pp.667–670.

Davies, G., 2011. Writing biology with mutant mice: The monstrous potential of post genomic life. *Geoforum*, (0). Available at: http://www.sciencedirect.com/science/article/pii/S0016718511000406 [Accessed March 23, 2012].

Decker, W.M., 1998. *Epistolary practices letter writing in America before telecommunications*, Chapel Hill: University of North Carolina Press.

Delborne, J.A., 2008. Transgenes and Transgressions: Scientific Dissent as Heterogeneous Practice. *Social Studies of Science*, 38(4), pp.509–541.

Denef, V.J., Mueller, R.S. & Banfield, J.F., 2010. AMD biofilms: using model communities to study microbial evolution and ecological complexity in nature. *The ISME Journal*, 4(5), pp.599–610.

Derrida, J., 1996. *Archive fever: a Freudian impression*, Chicago: University of Chicago Press.

Devos, D. & Valencia, A., 2001. Intrinsic errors in genome annotation. *Trends in Genetics*, 17(8), pp.429–431.

Dillon, M. & Lobo-Guerrero, L., 2009. The Biopolitical Imaginary of Species-Being. *Theory, Culture & Society*, 26(1), pp.1–23.

Doctorow, C., 2008. Big data: Welcome to the petacentre. *Nature*, 455(7209), pp.16–21.

Dombrowski, S.M. & Maglott, D.R., 2003. Using the Map Viewer to explore genomes. In J. McEntyre & J. Ostell, eds. *The NCBI Handbook*. Bethesda, MD: NCBI. Available at: http://www.ncbi.nlm.nih.gov/books/NBK21089/ [Accessed June 17, 2012].

Drexler, J.F. et al., 2012. Bats host major mammalian paramyxoviruses. *Nature Communications*, 3, p.796.

Dupré, J., 2004. Understanding contemporary genomics. *Perspectives on Science*, 12(3), pp.320–338.

Dwight, S.S. et al., 2004. Saccharomyces Genome Database: Underlying principles and organisation. *Briefings in Bioinformatics*, 5(1), pp.9–22.

Ebach, M.C. & Carvalho, M.R. de, 2010. Anti-intellectualism in the DNA barcoding enterprise. *Zoologia (Curitiba)*, 27(2), pp.165–178.

Ebert, L.B., 2008. To wikify GenBank? *IPBIZ*. Available at: http://ipbiz.blogspot.com/2008/06/to-wikify-genbank.html.

EBI-EMBL Group, 2012. *European Bioinformatics Institute: Annual Scientific Report 2011*, Hinxton: EBI-EMBL.

Eddy, S., 2009. Computational challenges in the future of large-scale sequencing. *Cryptogenomicon*. Available at: http://selab.janelia.org/people/eddys/blog/?p=86.

Editors, T., 2001. Is a Government Archive the Best Option? *Science*, 291(5512), pp.2318–2319.

Edwards, P.N., 2010. *A vast machine: computer models, climate data, and the politics of global warming*, Cambridge, MA: MIT Press.

Eglash, R., 2011. Multiple objectivity: an anti-relativist approach to situated knowledge. *Kybernetes*, 40(7/8), pp.995–1003.

Eisenstein, E.L., 1979. *The printing press as an agent of change: communications and cultural transformations in early modern Europe*, Cambridge; New York: Cambridge University Press.

Ellis, R. & Waterton, C., 2005. Caught between the cartographic and the ethnographic imagination: the whereabouts of amateurs, professionals, and nature in knowing biodiversity. *Environment and Planning D: Society and Space*, 23(5), pp.673–693.

EMBL-EBI, 2007. *The East Wing and a new dawn for the EMBL-EBI*, Wellcome Trust Genome Campus, Hinxton.

Emerson, D., Fleming, E.J. & McBeth, J.M., 2010. Iron-oxidizing bacteria: an environmental and genomic perspective. *Annual Review of Microbiology*, 64, pp.561–583.

Enwezor, O., 2008. *Archive fever: uses of the document in contemporary art*, New York: Steidl Publishers.

Essig, R.-B. & Nickisch, R.M.G. eds., 2007. *"Wer schweigt, wird schuldig!" : offene Briefe von Martin Luther bis Ulrike Meinhof*, Göttingen: Wallstein.

Fauquet, C.M. & Fargette, D., 2005. International Committee on Taxonomy of Viruses and the 3,142 unassigned species. *Virology*, 2(64).

Fausto-Sterling, A., 1985. *Myths of gender: biological theories about women and men*, New York: Basic Books.

Featherstone, M., 2006. Archive. *Theory, Culture & Society*, 23(2-3), pp.591–596.

Felt, U. ed., 2009. *Knowing and living in academic research: convergences and heterogeneity in research cultures in the European context*, Prague: Institute of Sociology of the Academy of Sciences of the Czech Republic.

Field, D. et al., 2009. 'Omics Data Sharing. *Science*, 326(5950), pp.234 –236.

Filip, G.M. & Ganio, L.M., 2004. Early Thinning in Mixed-Species Plantations of Douglas-Fir, Hemlock, and True Fir Affected by Armillaria Root Disease in Westcentral Oregon and Washington: 20 Year Results. *Western Journal of Applied Forestry*, 19(1), pp.25–33.

Finnegan, D., 2008. The Spatial Turn: Geographical Approaches in the History of Science. *Journal of the History of Biology*, 41(2), pp.369–388.

Fisher, P., 1998. *Wonder, the rainbow, and the aesthetics of rare experiences*, Cambridge, MA: Harvard University Press.

Fleischmann, R. et al., 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, 269(5223), pp.496 –512.

Flower, M.J. & Heath, D., 1993. Micro-anatomo politics: Mapping the human genome project. *Culture, Medicine and Psychiatry*, 17(1), pp.27–41.

Foster, H., 2004. An Archival Impulse. *October*, 110, pp.3–22.

Foucault, M., 1979. *Discipline and punish: the birth of the prison*, New York: Vintage Books.

Foucault, M., 1994. *The birth of the clinic: an archaeology of medical perception*, New York, NY: Vintage Books.

Foucault, M., 1970. *The order of things: an archaeology of the human sciences*, London: Tavistock.

Fox Keller, E., 1983. *A feeling for the organism: the life and work of Barbara McClintock*, San Francisco: W.H. Freeman.

Fox Keller, E., 1995. *Refiguring life: metaphors of twentieth-century biology*, New York: Columbia University Press.

Fox Keller, E., 2000. *The century of the gene*, Cambridge, MA: Harvard University Press.

Fox Keller, E., 2010. *The mirage of a space between nature and nurture*, Durham: Duke University Press.

Fox Keller, E. & Longino, H.E., 1996. *Feminism and science*, Oxford; New York: Oxford University Press.

Franklin, S., 2000. Global Nature and the Genetic Imaginary. In C. Lury, J. Stacey, & S. Franklin, eds. *Global nature, global culture*. London; Thousand Oaks; New Delhi: SAGE Publications, pp. 188–227.

Franklin, S., 2006. The cyborg embryo: Our path to transbiology. *Theory, Culture & Society*, 23(7-8), pp.167–187.

Franklin, S., Lury, C. & Stacey, J. eds., 2000. *Global nature, global culture*, London; Thousand Oaks; New Delhi: SAGE Publications.

Fraser, M., 2006. Event. *Theory, Culture & Society*, 23(2-3), pp.129–132.

Fraser, M., 2009. Experiencing Sociology. *European Journal of Social Theory*, 12(1), pp.63–81.

Fraser, M., 2010. Facts, Ethics and Event. In C. B. Jensen & K. Rödje, eds. *Deleuzian intersections: science, technology, anthropology*. New York: Berghahn Books, pp. 57–82.

Fredrickson, D.S., 1978. The National Institutes of Health yesterday, today, and tomorrow. *Public Health Reports*, 93(6), pp.642–647.

Fujimura, J.H., 1987. Constructing "do-able" problems in cancer research: articulating alignment. *Social Studies of Science*, 17(2), pp.257–293.

Fujimura, J.H., 2005. Postgenomic futures: translations across the machine-nature border in systems biology. *New Genetics and Society*, 24(2), pp.195–226.

Fujimura, J.H. & Fortun, M., 1996. Constructing Knowledge across Social Worlds: The Case of DNA Sequence Databases in Molecular Biology. In L. Nader, ed. *Naked science: anthropological inquiry into boundaries, power, and knowledge*. New York: Routledge, pp. 160–173.

Fujimura, J.H. & Rajagopalan, R., 2011. Different Differences: The Use of "genetic Ancestry" Versus Race in Biomedical Human Genetic Research. *Social Studies of Science*, 41(1), pp.5–30.

Fuller, M., 2009. Active data and its afterlives. *Take Away Media Festival*. Available at: http://www.spc.org/fuller/texts/active-data-and-its-afterlives/.

Fuller, S., 2009. Life beyond Darwin: Unbinding biology's time and space. *Progress in Human Geography*, 33(2), pp.147–153.

Gabrys, J., 2011a. *Digital rubbish: a natural history of electronics*, Ann Arbor: University of Michigan Press.

Gabrys, J., 2011b. Material politics: plastic, carbon and the work of the biodegradable. In Accumulation: The Material Ecologies and Economies of Plastic. Goldsmiths, London. Available at: http://jupiter.gold.ac.uk/sociology/calendar/?id=4452 [Accessed July 4, 2012].

Galas, D.J., 2001. Making Sense of the Sequence. *Science*, 291(5507), pp.1257 –1260.

Galison, P. & Jones, C.A., 1999. Factory, laboratory, studio: Dispersin sites of production. In P. Galison & E. A. Thompson, eds. *The architecture of science*. Cambridge, MA: MIT Press, pp. 497–539.

Galison, P. & Thompson, E.A. eds., 1999. *The architecture of science*, Cambridge, MA: MIT Press.

Gane, N. & Latour, B., 2004. Bruno Latour: The Social as Association. In *The future of social theory*. London; New York: Continuum.

García-Sancho, M., 2011. From metaphor to practices: the introduction of "information engineers" into the first DNA sequence database. *History and Philosophy of the Life Sciences*, 33(1), pp.71–104.

Garfinkel, S., 2000. *Database nation: the death of privacy in the 21st century*, Beijing; Cambridge: O'Reilly.

Garrity, G. et al., 2010. NamesforLife Semantic Resolution Services for the Life Sciences. *Nature Precedings*. Available at: http://precedings.nature.com/doifinder/10.1038/npre.2010.5137.1 [Accessed July 21, 2012].

Gaudillière, J.-P. & Rheinberger, H.-J., 2004. *From molecular genetics to genomics: the mapping cultures of twentieth-century genetics*, London; New York: Routledge.

Geertz, C., 1973. *The interpretation of cultures: selected essays*, New York: Basic Books.

Gere, C. & Parry, B., 2006. The Flesh Made Word: Banking the Body in the Age of Information. *BioSocieties*, 1(01), pp.41–54.

Gibbons, S.M., 2007. Are UK genetic databases governed adequately? A comparative legal analysis. *Legal Studies*, 27(2), pp.312–342.

Gieryn, T., 2006. City as Truth-Spot: Laboratories and Field-Sites in Urban Studies. *Social Studies of Science*, 36, pp.5–38.

Giles, J., 2007. Key biology databases go wiki. *Nature*, 445(7129), pp.691–691.

Gisler, P. & Michael, M., 2011. Companions at a distance: Technoscience, blood, and the horseshoe crab. *Society and Animals*, 19(2), pp.115–136.

Giustini, D., 2006. How Web 2.0 is changing medicine. *BMJ*, 333(7582), pp.1283–1284.

Gleick, J., 1994. *Genius: Richard Feynman and Modern Physics*, New York: Vintage Books.

Le Goff, J., 1992. *History and memory*, New York: Columbia University Press.

Golinski, J., 1998. *Making natural knowledge: constructivism and the history of science*, Cambridge; New York: Cambridge University Press.

Gooday, G., 2008. Placing or replacing the laboratory in the history of science? *Isis*, 99(4), pp.783–795.

Goodeve, T.N. & Haraway, D.J., 1999. *How like a leaf: an interview with Donna Haraway*, New York; London: Routledge.

Goodwin, C., 1994. Professional Vision. *American Anthropologist*, 96(3), pp.606–633.

Gorey, E., 1957. *The Doubtful Guest*, New York: Harcourt Brace Jovanovich.

Gottweis, H., 1998. *Governing molecules: the discursive politics of genetic engineering in Europe and the United States*, Cambridge Mass.: MIT Press.

Gottweis, H. & Petersen, A.R. eds., 2008. *Biobanks governance in comparative perspective*, Abingdon; New York: Routledge.

Grabowski, M. et al., 2007. Structural genomics: keeping up with expanding knowledge of the protein universe. *Current Opinion in Structural Biology*, 17(3), pp.347–353.

Gray, N.F., 1997. Environmental impact and remediation of acid mine drainage: a management problem. *Environmental Geology*, 30(1-2), pp.62–71.

Greenbaum, D. et al., 2011. Genomics and Privacy: Implications of the New Reality of Closed Data for the Field. *PLoS Comput Biol*, 7(12), p.e1002278.

Greenhough, B., 2006. Tales of an island-laboratory: defining the field in geography and science studies. *Transactions of the Institute of British Geographers*, 31(2), pp.224–237.

Greenpeace, 2012. *How Clean is Your Cloud?*, Available at: http://www.greenpeace.org/international/en/publications/Campaign-reports/Climate-Reports/How-Clean-is-Your-Cloud/ [Accessed July 4, 2012].

Greenpeace, 2011. *How dirty is your data?*, Available at: http://www.greenpeace.org/international/en/publications/reports/How-dirty-is-your-data/ [Accessed July 4, 2012].

Grier, D.A., 2005. *When computers were human*, Princeton: Princeton University Press.

Griesemer, J.R., 1990. Modeling in the museum: On the role of remnant models in the work of Joseph Grinnell. *Biology & Philosophy*, 5(1), pp.3–36.

Griesemer, J.R. & Gerson, E.M., 1993. Collaboration in the Museum of Vertebrate Zoology. *Journal of the History of Biology*, 26(2), pp.185–203.

Grinnell, J., 1968. *Joseph Grinnell's philosophy of nature: selected writings of a western naturalist*, Berkeley, CA: University of California Press.

Grinnell, J. & Storer, T.I., 1924. *Animal Life in the Yosemite: An account of the mammals, birds, reptiles, and amphibians in a cross-section of the Sierra Nevada*, Berkeley, CA: University of California Press.

Gugerli, D., 2009. *Suchmaschinen: die Welt als Datenbank*, Frankfurt am Main: Suhrkamp.

Guyer, J.I., 2004. *Marginal gains: monetary transactions in Atlantic Africa*, Chicago: University of Chicago Press.

Hacking, I., 1992. The self-vindication of the laboratory sciences. In A. Pickering, ed. *Science as practice and culture*. Chicago: University of Chicago Press, pp. 29–64.

Hagen, J.B., 2011. The origin and early reception of sequence databases. In M. Hamacher, M. Eisenacher, & C. Stephan, eds. *Data mining in proteomics: from standards to applications*. New York: Humana Press, pp. 61–77.

Hagen, J.B., 2000. The origins of bioinformatics. *Nature Reviews Genetics*, 1(3), pp.231–236.

Hamm, G.H. & Cameron, G.N., 1986. The EMBL Data Library. *Nucleic Acids Research*, 14(1), pp.5–9.

Haraway, D.J., 1997. *Modest–Witness@Second–Millennium.FemaleMan–Meets–OncoMouse: feminism and technoscience*, New York; London: Routledge.

Haraway, D.J., 1991. *Simians, cyborgs, and women: the reinvention of nature*, New York: Routledge.

Haraway, D.J., 2003. *The companion species manifesto: dogs, people, and significant otherness*, Chicago: Prickly Paradigm Press.

Haraway, D.J., 1992. The Promises of Monsters: A Regenerative Politics for Inappropriate/d Others. In L. Grossberg, C. Nelson, & P. A. Treichler, eds. *Cultural studies*. New York: Routledge, pp. 295–337.

Harding, A. ed., 2002. *Potential: ongoing archive*, Amsterdam: Artimo.

Harding, S.G., 1986. *The science question in feminism*, Ithaca: Cornell University Press.

Harvey, M. & Mcmeekin, A., 2002. The formation of bioinformatic knowledge markets: An "economies of knowledge" approach. *Revue d'économie industrielle*, 101(1), pp.47–64.

Häyry, M. ed., 2007. *The ethics and governance of human genetic databases : European perspectives*, Cambridge; New York: Cambridge University Press.

Helmreich, S., 2011. Nature/Culture/Seawater. *American Anthropologist*, 113(1), pp.132–144.

Herman, S.G. & Grinnell, J., 1986. *The naturalist's field journal : a manual of instruction based on a system established by Joseph Grinnell*, Vermillion: Buteo Books.

Hess, D., 2010. Ethnography and the Development of Science and Technology Studies. In P. Atkinson et al., eds. *Handbook of ethnography*. Los Angeles; London: SAGE, pp. 234–245.

Heyl, B.S., 2010. Ethnographic interviewing. In P. Atkinson et al., eds. *Handbook of ethnography*. London; Thousand Oaks; New Delhi: SAGE Publications, pp. 369–383.

Hilgartner, S., 1995. Biomolecular databases: new communication regimes for biology? *Science Communication*, 17(2), pp.240–263.

Hilgartner, S., 2004. Making Maps and Making Social Order: Governing American Genome Centers, 1988-1993. In J.-P. Gaudillière & H.-J. Rheinberger, eds. *From molecular genetics to genomics: the mapping cultures of twentieth-century genetics*. London; New York: Routledge, pp. 113–128.

Hilgartner, S. & Brandt-Rauf, S.I., 1994. Data access, ownership, and control toward empirical studies of access practices. *Science Communication*, 15(4), pp.355–372.

Hinchliffe, S., 2007. *Geographies of nature: societies, environments, ecologies*, London; Thousand Oaks; New Delhi: SAGE Publications.

Hinchliffe, S., 2004. Viruses. In S. Harrison, S. Pile, & N. J. Thrift, eds. *Patterned ground: entanglements of nature and culture*. London: Reaktion Books, pp. 228–230.

Hinchliffe, S. & Whatmore, S., 2006. Living cities: Towards a politics of conviviality. *Science as Culture*, 15(2), pp.123–138.

Hine, C., 2006. Databases as scientific instruments and their role in the ordering of scientific work. *Social Studies of Science*, 36(2), pp.269–298.

Hine, C., 2007. Multi-sited Ethnography as a Middle Range Methodology for Contemporary STS. *Science, Technology & Human Values*, 32, pp.652–671.

Hine, C., 2000. *Virtual ethnography*, London; Thousand Oaks; New Delhi: Sage Publications.

Hird, M.J., 2010. Meeting with the microcosmos. *Environment and Planning D: Society and Space*, 28(1), pp.36–39.

Hoeyer, K., 2003. "Science is really needed—that"s all I know': informed consent and the non-verbal practices of collecting blood for genetic research in northern Sweden. *New Genetics and Society*, 22(3), pp.229–244.

Hoeyer, K.L. & Tutton, R., 2005. "Ethics was here": Studying the language-games of ethics in the case of UK Biobank. *Critical Public Health*, 15(4), pp.385–397.

Holmberg, T., 2005. Questioning "the number of the beast": Constructions of humanness in a Human Genome Project (HGP) narrative. *Science as Culture*, 14(1), pp.23–37.

Hu, J.C. et al., 2008. The Emerging World of Wikis. *Science*, 320(5881), p.1289b–1290b.

Hugenholtz, P. & Tyson, G.W., 2008. Microbiology: metagenomics. *Nature*, 455(7212), pp.481–483.

Huttenhower, C. & Hofmann, O., 2010. A Quick Guide to Large-Scale Genomic Data Mining. *PLoS Comput Biol*, 6(5), p.e1000779.

Idnurm, A. et al., 2006. The Phycomyces madA gene encodes a blue-light photoreceptor for phototropism and other light responses. *Proceedings of the National Academy of Sciences of the United States of America*, 103(12), pp.4546 –4551.

Ingold, T., 2011. *Being alive: essays on movement, knowledge and description*, London; New York: Routledge.

Ingold, T., 2000. *The perception of the environment: essays on livelihood, dwelling & skill*, London; New York: Routledge.

Ingold, T., 1996. The temporality of the Landscape. In R. Preucel & I. Hodder, eds. *Contemporary archaeology in theory*. Oxford: Blackwell, pp. 59–76.

Ionesco, E., 1965. Amédée, or how to get rid of it. In M. Esslin, ed. *Absurd drama*. Harmondsworth: Penguin Books.

Jasanoff, S., 2007. *Designs on nature: science and democracy in Europe and the United States*, Princeton: Princeton University Press.

Jerzak, G. et al., 2005. Genetic variation in West Nile virus from naturally infected mosquitoes and birds suggests quasispecies structure and strong purifying selection. *The Journal of general virology*, 86(Pt 8), pp.2175–2183.

Jones, C.E., Brown, A.L. & Baumann, U., 2007. Estimating the annotation error rate of curated GO database sequence annotations. *BMC Bioinformatics*, 8(1), p.170.

Karsch-Mizrachi, I., 2007. GenBank: The Nucleotide Sequence Databe. In J. McEntyre & J. Ostell, eds. *The NCBI Handbook*. Bethesda, MD. Available at: http://www.ncbi.nlm.nih.gov/books/NBK21105/.

Karsch-Mizrachi, I. et al., 2011. The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Research*, 40(D1), pp.D33–D37.

Kawashima, Y. et al., 2005. Epidemiological study of mumps deafness in Japan. *Auris Nasus Larynx*, 32(2), pp.125–128.

Kay, L.E., 2000. *Who wrote the book of life? A history of the genetic code*, Stanford: Stanford University Press.

Kember, S., 1998. *Virtual anxiety: photography, new technologies, and subjectivity*, Manchester; New York: Manchester University Press.

Kevles, D.J. & Geison, G.J., 1995. The experimental life sciences in the twentieth century. *Osiris*, 10, pp.97–121.

Knorr Cetina, K. & Bruegger, U., 2002. Global microstructures: The virtual societies of financial markets. *American Journal of Sociology*, 107(4), pp.905–950.

Knorr-Cetina, K., 1999. *Epistemic cultures: how the sciences make knowledge*, Cambridge, MA: Harvard University Press.

Knorr-Cetina, K., 1992. The Couch, the Cathedral and the Lab: On the Relationship between Experiment and Laboratory Science. In A. Pickering, ed. *Science as practice and culture*. Chicago: University of Chicago Press.

Knorr-Cetina, K., 1981. *The manufacture of knowledge: an essay on the constructivist and contextual nature of science*, Oxford; New York: Pergamon Press.

Kohler, R., 2006. *All creatures: naturalists, collectors, and biodiversity, 1850-1950*, Princeton: Princeton University Press.

Kohler, R., 2002a. *Landscapes & labscapes: exploring the lab-field border in biology*, Chicago, IL: University of Chicago Press.

Kohler, R., 2002b. Place and practice in field biology. *History of Science*, 40(2), pp.189–210.

Korts, K., 2004. Introducing gene technology to the society: Social implications of the Estonian Genome Project. *Trames*, 8(1-2), pp.241–253.

Krige, J., 2002. The birth of EMBO and the difficult road to EMBL. *Studies in History and Philosophy of Science Part C: Biological and Biomedical Sciences*, 33(3), pp.547–564.

Kvale, S., 2007. *Doing interviews*, London; Thousand Oaks; New Delhi: SAGE Publications.

Larhammar, D. & Milner, R.J., 1989. Phylogenetic relationship of birds with crocodiles and mammals, as deduced from protein sequences. *Molecular Biology and Evolution*, 6(6), pp.693 –696.

Lathe, W.C.I., 2008. Wikification of GenBank. *The OpenHelix Blog*. Available at: http://blog.openhelix.eu/?p=275 [Accessed March 15, 2012].

Latour, B., 1990. Drawing Things Out. In M. Lynch & S. Woolgar, eds. *Representation in scientific practice*. Cambridge, MA: MIT Press.

Latour, B., 1983. Give me a laboratory and I will move the world. In K. Knorr-Cetina & M. J. Mulkay, eds. *Science observed: perspectives on the social study of science*. London; Beverly Hills; New Delhi: SAGE Publications, pp. 141–170.

Latour, B., 1999. *Pandora's hope: essays on the reality of science studies*, Cambridge, MA: Harvard University Press.

Latour, B., 2004. *Politics of nature: how to bring the sciences into democracy*, Cambridge, MA: Harvard University Press.

Latour, B., 2005. *Reassembling the social: an introduction to actor-network-theory*, Oxford; New York: Oxford University Press.

Latour, B., 1987. *Science in action: how to follow scientists and engineers through society*, Cambridge, MA: Harvard University Press.

Latour, B., 1988. *The pasteurization of France*, Cambridge, MA: Harvard University Press.

Latour, B., 2011. Waiting for Gaia: Composing the common world through arts and politics. French Institute, London, 21 November 2011.

Latour, B. & Woolgar, S., 1986. *Laboratory life: the construction of scientific facts*, Princeton: Princeton University Press.

Latour, Bruno, 2004. Why has critique run out of steam? From matters of fact to matters of concern. *Critical Inquiry*, 30(2), pp.225–248.

Law, J., 2004. *After method: mess in social science research*, London: Routledge.

Law, J., 1986. On power and its tactics: a view from the sociology of science. *The Sociological Review*, 34(1), pp.1–38.

Law, J., 2010. The materials of STS. In D. Hicks & M. Beaudry, eds. *The Oxford Handbook of Material Cultural Studies*. pp. 171–186.

Law, J. & Mol, A., 1995. Notes on materiality and sociality. *The Sociological Review*, 43(2), pp.274–294.

Law, J. & Mol, A., 2001. Situating technoscience: an inquiry into spatialities. *Environment and Planning D: Society and Space*, 19(5), pp.609–621.

Law, J. & Singleton, V., 2005. Object Lessons. *Organization*, 12(3), pp.331–355.

Leinonen, R. et al., 2010. The European Nucleotide Archive. *Nucleic Acids Research*, 39(Database), pp.D28–D31.

Lenoir, T., 1999. Shaping biomedicine as an information science. In M. E. Bowden, T. B. Hahn, & R. V. Williams, eds. *Proceedings of the 1998 Conference on the History and Heritage of Science Information Systems*. Conference on the History and Heritage of Science Information Systems. Medford, NJ: Published for the American Society for Information Science and the Chemical Heritage Foundation by Information Today, pp. 27–45.

Leonelli, S., 2007a. Arabidopsis, the botanical Drosophila: from mouse cress to model organism. *Endeavour*, 31(1), pp.34–38.

Leonelli, S., 2008. Performing abstraction: two ways of modelling Arabidopsis thaliana. *Biology and Philosophy*, 23(4), pp.509–528.

Leonelli, S., 2007b. *Weed for Thought: Using Arabidopsis thaliana to Understand Plant Biology*. Ph.D. Amsterdam: Vrije Universiteit Amsterdam.

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

Levitt, M., 2009. Nature of the protein universe. *Proceedings of the National Academy of Sciences*, 106(27), pp.11079–11084.

Lewis, K., 2010. The Uncultured Bacteria. *Small Things Considered: The Microbe Blog*. Available at: http://schaechter.asmblog.org/schaechter/2010/07/the-uncultured-bacteria.html [Accessed March 31, 2011].

Li, Z. et al., 2006. Beilong virus, a novel paramyxovirus with the largest genome of non-segmented negative-stranded RNA viruses. *Virology*, 346(1), pp.219–228.

Liss, A., 2009. *Feminist art and the maternal*, Minneapolis: University of Minnesota Press.

Livingstone, D.N., 2003. *Putting science in its place: geographies of scientific knowledge*, Chicago: University of Chicago Press.

Lord, P. & MacDonald, A., 2003. *e-Science Curation Report*, Twickenham: The JISC Committee for the Support of Research.

Luyt, B., 2008. Centres of calculation and unruly colonists: the colonial library in Singapore and its users, 1874. *Journal of Documentation*, 64(3), pp.386 – 396.

Lynch, Michael, 1999. Archives in formation: Privileged spaces, popular Archives and paper trails. *History of the Human Sciences*, 12(2), pp.65–87.

Lynch, Michael, 1985. *Art and artifact in laboratory science: a study of shop work and shop talk in a research laboratory*, London: Routledge & Kegan Paul.

Lynch, Michael, 2008. Ontography: Investigating the Production of Things, Deflating Ontology. In Oxford Ontologies Workshop. Said Business School, Oxford.

Lynch, M.E., 1988. Sacrifice and the transformation of the animal body into a scientific object: Laboratory culture and ritual practice in the neurosciences. *Social Studies of Science*, 18(2), pp.265–289.

Lyons, M., 1999. Love Letters and Writing Practices: On Écritures Intimes in the Nineteenth Century. *Journal of Family History*, 24(2), pp.232–239.

Mackenzie, A., 2003a. Bringing sequences to life: how bioinformatics corporealizes sequence data. *New Genetics and Society*, 22(3), pp.315–332.

Mackenzie, A., 2006. *Cutting code: software and sociality*, New York: Peter Lang.

Mackenzie, A., 2005a. Problematising the Technological: The Object as Event? *Social Epistemology*, 19(4), pp.381–399.

Mackenzie, A., 2003b. These things called systems: Collective imaginings and infrastructural software. *Social Studies of Science*, 33, pp.365–387.

Mackenzie, A., 2005b. Untangling the unwired Wi-Fi and the cultural inversion of infrastructure. *Space and Culture*, 8(3), pp.269–285.

Malek, J.A. & Haft, D.H., 2001. Conserved protein domains are maintained in an average ratio to proteome size. *Genome Biology*, 2(5). Available at: http://genomebiology.com/2001/2/5/preprint/0004 [Accessed March 13, 2012].

Manning, E., 2010. Always More than One: The Collectivity of a Life. *Body & Society*, 16(1), pp.117–127.

Manovich, L., 2001. *The language of new media*, Cambridge, MA: MIT Press.

Marcus, G., 1998. *Ethnography through thick and thin*, Princeton: Princeton University Press.

Marks, L.U., 2002. *Touch: sensuous theory and multisensory media*, Minneapolis, MN: University of Minnesota Press.

Marres, N., 2005. *No Issue, No Public: Democratic Deficits after the Displacement of Politics*. Amsterdam: University of Amsterdam.

Marres, N., 2009. Testing powers of engagement: Sustainable living experiments, the object turn, and the undoability of public involvement. In *APSA 2009 Toronto Meeting Paper*. APSA. Toronto. Available at: http://papers.ssrn.com/Sol3/papers.cfm?abstract_id=1449735 [Accessed March 20, 2012].

Marres, N., 2007. The issues deserve more credit: Pragmatist contributions to the study of public involvement in controversy. *Social Studies of Science*, 37(5), pp.759–780.

Massumi, B., 1997. The Autonomy of Affect. In P. Patton, ed. *Deleuze: a critical reader*. Oxford: Blackwell, pp. 217–239.

Maurer, B., 2005. *Mutual life, limited: Islamic banking, alternative currencies, lateral reason*, Princeton: Princeton University Press.

Mayberry, M., Subramaniam, B. & Weasel, L.H., 2001. *Feminist science studies: a new generation*, New York: Routledge.

Meyers, B.C., Scalabrin, S. & Morgante, M., 2004. Mapping and sequencing complex genomes: let's get physical! *Nature Reviews Genetics*, 5(8), pp.578–588.

Michael, M., 2009. Publics performing publics: of PiGs, PiPs and politics. *Public Understanding of Science*, 18(5), pp.617–631.

Michael, M., 2004. Roadkill: Between humans, nonhuman animals, and technologies. *Society and Animals*, 12(4), pp.277–298.

Michael, M., 2010. Some disjointed thoughts on transbiology: From spaces to prepositions. In The Spaces of Transbiology. Wellcome Trust, London.

Michael, M., 2006. *Technoscience and everyday life: the complex simplicities of the mundane*, Maidenhead; New York: Open University Press.

Michael, M. & Brown, N., 2004. The Meat of the Matter: Grasping and Judging Xenotransplantation. *Public Understanding of Science*, 13(4), pp.379–397.

Michael, M., Wainwright, S.P. & Williams, C., 2007. Temporality and Prudence: On Stem Cells as "Phronesic Things." *Configurations*, 13(3), pp.373–394.

Miller, P. & O'Leary, T., 1994. The Factory as Laboratory. *Science in Context*, 7(03), pp.469–496.

Miller, P. & Rose, N., 1990. Governing economic life. *Economy and Society*, 19(1), pp.1–31.

Mohrhardt, F.E., 1962. A Building for the National Library of Medicine. *Libri*, 12(3), pp.235–239.

Mol, A., 2002. *The body multiple: ontology in medical practice*, Durham: Duke University Press.

Mol, A., 2008. *The logic of care: health and the problem of patient choice*, London; New York: Routledge.

Mollenkopf, J.H. & Castells, M. eds., 1991. *Dual city: restructuring New York*, New York: Russell Sage Foundation.

Molyneux, A.J. & Coghill, S.B., 1994. Cell Lysis Due to Ultrasound Gel In Fine Needle Aspirates; an Important New Artefact In Cytology. *Cytopathology*, 5(1), pp.41–45.

Moretti, S., 2011. In silico experiments in scientific papers on molecular biology. *Science Studies*, 24(2). Available at: http://www.sciencestudies.fi/v24n2MorettiPDF.

Morgan, M. & Morrison, M. eds., 1999. *Models as mediators: perspectives on natural and social sciences*, Cambridge; New York: Cambridge University Press.

Murphy, M., 2009. Scale, topography, origami. Workshop on Scaleography. University of Oxford, 8 July 2009.

Murphy, M., 2006. *Sick building syndrome and the problem of uncertainty: environmental politics, technoscience, and women workers*, Durham: Duke University Press.

Myers, N., 2008. Molecular Embodiments and the Body-work of Modeling in Protein Crystallography. *Social Studies of Science*, 38(2), pp.163–199.

Myers, N. & Dumit, Joe, 2011. Haptics. In F. E. Mascia-Lees, ed. *A companion to the anthropology of the body and embodiment*. Hoboken: Wiley-Blackwell, pp. 239–261.

Nair, S.P., 2005. Native Collecting and Natural Knowledge (1798-1832): Raja Serfoji II of Tanjore as a "Centre of Calculation." *Journal of the Royal Asiatic Society*, 15(03), pp.279–302.

Nelkin, D. & Anker, S., 2003. *The molecular gaze: art in the genetic age*, New York; Oxford: Cold Spring Harbor Laboratory Press.

Nicholls, H., 2007. Sorcerer II: The Search for Microbial Diversity Roils the Waters. *PLoS Biol*, 5(3), p.e74.

O'Malley, M.A. & Dupré, J., 2005. Fundamental issues in systems biology. *BioEssays*, 27(12), pp.1270–1276.

Olszewski, R.T., 2003. Bayesian Classification of Triage Diagnoses for the Early Detection of Epidemics. In *Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference*. International Florida Artificial Intelligence Research Society Conference. St. Augustine, FL: AAAI Press, pp. 412–416.

Ophir, A. & Shapin, S., 1991. The Place of Knowledge A Methodological Survey. *Science in Context*, 4(01), pp.3–22.

Osborne, T., 1999. The ordinariness of the archive. *History of the Human Sciences*, 12(2), pp.51–64.

Ouzounis, C.A., 2012. Rise and demise of bioinformatics? promise and progress. *PLoS Comput Biol*, 8(4), p.e1002487.

Palsson, B., 2000. The challenges of in silico biology. *Nature Biotechnology*, 18(11), pp.1147–1150.

Pálsson, G., 2008. The rise and fall of a biobank: The case of Iceland. In H. Gottweis & A. R. Petersen, eds. *Biobanks governance in comparative perspective*. Abingdon; New York: Routledge, pp. 41–55.

Pálsson, G. & HarÐardóttir, K.E., 2002. For whom the cell tolls: Debates about biomedicine. *Current Anthropology*, 43(2), pp.271–301.

Parisi, L., 2010. Event and evolution. *The Southern Journal of Philosophy*, 48, pp.147–164.

Parry, B., 2004. *Trading the genome: investigating the commodification of bio-information*, New York: Columbia University Press.

Parry, B. & Gere, C., 2006. Contested bodies: property models and the commodification of human biological artefacts. *Science as Culture*, 15(2), pp.139–158.

Pasigraphy, 2008. Science Editors=Hippies, NCBI-ists=Commies, Fungi Bioinformaticians=Babies. *Pasygraphy*. Available at: http://pasigraphy.wordpress.com/2008/03/27/science-editorshippies-ncbi-istscommies-computational-biologistsbabies/ [Accessed March 15, 2012].

Paulson, W., 2001. For a comsopolitical philology: Lessons from science studies. *SubStance*, 30(3), pp.101–119.

Pearson, W.R. & Lipman, D.J., 1988. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, 85(8), pp.2444–2448.

Pennisi, E., 2008. Proposal to "Wikify" GenBank Meets Stiff Resistance. *Science*, 319(5870), pp.1598–1599.

Petersen, A., 2005. Biobanks: Challenges for "ethics." *Critical Public Health*, 15(4), pp.303–310.

Pevzner, P.A., Tang, H. & Waterman, M.S., 2001. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences*, 98(17), pp.9748 –9753.

Pico, A.R. et al., 2008. WikiPathways: Pathway Editing for the People. *PLoS Biol*, 6(7), p.e184.

Pignatelli, M. et al., 2008. Metagenomics reveals our incomplete knowledge of global diversity. *Bioinformatics*, 24(18), pp.2124–2125.

Pinch, T.J. & Bijker, W.E., 1984. The Social Construction of Facts and Artefacts: Or How the Sociology of Science and the Sociology of Technology Might Benefit Each Other. *Social Studies of Science*, 14(3), pp.399–441.

Porter, S., 2008a. Biologists vs. the Age of Information. *Discovering Biology in a Digital World*. Available at: http://scienceblogs.com/digitalbio/2008/06/biologists_vs_the_age_of_infor.php [Accessed March 15, 2012].

Porter, S., 2008b. Dinosaur DNA discovered in GenBank. *Discovering Biology in a Digital World*. Available at: http://scienceblogs.com/digitalbio/2008/04/dinosaur_dna_discovered_in_gen.php.

Poster, M., 1990. *The mode of information: poststructuralism and social context*, Chicago, IL: University of Chicago Press.

Poster, M., 1995. *The second media age*, Cambridge: Polity Press.

Pottage, A., 2006. Too Much Ownership: Bio-prospecting in the Age of Synthetic Biology. *BioSocieties*, 1(2), pp.137–158.

Povolotskaya, I.S. & Kondrashov, F.A., 2010. Sequence space and the ongoing expansion of the protein universe. *Nature*, 465(7300), pp.922–926.

Powell, R.C., 2007. Geographies of Science: Histories, Localities, Practices, Futures. *Progress in Human Geography*, 31(3), pp.309–329.

Pruitt, K.D. et al., 2002. The Reference Sequence (RefSeq) Database. In J. McEntyre & J. Ostell, eds. *The NCBI Handbook*. Bethesda, MD: National center for Biotechnololgy Information. Available at: http://www.ncbi.nlm.nih.gov/books/NBK21091/.

Pynchon, T., 1997. *Mason & Dixon*, New York: Henry Holt.

Rabinow, P., 1996. *Making PCR: a story of biotechnology*, Chicago: University of Chicago Press.

Reed, J.L. et al., 2006. Towards multidimensional genome annotation. *Nature Reviews Genetics*, 7(2), pp.130–141.

Renfro, D.P. et al., 2011. GONUTS: the Gene Ontology Normal Usage Tracking System. *Nucleic Acids Research*, 40(D1), pp.D1262–D1269.

Rheinberger, H.-J., 1998. Experimental systems–graphematic spaces. In T. Lenoir & H. U. Gumbrecht, eds. *Inscribing science: Scientific texts and the materiality of communication*. Stanford: Stanford University Press, pp. 285–303.

Rheinberger, H.-J., 1997. *Toward a history of epistemic things: synthesizing proteins in the test tube*, Stanford: Stanford University Press.

Rheinberger, H.-J. & Gaudillière, J.-P. eds., 2004. *Classical genetic research and its legacy: the mapping cultures of twentieth-century genetics*, London; New York: Routledge.

Richards, T., 1993. *The imperial archive: knowledge and the fantasy of empire*, London: Verso.

Richmond, M.L., 1997. "A lab of one's own": the Balfour biological laboratory for women at Cambridge University, 1884-1914. *Isis*, 88(3), pp.422–455.

Riles, A., 2006. Introduction: In response. In A. Riles, ed. *Documents: artifacts of modern knowledge*. Ann Arbor: University of Michigan Press, pp. 1–38.

Riles, A., 2000. *The network inside out*, Ann Arbor: University of Michigan Press.

Rimmer, M., 2009. The Sorcerer II Expedition: Intellectual Property and Biodiscovery. *Macquarie Journal of International and Comparative Environmental Law*, 6, pp.147–87.

Roberts, R.J. et al., 2001. Building A "GenBank" of the Published Literature. *Science*, 291(5512), pp.2318–2319.

Roberts, R.J. et al., 2010. COMBREX: a project to accelerate the functional annotation of prokaryotic genomes. *Nucleic Acids Research*, 39(Database), pp.D11–D14.

Rose, H., 2006. From Hype to Mothballs in Four Years: Troubles in the Development of Large-Scale DNA Biobanks in Europe. *Community Genetics*, 9(3), pp.184–189.

Rose, H., 2001. *The Commodification of bioinformation: the Icelandic Health Sector Database*, [London]: Wellcome Trust.

Rose, N., 2007. Molecular Biopolitics, Somatic Ethics and the Spirit of Biocapital. *Social Theory & Health*, 5, pp.3–29.

Rose, N. & Novas, C., 2008. Biological citizenship. In A. Ong & S. J. Collier, eds. *Global Assemblages*. Blackwell Publishing Ltd, pp. 439–463.

Rosengarten, M., 2009. *HIV interventions: biomedicine and the traffic between information and flesh*, Seattle, WA: University of Washington Press.

Ross, A., 2003. *No-collar: the humane workplace and its hidden costs*, New York: Basic Books.

Rouse, J., 2007. Social practices and normativity. *Philosophy of the Social Sciences*, 37(1), pp.46–56.

Rusch, D.B. et al., 2007. The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol*, 5(3), p.e77.

Russell, C., 1999. *Experimental ethnography: the work of film in the age of video*, Durham, NC; London: Duke University Press.

Said, E.W., 1978. *Orientalism*, New York, NY: Pantheon Books.

Salzberg, S.L., 2007. Genome re-annotation: a wiki solution? *Genome Biology*, 8(1), p.102.

Sasaki, T. & Tsunoda, K., 2009. Time to revisit mumps vaccination in Japan? *The Lancet*, 374(9702), p.1722.

Sassen, S., 2002. Towards a sociology of information technology. *Current Sociology*, 50(3), pp.365–388.

Sayers, E.W. et al., 2010. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 39(Database), pp.D38–D51.

Schierwater, B., 2005. My favorite animal, Trichoplax adhaerens. *BioEssays*, 27(12), pp.1294–1302.

Schneider, J.W., 2005. *Donna Haraway: live theory*, New York: Continuum.

Serres, M., 1995a. *Angels, a modern myth*, Paris: Flammarion.

Serres, M., 2009. *The five senses: a philosophy of mingled bodies (I)*, London; New York: Continuum.

Serres, M., 1995b. *The natural contract*, Ann Arbor: University of Michigan Press.

Serres, M., 1982. *The parasite*, Baltimore, MD: Johns Hopkins University Press.

Serres, M. & Latour, B., 1995. *Conversations on science, culture, and time*, Ann Arbor: University of Michigan Press.

Shackleford, J., 1993. Tycho Brahe, laboratory design, and the aim of science: Reading plans in context. *Isis*, 84(2), pp.211–230.

Shapin, S., 1988. The house of experiment in seventeenth-century England. *Isis*, 79(3).

Shapin, S. & Schaffer, S., 1989. *Leviathan and the air-pump: Hobbes, Boyle, and the experimental life*, Princeton: Princeton University Press.

Shiel, W., 2008. *Webster's new world medical dictionary.*, Hoboken: Wiley-Interscience.

Simmel, G., 1908. Der Brief. *Österreichische Rundschau*, 15(5), pp.334–336.

Sjölander, K. et al., 2011. Ortholog identification in the presence of domain architecture rearrangement. *Briefings in Bioinformatics*, 12(5), pp.413–422.

Slater, H., 2000. The art of governance: The Artist Placement Group 1966-1989. *Variant*, 11.

Sloterdijk, P., 1999. *Regeln für den Menschenpark*, Frankfurt am Main: Suhrkamp Verlag.

Smelik, A. & Lykke, N. eds., 2008. *Bits of life : feminism at the intersections of media, bioscience, and technology*, Seattle: University of Washington Press.

Smith, B.H., 1997. Microdynamics of incommensurability: Philosophy of science meets science studies. In A. Plotnitsky, ed. *Mathematics, science, and postclassical theory*. Durham: Duke University Press, pp. 243–266.

Smith, N., 1992. Contours of a Spatialized Politics: Homeless Vehicles and the Production of Geographical Scale. *Social Text*, 33, pp.54–81.

Smith, T.F., 1990. The history of the genetic sequence databases. *Genomics*, 6(4), pp.701–707.

Spinney, J., 2006. A place of sense: a kinaesthetic ethnography of cyclists on Mont Ventoux. *Environment and Planning D: Society and Space*, 24(5), pp.709–732.

Spinney, J., 2009. Cycling the City: Movement, Meaning and Method. *Geography Compass*, 3(2), pp.817–835.

St John, J. et al., 2012. Sequencing three crocodilian genomes to illuminate the evolution of archosaurs and amniotes. *Genome Biology*, 13(1), p.415.

St. Pierre, S. & McQuilton, P., 2009. Inside FlyBase. *Fly*, 3(1), pp.112–114.

Star, S.L., 2002. Infrastructure and ethnographic practice: Working on the fringes. *Scandinavian Journal of Information Systems*, 14(2), pp.107–122.

Star, S.L., 1999. The Ethnography of Infrastructure. *American Behavioral Scientist*, 43(3), pp.377–391.

Star, S.L. ed., 1995a. *Ecologies of knowledge: work and politics in science and technology*, Albany: State University of New York Press.

Star, S.L. ed., 1995b. The politics of formal representations: Wizards, gurus, and organizational complexity. In *Ecologies of knowledge: work and politics in science and technology*. Albany: State University of New York Press, pp. 88–118.

Star, S.L. & Griesemer, J.R., 1989. Institutional Ecology, `Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science*, 19(3), pp.387 –420.

Star, S.L. & Lynch, Michael eds., 1995. Laboratory space and the technological complex. In *Ecologies of knowledge : work and politics in science and technology*. Albany: State University of New York Press, pp. 226–256.

Star, S.L. & Ruhleder, K., 1996. Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces. *Information Systems Research*, 7(1), pp.111–134.

Steedman, C., 2002. *Dust: the archive and cultural history*, New Brunswick: Rutgers University Press.

Stein, L., 2001. Genome annotation: from sequence to biology. *Nature Reviews Genetics*, 2(7), pp.493–503.

Stengers, I., 2010. *Cosmopolitics I: I. the science wars, II. the invention of mechanics, III. thermodynamics*, Minneapolis: University of Minnesota Press.

Stengers, I., 2008. Experimenting with refrains: Subjectivity and the challenge of escaping modern dualism. *Subjectivity*, 22(1), pp.38–59.

Stengers, I., 1997. *Power and invention: situating science*, Minneapolis: University of Minnesota Press.

Stengers, I., 2005. The cosmopolitical proposal. In Bruno Latour & P. Weibel, eds. *Making things public: atmospheres of democracy*. Cambridge, MA; Karlsruhe: MIT Press and ZKM.

Stengers, I., 2000. *The invention of modern science*, Minneapolis: University of Minnesota Press.

Stockinger, H., Krüger, M. & Schüßler, A., 2010. DNA barcoding of arbuscular mycorrhizal fungi. *New Phytologist*, 187(2), pp.461–474.

Stoler, A., 2002. Colonial archives and the arts of governance. *Archival Science*, 2(1), pp.87–109.

Strasser, B., 2010. Collecting, Comparing, and Computing Sequences: The Making of Margaret O. Dayhoff's Atlas of Protein Sequence and Structure, 1954–1965. *Journal of the History of Biology*, 43(4), pp.623–660.

Strasser, B., 2011. The experimenter's museum: GenBank, natural history, and the moral economies of biomedicine. *Isis*, 102(1), pp.60–96.

Strasser, B., 2003. The transformation of the biological sciences in post-war Europe. *EMBO Reports*, 4(6), pp.540–543.

Strathern, M., 2004a. *Commons and borderlands: working papers on interdisciplinarity, accountability and the flow of knowledge*, Wantage: Sean Kingston.

Strathern, M., 2004b. Laudable aims and problematic consequences, or: The "flow" of knowledge is not neutral. *Economy and Society*, 33(4), pp.550–561.

Strathern, M., 1991. *Partial connections*, Savage: Rowman & Littlefield Publishers.

Strathern, M., 2000. The Tyranny of Transparency. *British Educational Research Journal*, 26(3), pp.309–321.

Strauss, A. et al., 1985. *Social organization of medical work*, Chicago: University of Chicago Press.

Subramaniam, B., 2009. Moored Metamorphoses: A Retrospective Essay on Feminist Science Studies. *Signs*, 34(4), pp.951–980.

Suh, A. et al., 2011. Mesozoic retroposons reveal parrots as the closest living relatives of passerine birds. *Nat Commun*, 2, p.443.

Tedersoo, L. et al., 2011. Tidying Up International Nucleotide Sequence Databases: Ecological, Geographical and Sequence Quality Annotation of ITS Sequences of Mycorrhizal Fungi. *PLoS ONE*, 6(9).

Thacker, E., 2005. *The global genome: biotechnology, politics, and culture*, Cambridge, MA: MIT Press.

Thoreau, H., 2009. *The journal, 1837-1861*, New York: New York Review Books.

Thrift, N., 2005. From born to made: technology, biology and space. *Transactions of the Institute of British Geographers*, 30(4), pp.463–476.

Thrift, N., 2006. Re-inventing invention: new tendencies in capitalist commodification. *Economy and Society*, 35(2), pp.279–306.

Tooze, J., 1974. EMBO: The European Molecular Biology Organisation. *Biochemical Education*, 2(2), pp.22–22.

Traweek, S., 1988. *Beamtimes and lifetimes: the world of high energy physicists*, Cambridge, MA: Harvard University Press.

Tringe, S.G. et al., 2005. Comparative Metagenomics of Microbial Communities. *Science*, 308(5721), pp.554 –557.

Tutton, R. & Corrigan, O. eds., 2004. *Genetic databases: socio-ethical issues in the collection and use of DNA*, London; New York: Routledge.

Tutton, R., Kaye, J. & Hoeyer, K., 2004. Governing UK Biobank: the importance of ensuring public trust. *Trends in Biotechnology*, 22(6), pp.284–285.

Tweedie, S. et al., 2009. FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Research*, 37(Database), pp.D555–D559.

Ulane, C.M. et al., 2003. STAT3 Ubiquitylation and Degradation by Mumps Virus Suppress Cytokine and Oncogene Signaling. *Journal of Virology*, 77(11), pp.6385–6393.

Varda, A., 2000. *The Gleaners and I*, Cine-Tamaris.

Venter, J.C. et al., 2004. Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science*, 304(5667), pp.66–74.

Verran, H., 2009. On assemblage. *Journal of Cultural Economy*, 2(1-2), pp.169–182.

Verran, H., 2001. *Science and an African logic*, Chicago: University of Chicago Press.

Wainwright, S. & Williams, C., 2008. Spaces of speech and places of performance: an outline of a geography of science approach to embryonic stem cell research and diabetes. *New Genetics & Society*, 27(2), pp.161–173.

Wainwright, S.P., Michael, M. & Williams, C., 2008. Shifting paradigms? Reflections on regenerative medicine, embryonic stem cells and pharmaceuticals. *Sociology of Health & Illness*, 30(6), pp.959–974.

Wake, M.H., 2003. What is "Integrative Biology"? *Integr. Comp. Biol.*, 43(2), pp.239–241.

Waldby, C., 2009. Singapore Biopolis: Bare Life in the City-State. *East Asian Science, Technology and Society*, 3(2-3), pp.367 –383.

Waldby, C., 1997. The body and the digital archive: the Visible Human Project and the computerization of medicine. *Health:*, 1(2), pp.227–243.

Waldby, C., 2000. *The Visible Human Project: informatic bodies and posthuman medicine*, London; New York: Routledge.

Waldrop, M., 2008. Big data: wikiomics. *Nature News*, 455(7209), pp.22–25.

Walgate, R., 1982. Europe leads on sequences. *Nature*, 296(5858), pp.596–596.

Walz, M., 2006. Genome campus grows on all fronts. *R&D Magazone*.

Wang, K., 2006. Gene-function wiki would let biologists pool worldwide resources. *Nature*, 439(7076), pp.534–534.

Warner, Marina, 2011. *Stranger magic: charmed states & the Arabian nights*, London: Chatto & Windus.

Warner, Michael, 2002. *Publics and counterpublics*, New York: Zone Books.

Whatmore, S., 2002. *Hybrid geographies: natures, cultures, spaces*, London; Thousand Oaks; New Delhi: SAGE Publications.

Whatmore, S., 2009. Mapping knowledge controversies: science, democracy and the redistribution of expertise. *Progress in Human Geography*, 33(5), pp.587–598.

Wiegler, L., 2007. EBI Expands Hinxton Facility to Keep Ahead of the Bioinformatics Data Curve. *Bioinform*, 11(42).

Wilkie, A. & Michael, M., 2009. Expectation and mobilisation enacting future users. *Science, Technology & Human Values*, 34(4), pp.502–522.

Williams, A.J. ed., 2011. *Nancy Holt: sightlines*, Berkeley, CA: University of California Press.

Woolgar, S. et al., 2009. From scale to scalography: a provocation piece. In International workshop on scalography. Said Business School, Oxford.

Woolgar, S., 1991. *Knowledge and reflexivity: new frontiers in the sociology of knowledge*, London; Thousand Oaks; New Delhi: Sage.

Wynne, B., 2005. Reflexing complexity post-genomic knowledge and reductionist returns in public science. *Theory, Culture & Society*, 22(5), pp.67–94.

Ybema, S. & Kamsteeg, F., 2009. Making the Familiar Strange: A Case for Disengaged Organizational Ethnography. In S. Ybema et al., eds. *Organizational ethnography : studying the complexities of everyday life*. Los Angeles; London: SAGE.

Yusoff, K., 2009. Excess, catastrophe, and climate change. *Environment and Planning D: Society and Space*, 27(6), pp.1010 – 1029.

Zehetner, G. & Lehrach, H., 1994. The Reference Library System — sharing biological material and experimental data. *Nature*, 367(6462), pp.489–491.

Zimmerman, A.S., 2008. New knowledge from old data: The role of standards in the sharing and reuse of ecological data. *Science, Technology & Human Values*, 33(5), pp.631–652.