Title

PhD Thesis - From Open Data to Knowledge Production:Biomedical Data Sharing and Unpredictable Data Reuses

Author

Pasquetto, Irene

Publication Date

2018

UNIVERSITY OF CALIFORNIA

Los Angeles

From Open Data to Knowledge Production:

Biomedical Data Sharing and Unpredictable Data Reuses

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Information Studies

by

Irene Pasquetto

2018

ABSTRACT OF THE DISSERTATION

From Open Data to Knowledge Production:

Biomedical Data Sharing and Unpredictable Data Reuses

by

Irene Pasquetto

Doctor of Philosophy in Information Studies

University of California, Los Angeles, 2018

Professor Christine L. Borgman, Chair

Using a US consortium for data sharing as the primary field site, this three-year ethnographic research project examines the socio-technical, epistemic, and ethical challenges of making biomedical research data openly available and reusable. Public policy arguments for releasing scientific data for reuse by others include increasing trust in science and leveraging public investments in research. In most types of scientific research, data release occurs in parallel with associated publications, after peer-review. In the consortium studied for this project, datasets may also be released independently without an associated publication. Such research datasets are conceptualized as "hypothesis free" resources from which novel knowledge can be extracted indefinitely. Among the findings of this project are that biomedical researchers do not download and re-analyze "hypothesis free" research data from open repositories as a regular practice. Data reuse is a complex, delicate, and often time-consuming process. Metadata and

ontology schemas appear to be necessary but not sufficient for data reuse processes. For scientists to test new hypotheses on "old" data, they depend on access to peer-reviewed primary analyses, pre-existing trusted relationships with the data creators, and shared research agendas. Data donors (patients, study participants, etc.), on the other hand, retain little control over how open research data are reused. Findings suggest that, in practice, it is impossible to predict – and consequently to regulate – how datasets might be reused once made openly available. Unintended consequences of reusing this consortium's open data already are emerging, to the concern of some participants.

This dissertation of Irene Pasquetto is approved.

Soraya de Chadarevian

Christopher M. Kelty

Leah A. Lievrouw

Christine L. Borgman, Committee Chair

University of California, Los Angeles

2018

Table of Contents

My doctoral work at UCLA was supported by Alfred P. Sloan Foundation grant *If Data Sharing is the Answer, What is the Question?* Christine L. Borgman, Principal Investigator, Award# 2015-14001 (July 2015-June 2018), and by a Dissertation Year Fellowship, awarded by the UCLA Graduate Division.

Irene Pasquetto is a Ph.D. Candidate in the Department of Information Studies at UCLA, and a research assistant at the UCLA Center for Knowledge Infrastructures and also at the Participation Lab (PartLab) in the UCLA Institute for Society and Genetics. She teaches and co-teaches classes on data management, information ethics, and data economies. Her overarching research interest lies in the analysis of data-centric practices and technologies, especially in relation to science and technology policy-making. For her dissertation research, Irene investigated the epistemic, socio-technical, and ethical challenges related to the open circulation of biomedical data and software. She has also published on issues of genetic ancestry testing, open data, and identity, on the challenges of archiving and reusing climate-change data. With her work, Irene aims at informing the design and implementation of governance models for data and code infrastructures.

1. Introduction

"Fungibility" is the property of a good or a commodity to be freely exchangeable or replaceable, in whole or in part, for another of like nature or kind. Oil is fungible.

The primary goal of this study is to construct an account of how biomedical researchers evaluate and reuse "open data" – specifically data they did not collect themselves – to produce novel scientific knowledge, and to show how this delicate process is connected to issues such as scientific credibility, academic capital, and the effective creation of novel scientific hypotheses. In this project, I use the expression *open data* to refer to publicly funded research data made freely available in open repositories for anyone to reuse.

Funding bodies in the biomedical domain want to know how the scientists reuse open data. Historically, biomedical research is a data rich domain (García-Sancho, 2015; Leonelli, 2016; Strasser, 2012). In the last few decades, the field heavily invested in the collection and accumulation of "data resources" that the research community uses to advance biomedical knowledge. Examples include the many databases and tools accessible through the National Center for Biotechnology Information (NCBI) website. Today, the National Institutes of Health (NIH), the major US funding agency for biomedical research, increasingly promotes and encourages the reuse of existing open

data resources (National Institutes of Health, 2017a). The NIH's focus on data reuse is motivated first of all by the fact that maintaining many, large, and diverse sets of data (i.e., Big Data) "alive" is extremely expensive, time consuming, and requires a specialized workforce. NIH's 2018 *Strategic Plan for Data Science* identified a set of current challenges for the advancement of biomedical research (National Institutes of Health (NIH), 2018). At the top of the list are the growing costs of managing biomedical large-scale datasets, and the difficulties to integrate "siloed" data resources (Wilkinson et al., 2016). Initiatives such as the NIH Big Data to Knowledge (BD2K) and the NIH Data Commons programs aim at making sure that novel knowledge is harvested from NIH-funded data repositories (Margolis et al., 2014).

The main case study of this dissertation project is a NIH-funded consortium for data sharing that I refer to as the DataFace Consortium (DF). I use the fictional name "DataFace" to protect the confidentiality of the research participants. The researchers participating in DataFace collect experimental and observational biomedical craniofacial data from humans and three model organisms. Data types include facial images, facial measurements, metrics and statistics for facial traits analysis, gene/RNA expression data, whole-genome sequences, association results from Genome-Wide Association Studies (GWAS), and results from function-validation studies. I refer to these data as "craniofacial research data." In collaboration with "the DataFace engineering hub," the researchers release the datasets in open access after data collection. The datasets are available in a digital repository for anyone to reuse, in both research and commercial settings. I employ this case study to investigate the following research questions:

1. What motivates the design of policies and infrastructures for open research data?

2. How do researchers reuse open research data for knowledge production?

3. What are the societal implications of making available and reusing open biomedical data across contexts of production?

This is a multi-layer investigation of interdisciplinary nature. First, this project brings to surface the existence of different "data governance regimes" that regulate how and when biomedical research data are made publicly available for reuse. By engaging with the literature from the history of science and the social studies of science, I examined the emergence of multiple data governance regimes in the biomedical domain during the twentieth-century in the US (De Chadarevian, 2004; Hilgartner, 2017; Kelty, 2012; Stevens, 2013; Strasser, 2011). This analysis suggests that the biomedical domain is shifting from a "semi-openness" toward a "radical openness" regime for data sharing. In a radical openness regime, research data are conceptualized as "hypothesis free" resources and made available in open repositories, sometimes prior to the publication of the associated primary analyses. As I discuss, this regime is challenging well-established data practices and community norms on several fronts.

Second, this study examines the socio-technical and epistemic factors that scientists take into account when they reuse others' data available in the public domain. I employed concepts and methodologies from the interdisciplinary field of *data studies* to develop a typology of data reuse practices for biomedical research data (Borgman, 2015; Leonelli, 2016; Wallis, Rolando, & Borgman, 2013). Scholars in the field of data studies posed the

question: How do scientists trust others' data? In my examination, I take a step back and I ask: How do scientists reuse others' data? From the analysis of observational and interview data, I derive a typology of data reuse practices, which differentiates between "background" and "foreground" reuse of others' data. This typology is used to explain how the epistemic and socio-technical challenges of data reuse vary depending on *the purpose* of the reuse. The vast majority of data reuse is for background purposes such as comparing experimental data to validated measures from trusted sources. Researchers conduct reviews of datasets much as they conduct literature reviews, and often assemble summary-level datasets and literature in combination. Reanalyzing a "raw dataset" or integrating data from multiple sources is a much more complex endeavor and done much less often (i.e., foreground reuse). When scientists wish to reanalyze data from other laboratories to produce novel knowledge, which is foreground use, they typically collaborate directly with the data creators, coauthoring papers.

Third, my investigation adds a critical lens to the current debate on the public benefits of open research data. Often, open data initiatives symbolize a reaction against the view of scientific knowledge production as an esoteric, technical and overspecialized process and promote the idea that scientific knowledge can and should be investigated as a whole (Leonelli, 2016). Also, making data freely and legally available is seen as a way to foster public trust in science as a source of reliable knowledge and legitimate source of information (Borgman, 2015). In the hype for open data, it is important to recognize that there could be unforeseen societal burdens to making research data openly available, especially when research data are collected from human subjects. Beyond obvious issues

of privacy, data reusers need to face novel questions of data ownership and control (Radin, 2017b, 2017a). DataFace human subjects datasets (DNA samples and 3D facial images) are being reused to train machine-learning algorithms that associate human facial features to genetic markers. In turn, these algorithms inform research on the design of Forensic DNA Phenotyping (FDP) technologies, which are employed by law enforcement agencies to reconstruct faces of suspects from DNA samples left at the crime scenes. The use of FDPs services in criminal investigations has been criticized by scholars in several fields, including biomedicine, bio ethics, legal studies, and the social studies of science (Toom et al., 2016). By "following the data," I examined the process through which novel analyses of few DataFace datasets informed research on DNA-based facial reconstruction, and I further discuss the scientific debate surrounding this case study. My findings show that it is impossible to predict how research data will be reused once these are made openly available.

2. Theoretical framework

Scientific knowledge does not originate in a vacuum, but is constructed in a net of social relationships, artifacts, technical innovations, and personal and collective interests (Kuhn, 1970). In this project, the expression "knowledge infrastructure" (KI) refers to the complex socio-technical environment in which scientific knowledge is produced, shared, used and reused (Borgman, Darch, Sands, & Golshan, 2016; Bowker, 2005; P. N. Edwards, 2010; Star, 1995). In what follows, I describe the characteristics of a KI, and of how it differs from other types of infrastructures. I then summarize few useful conceptualizations of what constitute "research data," and of how research data might be distinguished from other types of data. Finally, I operationalize three expressions that constitute the building blocks of this project: Data Reuse, Data Sharing, and Open Data.

Knowledge infrastructures in science

In social science, "infrastructure" is defined as a relational entity that emerges for people in practice and structure (Star & Ruhleder, 1994). In a foundational paper, Star and Ruhleder (1994) provided a list of key features that characterize infrastructures. Infrastructures are *relational* and *distributed* systems. Different perspectives, standards, conventions of practice, and cultural and organizational challenges need to be in place for the infrastructure in order to function. An infrastructure is built upon other layers, and, at the same time, is shaped and constrained by its relations to them. In this sense, infrastructures are embedded in other structures, social arrangements and technologies. To actively participate in an infrastructure is neither "natural" nor "automatic" for participants, but is something that is *learned* as a part of a membership within particular professional, social or cultural communities. In order to function, an infrastructure needs

to be woven into the *daily practices* of the workers. Because technologies, humans and policies involved in the infrastructure constantly change, also infrastructures themselves keep evolving. Indeed, infrastructures are *unstable* systems. Reformulating Castells and Hughes' work in the history of science and technology, researchers developed a general model for understanding the evolutions of infrastructures over time (P. N. Edwards, 2010; Hughes, 1983; Jackson, Edwards, Bowker, & Knobel, 2007; Castells, 1996). Scholars showed that infrastructures are initially about the accomplishment of *scale*, as they grow into *networks*. During their formation, infrastructures are sites of intense conflict. Discrepancies in the fundamental experience and vision of infrastructures start to emerge and materialize especially in the relation to designer assumptions and user expectations (Bowker, Baker, Millerand, & Ribes, 2009). Also, developing infrastructures can disconnect existing institutional, legal, and property regimes. If they survive this initial phase, infrastructures go through an intermediate phase in which they adapt and mutate, to finally become heterogeneous systems linked to each other via the consolidation of agreed upon gateways (e.g. standardizations). Once in place, provisional winners and losers are established. At this stage the infrastructure may reach a moment of apparent stability; it is in this stage that the infrastructure "disappears," and their products are taken for granted (Bowker, 2005; P. N. Edwards, Jackson, Bowker, & Knobel, 2007). To provide an example, let us consider a scientific database as an element of an infrastructure. Once a database is widely adopted, as in the case of GenBank in the life science domain, it may be taken for granted because scientists know they can rely on it. However, once the database breaks down or stops functioning as expected, its existence suddenly "reappears" to its users. During moments of breakdown or upheaval, different

layers of the infrastructure are exposed. These phases represent unique opportunities for scholars who study infrastructures. By *going backstage* to observe the infrastructure in its making, practicing, and breaking, scholars analyze the relationships across its multiple components, a process often refers to as the method of *infrastructural inversion* (Bowker, 2005). Finally, I want to draw attention to the concept of *momentum*, which signifies the necessary condition under which the infrastructure develops in trajectories and path dependencies (Jackson, Ribes, & Buyuktur, 2010). Because KIs evolve in path dependencies, once the infrastructure takes a direction for its grow it is hard to change it (Jackson et al., 2010).

Infrastructures can be of many types. In this dissertation, I analyze how data are shared and reused in a "knowledge infrastructure" (KI) in the biomedical science. Contrary to cities' infrastructures, a KI's main goal is not to move cars or metro cabins, but to allow the production and flows of knowledge. Beyond science, the notion of KI can be used to study knowledge production and circulation in a variety of contexts, such as in finance, education, and so on. In here, I employ Edwards's definition of a KI (2010, p. 17): "Robust networks of people, artifacts, and institutions that generate, share and maintain specific knowledge about human and natural worlds." In a knowledge infrastructure, participants often adopt computer-based technologies to move knowledge around. KIs' goals can include to organize and provide online access to digital resources (data, code, visualization tools etc.), foster multidisciplinary collaborations, and allow remote work interactions (Borgman, Wallis, Mayernik, & Pepe, 2007; Jackson et al., 2007; Olson & Olson, 2000). In science computer-based KIs, technologists and domain

experts need to collaborate to make knowledge flowing. In a KI, those who are in charge of building the technical infrastructure (computer engineers, software developers, data managers etc.) need to collaborate and exchange information with those who will be using the information to produce knowledge (scientists, clinicians etc.) (Bowker et al., 2009; Ribes & Bowker, 2009).

Sometimes, KIs are referred to as Cyberinfrastructures (CIs), especially in US scientific parlance (Bietz, Baumer, & Lee, 2010; Borgman, Bowker, Finholt, & Wallis, 2009; Jackson et al., 2007; Ribes & Lee, 2010). In policy and technological environments, CI is sometimes used to refer exclusively to the physical or technological aspects of an infrastructure (Atkins et al., 2003). Also, while "CI" is mostly an American terminology, in Europe researchers often refer to science collaborative projects as e-science projects (Atkins et al., 2010; Carusi et al., 2010; David & Spence, 2003).

In studying scientific KIs, researchers in information studies and in the social studies of science adopt a "socio-technical approach" to the analysis of the work practices of the scientists and their artifacts, such as notebooks, datasets, software, and so on. The term "socio-technical" is used to refute the idea that social and technical phenomena are distinct and contradictory (Star, 1995). By taking distance from this dichotomy, social researchers aim to acknowledge that social and technical problems and solutions always exist in a relation of co-dependency. As Bowker et al. (2009) pointed out, in studying scientific practices, instead of wondering whether a problem is of a social or technical nature, we should focus on understanding if the proposed solution, for any give problem,

is primarily social, technical, or a combination of both. Sociotechnical analyses of scientific practices normally rely upon qualitative methods for data collection and analysis, typically ethnographic observations and interviews.

Research data and evidentiary power

To provide an in-depth analysis of the notion of "data" is beyond the scope of this dissertation project. However, it is in its interest to discuss whether *research data* are in any ways different from other kinds of data.

Scholars in the social studies of science investigated what is called "the demarcation problem" for a very long time. The demarcation problem refers to the possibility of defying scientific knowledge as in any way different from non-scientific knowledge. While this is an open and complex debate (Collins & Evans, 2008), we can at least say that – generally – scientific knowledge is researched by social scientists as a kind of knowledge that is produced and "certified" by individuals that are generally recognized as "experts" in a field of research. Building on this general statement, we can define research data as "entities used as evidence of phenomena for the purposes of research or scholarship (i.e., science)" (Borgman, 2015, p. 29). Theoretically, as Borgman observes, any artifact can be a used as a piece of research data; it all depends on what counts as a piece of evidence in the first place. In the biomedical community, data are certainly treated as a piece of evidence when they are published along with peer-reviewed academic publications. Indeed, the whole point of replication studies is to re-calculate others' data to verify consistency of results. It is in this sense that research data are used as evidence for reproducible knowledge.

The key role that certified biomedical knowledge plays in our everyday lives does not need much explanation. If a group of non-experts creates a new drug with their own Do It Yourself (DIY) biology kit, and they want to sell it, they can't. Obviously, to be commercialized, drugs need, among other things, to be approved by the Food and Drugs Administration (FDA) agency. Similarly, if I want to try that new shiny drug I saw in a TV commercial the other day, I need a prescription from my primary care physician. The FDA and the physicians rely on certified medical knowledge – peer-reviewed publications – to decide whether to approve or not a drug for commercial use, or to prescribe or not a drug to a patient. The final evidence that a certain drug does what it is supposed to do rests in the research data (and in the research design) underlying these publications.

We can then think that the defining trait of research data – as opposed to any other kind of data – is that they perform as evidence for certified knowledge, which means knowledge that is peer-reviewed and published. However, the majority of data used for research purposes, especially raw and negative data, are not available in public repository and cannot be consulted as pieces of evidence. Most research data never leave private laptops and universities servers (Wallis et al., 2013). Indeed, only the result data are usually published along with academic publications (Borgman, 2015). Under certain circumstances (e.g. data abundance), researchers share their "unpublished" data with their colleagues prior to publication, traditionally in the context of closed collaborations.

One could also suppose that there is no stable defying trait in any kind of data, including research data. Following a well-established tradition in the social studies of science, scholars have argued that all types of data are imperfect and arbitrary representations of facts in the world. As pointed out by media scholars Gitelman and Jackson (2013), data are always "cooked," which means subjected to human manipulation and, consequently, data are in a way or another biased interpretations and not objective representations of reality. Following this argument, we can then affirm that knowledge is always produced rather than innocently discovered (Gitelman, 2013, p. 4). Deriving from Daston and Galison's (2007) conceptualization of mechanical objectivity, Gitelman and Jackson argued that objectivity is historically situated and culturally specific: "It comes from somewhere and it is the result of ongoing changes to the conditions of inquiry, conditions that are material, social and ethical" (Gitelman, 2013, p. 4). Sociologists of science famously argued that scientists use data as "rhetorical devices" to validate, certify and mobilize knowledge (Day, 2014; Latour, 1987; Rosenberg, 2013). Science data are then "cooked" by individuals working in highly specialized cultures of practice (Galison, 1987; Knorr-Cetina, 1999; Rheinberger, 1997). It has also been argued that "translating" data between contexts requires a compromise on issues of quality and accuracy. In this perspective, data go through a process of simplification and approximation to be shared and reused by scientists operating in different communities (Bowker & Star, 1999; H. M. Collins, Evans, & Gorman, 2007; Latour, 1987).

In her latest book "Data-centric Biology," Leonelli (2016) proposes that research data do not have stable defying traits, but their evidentiary power lies instead in the ways in

which these are "packaged" for reuse. For Leonelli, who specifically writes about biology data produced in the context of animal model communities, the evidentiary power of research data does not lie in their apparent "immobility," as Latour and Rheinberger argued (Latour, 1987; Rheinberger, 1997), but it lies in the possibility of repurposing them to ask novel research questions, across different experimental settings. Interestingly, Leonelli observes, the fact that a given research dataset can be used "to prove" the existence of multiple biological phenomena at once does not take anything away from its intrinsic validity. Indeed, Leonelli argues, researchers are well aware of the fact that research data are not simply "found" in nature, but they are "crafted" in ways that make them usable for research purposes. Data creators craft, or cook, the data themselves, and they do so intentionally. The trick lies in being aware of how data are crafted, knowing how they are crafted allow us to make judgment about their validity. In this frame, the data's evidentiary power, and their potential for reuse, is strictly dependent on how the data are described by use of metadata and ontologies. By using ontologies and metadata, data professionals describe and characterize data in ways that make them potentially useful to ask a number of questions, by a number of researchers. This means that multiple "facts" can be associated to a single dataset at once.

Terminology specifications: data reuse, data sharing, open data

As it is often the case, data reuse can mean many different things, for different people. The most fundamental problem in defying data reuse is to distinguish between a "use" and a "reuse." In the simplest situation, data are collected by one individual, for a specific research project, and the first "use" is by that individual to ask a specific research question. If that same individual returns to that same dataset later, whether for the same

or a later project, that can be considered a "reuse." However, this is not the kind of "reuse" that I aim to study in this dissertation. When that dataset is contributed to a repository, retrieved by someone else, and deployed for another project, I then consider it a case of "data reuse." Not by chance, in the common parlance of data practices studies, reuse usually implies the usage of a dataset by someone other than the originator (Carlson & Anderson, 2007; Rung & Brazma, 2012; Zimmerman, 2008).

This type of data reuse – *reuse by others* – is probably the most troublesome of all data reuses. This is because it implies that researchers need to trust data that have been collected by someone else, data whose production they did not personally witness. While data reuse is a well established practice in certain research communities, such as among model organism communities (Leonelli, 2016), it is perceived as a real challenge among others. Socio-technical studies on data reuse practices unpacked patterns of data reuse for open repositories (or digital archives) of research data. This line of research showed that scientists carefully evaluate others' data in order to take decisions on whether reusing them, or not. Scientists need to be able to "understand" and especially "trust" others' data in order to reuse them (Faniel & Jacobsen, 2010; Zimmerman, 2007). Several factors play key roles in enabling understanding and trust of others' data, such as the popularity of the repository in which data are stored, the granularity and accuracy of data curation, and the perceived reliability of the individuals who collected the data in the first place (Yoon & Kim, 2017). Overall, granular and detailed "data curation" (i.e., use of metadata to describe and characterize others data) is perceived as a main factor that enables data reuse

(Wilkinson et al., 2016). Also tools for data discovery and analysis might encourage data reuse (Goodman et al., 2014).

People can reuse others' data for many purposes, such as for outreach, in the context of replication studies, etc. In this dissertation project, I look at how scientists reuse *others* data for *knowledge production*, and, in particular, to ask novel research questions. The possibility of re-purposing old data to squeeze out new knowledge is endemic of the "big data" imaginary (boyd & Crawford, 2012). This is what Leonelli refers to as the promise of making data "fungible," meaning "interchangeable units" that travel from context to context, from research questions to research questions, by carrying infinite potential for novel knowledge extraction (Leonelli, 2016). In biomedical research, and in particular in policy context, the act of re-analyzing others' data to run novel research questions is referred to as "secondary analysis" of data (National Institutes of Health, 2017a). The "primary analysis" is the analysis conducted by the data creator herself, in relation to the research question the data were originally collected for. In this context, "secondary" does not have a demeaning connotation.

Another problematic term is *data sharing*. Data sharing generally refers to the act of releasing data in a form that can be used by other individuals. Data sharing thus encompasses many means of releasing data, but says little about the usability of those data. Examples of data sharing in science include private exchanges between researchers; posting datasets on researchers' or laboratory websites; depositing datasets in archives, repositories, domain-specific collections, or library collections; and attaching data as

supplemental materials in journal articles (Wallis et al., 2013). A relatively newer practice in many fields is to disseminate a dataset as a "data paper." Methods of data sharing vary by domain, data type, country, journal, funding agency, and other factors. The ability to discover, retrieve, and interpret shared data varies accordingly (Borgman, 2015; Leonelli, 2016; Palmer, Weber, & Cragin, 2011). In this dissertation project, I use "data sharing" to refer to the act of scientists consensually and intentionally sharing data within each other, upon-request and in the context of inter-laboratories collaborations.

I am especially eager to differentiate "data sharing" from "open data." Open data is a problematic term given the array of concepts and conditions to which it may refer (Levin, Leonelli, Weckowska, Castle, & Dupré, 2016; Pasquetto, Sands, Darch, & Borgman, 2016; Pomerantz & Peek, 2016). Baseline conditions for making scientific data "open" usually refer to "fewest restrictions" and "lowest possible costs." Legal and technical availability of data are also mentioned (Open Knowledge Foundation, 2015; Organisation for Economic Co-operation and Development, 2007). The OECD specifies 13 conditions for open data, only a few of which are likely to be satisfied in any individual situation (Organisation for Economic Co-operation and Development, 2007). Examples of "open data initiatives" in the academy include repositories and archives (e.g.,, GenBank, Protein Data Bank, Sloan Digital Sky Survey), federated data networks (e.g.,, World Data Centers, Global Biodiversity Information Facility; NASA Distributed Active Archive Centers), virtual observatories (e.g.,, International Virtual Observatory Alliance, Digital Earth), domain repositories (e.g.,, PubMedCentral, arXiv), and institutional repositories (e.g.,, University of California eScholarship). However, in these projects, openness varies

in many respects. Public data repositories may allow contributors to retain certain intellectual property rights over the deposited. Data may be open, but interpretable only with proprietary software. Data may be created with open source software, but require licensing for data use. Open data repositories may have long-term sustainability plans, but many depend on short-term grants or on the viability of business models. Keeping data open over the long term often requires continuous investments in curation to adapt to changes in the user community (Baker, Duerr, & Parsons, 2015). A promising new development to address the vagaries of open data is the FAIR standards – Findable, Accessible, Interoperable, and Reusable data (NIH, 2016). These standards apply to the repositories in which data are deposited. The FAIR standards were enacted by a set of stakeholders to enable open science, and they incorporate all parts of the "research object," from code, to data, to tools for interpretation (NIH, 2016; Wilkinson et al., 2016). Beyond idiosyncratic features of specific open data initiative, in here I employ the term "open data" to refer to those initiatives in which data are contributed to a repository of some sort, as opposed to "shared" upon request in between individuals and labs.

3. Literature review

The analysis of literature relevant to this dissertation project is divided in three sections. First, I review the emergence of different data sharing regimes in the field of biology during the twentieth-century. As observed by Hilgartner (2017) in his latest book on this topic, this story has received little attention compared to that paid to the concomitant rise of intellectual property rights. Hilgartner (2017) pointed out that the twentieth-century led to a profound historical transformation in the life sciences, namely the development of public access and data-sharing policies. Before the advent of the "genomics revolutions" in the late 80s, biomedical research data used to be made available in open repositories "post publication." Specialized communities of researchers (e.g., model organism communities) would curate and reuse these data in their daily research practices. The scientists would submit the data to the open repositories either on a voluntary basis, or – later – as a requirement for publication. Data would be shared "prior to publication" solely within semi-closed collaborations in which scientists would retain some control – and credit – over the reuses of their data. In the late 90s, the genomics research community saw the emergence of what Hilgartner refers to as the "knowledge-control regime" of Unpublished in Journal, Available in Databases (UJAD) research data. Toward the completion of the Human Genome Project, the NIH started to require scientists to deposit their sequence data in publicly available repositories right after data collection, which often meant "prior to publication," as a requirement for ensuring continuous funding. In what follows, I examine some of the factors that led to the adoption of policies for depositing research data in open repositories prior to publication, in the US.

In the second section I examine what it means to reuse research data "ethically," the problems of data ownership, and of the limits of Informed Consent. Finally, I conclude this literature review with an analysis of the epistemic implications of recent innovations in human genetics – "personalized medicine" and "population genetics" – that are used by the craniofacial researchers to engage and make sense of the DataFace datasets. I also review some ethical concerns that scholars in the fields of the social studies of science are raising in relation to these emerging methodologies.

Pre-genome perspectives on data sharing and reuse

I start my analysis by reviewing some "classics" in the history and anthropology of science that narrate the ways in which science stakeholders would regulate the practices of data sharing and reuse before the advent of the "genomics revolution" in the late 80s. The co-presence of two competing perspectives on "how and when" biomedical researchers should share and reuse data characterized the early practices of genetics research and, later, molecular biology. On one hand, some science actors promoted the idea of knowledge production as a collective effort, and of research data as "collective properties." Promoters of research data as collective properties would make data available in semi-open systems in which only those who belong and contribute to the community would have access to the data. On the other hand, the promoters of full open access to research data would argue that anyone should have access to anyone's data at all time, and that academic credit should go to those able to exploit the data, not to those who collected it. In this sense, research data are resources that anyone should be able to access and exploit – a logic similar to the one of free market and open competition. In

this regime of "full openness," the merit for the data analysis goes to the individual scientists, not to a collective of scientists. Importantly, while my analysis show the co-presence of different perspective on open data before the genomics era, it also pointed out at the fact that neither was at this point formalized in official policies for data sharing.

The drosophila community: a practice-driven openness

Practices of organizing, disseminating, and reusing data have long characterized biology research (Kelty, 2012; Clarke & Fujimura, 2014; De Chadarevian, 2004). Biologists, especially those working in model organism communities, have been sharing their data in between labs since the early days of molecular biology. Robert Kohler famously narrated the rise of data sharing and reuse practices among the members of the drosophila community in the USthroughout the first half of the 20th century (Kohler, 1994). At the time, the creation of the laboratory-standard fruit fly (i.e., the drosophila) reconfigured biology methodology and profession. The drosophila was one of the most productive of all laboratory animals. From 1910 to 1940, the center of Drosophila culture in America was the school of Thomas Hunt Morgan and his students Alfred Sturtevant and Calvin Bridges. Morgan's lab first created the standard flies, through inbreeding, and then organized a network for exchanging stocks of flies in between labs that spread their practices around the world. By rearticulating Edward Palmer Thompson's work, Kohler (1994, p. 12) argued that the drosophila research community had a "moral economy" of openness, defined as "a set of moral conventions that regulates the access to tools of trade and the distribution of credit and rewards for achievement." As explained by Kohler, the drosophila scientists were managing resources, such as stocks, tools, storage practices,

and data, in ways that allowed other labs working on the heredity research to access the same resources.

In the drosophila community, data sharing and reuse practices were implemented as a response to practical needs, and, at the same time, they functioned as a promotional means for spreading novel research methodologies. The drosophila community was producing a great amount of mutants to be examined, more than enough for everybody to study; openness helped to optimize this analysis. Most importantly, among drosophila researchers, sharing tools and reusing each other's data became a sign of affiliation with that community. Access and diffusion of the technical tools and data brought their inventors visibility, trust, and prestige, and the drosophila exchange system became the material basis for recruitment, employment and communication within the discipline. As a result, the free availability of drosophila helped to establish the epistemic consolidation of the new practice of "genetic mapping" as the dominant form of experimental heredity research. The drosophila community moral economy of openness was mainly motivated by practical needs of convenience and efficiency, the labs were simply producing more data than they could possibly analyze, and eventually became a widely adopted ethical standard across the whole drosophila community.

Competition and collaboration in semi-open communities

Among drosophila labs, research data and other materials "were regarded as communal property" (Kohler, 1994, p. 133). By employing the *Drosophila Information Service (DIS)* newsletter as "model" to study open science ethos, Kelty (2012) revealed interesting details behind this understanding of research data as "communal proprieties."

In the twentieth-century, drosophila scientists were using the DIS newsletter to communicate and monitor progress in the field. Kelty argued that the newsletter was regulated by a strategy of information sharing that was, at the same time, both open and closed. The newsletter, and its information, was not open "to just anyone." DIS's subscribers would have unrestricted access to others' flies, techniques, results and other information only at the condition that they would first share a list of mutants fruit flies available in their labs, and thereby were willing to share these mutants by mail or in person with other labs. By contributing with their own mutants, tools, and concepts, the DIS subscribers would become "actively engaged" members of the community. At the same time though, this mechanism isolated those geneticists who were not primarily working with drosophila and didn't have anything to contribute with. Kelty observed that, precisely because this semi-open strategy for information sharing was in place, "the newsletter became the *de facto locus* for the construction of a recognizable and stable research collective – a community, a paradigm, a tradition and so on with stable concepts and epistemic objects contributed by and collectively owned by Drosophila labs around the world" (Kelty, 2012, p.147).

Significantly, as Kelty pointed out, this mixed open/close regime for data sharing made the newsletter a relatively "safe place" for the scientists to share preliminary experimental results before these would be "good enough" for official publications. In Kelty's words: "Scientists could signal each other about problems they owned without fear of getting scooped" (Kelty, 2012, p. 147). Because subscribers were requested to share their data with the whole DIS community as a condition to access *others'* data,

power and vulnerability were equally distributed among all members of the community. At any time, anyone could scoop anyone else. In this context, a sort of "do not do unto others what you do not want others to do unto you" ethos was preventing scientists from reusing each other data in ways that could be perceived as unfair.

The notice printed front and center on the newsletter's cover seemed to reinforce such ethos: "This is not a publication – Unpublished material presented in this circular must not be used in publications without the specific permission of the author." As discussed by Kelty, the newsletter editors made explicit that the newsletter was a tool to disseminate information *privately* among the community members, but not to the general public, and that the newsletter would not engage, for this specific reason, in the practices of scholarly citations. Interestingly, Kelty noticed that already in the foreword of second number of the newsletter's editors felt the need of re-articulated the cover statement. Few readers apparently interpreted the request for permission before using others' material in publications as a barrier to the open circulation of knowledge, and, for this reason, as being in contradiction with the main goal of the newsletter itself. In the foreword, the authors asked to the subscribers to share only material and information that they would feel confortable sharing "by mail with another Drosophila worker." Most importantly, they specified that, actually, explicit permission from the authors for data and information reuse in novel publications was not needed, and that acknowledgment of the "source" of the material was enough. During the following years, norms and expectations concerning free exchange and acknowledgements would be frequently repeated in the foreword of the DIS. At the same time, no copyright notice was ever added to the DIS, and the

informal rule of acknowledgment successfully regulated the exchange of info via the DIS for over 50 years. The DIS became so popular that it was officially turned into a journal toward the end of the century.

As discussed by Kelty, at the heart of the success of the DIS it was its feature of being open and closed at the same time, of sharing data and materials only with "actively engaged" DIS members (i.e., members who were willing to share their materials in the first place). This "partial porousness" – as Kelty's calls it – of the DIS community allowed the scientists to work, at the same time, in an environment of both competitiveness and cooperation. In this particular knowledge-control regime, DIS scientists would manage the sharing and access to research data and materials as "collective proprieties."

Kelty also compared the moral economy of the DIS with the moral economy of the contemporary synthetic biology community. He found some similarities, especially the fact that also synthetic biology relies on a similar open/closed regime of knowledge access. However, Kelty explains, in the case of synthetic biology the mixed open/closed regime is motivated by a different ethos. Synthetic biology is indeed a "commercially driven science" whose data sharing and reuse practices are highly impacted by the pressures of contemporary intellectual property-saturated biotechnology market. To do their work, Kelty argues, synthetic biologists must depend on the biotechnology industry for tools and financial investments. The biotechnology industry is dominated by patents, as opposed to copyright, and a structured system of investment and return. In this frame,

data and other research objects are not "collectively owned" by synthetic biology communities, but by individuals and investors. He explains:

> Whereas the closure of newsletters was intended to facilitate the creation of collectively owned concepts in the service of a cumulative science, the intellectual property system recognizes no such thing: all concepts, techniques, objects, practices, must be individually owned – subject to the intellectual property regime's definition of an individual and his/her/its rights. Even though the intent of the intellectual property system may once have been to balance individual gain with public benefit (Boyle, 2008; Hyde, 2010), the reality of the system as implemented is that everything, down to the very mutant fly and its sequenced gene, must be individually owned in order to serve the growth of a competitive market (Kelty, 2012, p. 162).

Centralized open collections and community databases

Starting with the second half of the twentieth-century, the organization of biology knowledge in digital databases played a central role in enabling data distribution and reuse practices in the field (Kelty, 2012, p. 162). Strasser (2011) argued that contemporary biology inherited the practice of organizing research outputs into structured knowledge representation schemas, such the digital databases, from the natural history tradition of collection and cataloguing "natural facts" about world. Natural history approach relied on "collections" as primary means for knowledge production. Bringing specimens together in a single place and organizing them in a systematic way made comparisons among different sets of data possible. By facilitating analogical reasoning, databases enabled the identification and inscription of "differences" between datasets into broader theoretical systems. It is in the natural history tradition that the idea of bringing the specimens dispersed in the world to a central location originated, a knowledge organizational schema that still guides the design of biology data repositories today,

including the design of the DataFace repositories. The impetus for the creation of biology databases was parallel to that for the founding of so many natural history collections. It was a reaction to a perceived "information overload," augmented by a new recognition of the scientific promise of the knowledge such a database would contain and the potential for individual and institutional prestige that would accompany its development. Databases, as Strasser observed, are so essential to modern experimental practices, but at the same time they belong to a "natural historian" way of knowing that relies on the collection and comparison of natural facts, often across many species – like in the case of the DataFace Consortium.

The twentieth-century saw the rise of many different collections of experimental data organized in databases. Well before the sequence databases, model organism communities were pioneers in the creation and maintenance of "community databases," such as the *FlyBase*, the *WormBase*, and the *Mouse Genome Informatics*. These resources are still highly used for data sharing and deposit today, and constitute necessary means for knowledge production (Leonelli, 2016). Community databases bring together several data types, organized by kind of model organism. As in the case of the drosophila community, most model organism communities are have a strong sense of affiliation, a moral economy of openness, and a way of thinking about research data as "collective properties" (Leonelli, 2016).

Traditionally, the organization of research data in community databases follows a quite specific workflow that originated in second half of the last century, but still remains

the same for most community databases today. It is this workflow that enabled the wide adoption of community databases as primary means for knowledge representation and production among model organism communities. I want to bring attention to two specific constitutive features of this workflow. First, community databases are made of collections of research data made available "post publication." Second, these collections go through a process of granular curation and integration that makes their reuse quit effective.

In model organism communities, the organization of scientific knowledge in the community databases enables the production and wide distribution of what Sabina Leonelli calls "small facts." The database staff harvests scientific information and sets of research data from the academic publications. In alternative, the researchers themselves submit to the database their data. Once the data are submitted to the platform, a team of "bio-curators" organizes and describes the datasets. The bio-curators "curate" the data so these will meet the specific needs of specialized communities. It is this process of data curation that allows biology small facts to "journey" from one research situation to the next. Bio-curators are responsible for guaranteeing the "effective packaging" of biology data, which is achieved by describing research data through "relevance and reliability labels." These labels are, respectively, bio-ontologies and metadata. When biologists download reuse others' data to ask new research questions, data go through a process of both de-contextualization and re-contextualization. Relevance labels make data attractive to users in new contexts (bio-ontologies) by associating datasets with their "object of research" (the biological entity under study) and the known datasets. Reliability labels

provide information about the "quality" of the data, such as information about the data format, the organism used in the experiment, the instrument and methods used, and the laboratory conditions under which the data were obtained (Leonelli, 2016). When successfully applied, these labels allow biology data to work as boundary objects: objects that are both plastic enough to adapt to local needs and constraints of the several parties employing them, yet robust enough to maintain a common identity across sites (Star & Griesemer, 1989).

Data sharing in the molecular lab: pre and post-publication practices

Centralizing data from distributed datasets into centralized databases constitute one way in which scientists make their data available (Wallis et al., 2013). We know that a lot of data sharing also happen in between labs "upon request." As pointed out by the Hilgartner (2017), before the rise of genomics the molecular biology lab was characterized by two quite distinct data sharing practices. On one hand, the labs working on similar issues would share within each other data and material underlying published papers. Typically, this type of reuse would come with no strings attached, other than a courtesy citation to the data creators' previous work. On the other hand, sometimes labs would want to reuse others' data when the related papers were not published yet. In this case, labs would share and reuse each other data in the context of a collaboration, which would result in co-authorship.

In her study of the epistemic culture of molecular biology, Karin Knorr-Cetina famously described the organization of knowledge production in laboratories in the 80s (Hilgartner, 2017, p. 65; Knorr-Cetina, 1999). In Knorr-Cetina's representation,

molecular biology is a highly individualized culture. Research takes place in small laboratories, typically consisting of a laboratory head and few postdoctoral researchers, graduate students, and technicians. People work mostly independently at their "lab benches," where "bench-work" requires careful manipulation of small tools and volatile material. This type of work practice is typically referred to as "wet lab" work. The "head" of the laboratory is entitled to direct the research program, select and manage personnel, allocate projects to subordinates, authorize the expenditure of resources, and make internally authoritative judgments about the epistemic quality of knowledge objects. In simple terms, the head of the laboratory "speaks" for it. The head of the lab assigns postdoctoral researchers to projects. Graduate students and the technicians assist postdocs in the development of the projects.

As discussed by Hilgartner, "the head of the laboratory would enjoy strong managerial privileges over transfers of knowledge and resources, and holds a legitimate monopoly on representing the laboratory and its accomplishments to the wider world" (Hilgartner, 2017, p. 65). For example, during the race to find a gene, the head of the lab would decide whether or not to publish "intermediate results" that could help rivals. The head of the lab also decides whether, how, and under what circumstances to share the lab's data with other labs. In this "typical" lab framework, the most common way of sharing data "inter-laboratories" was to request data and material underlying published papers. Sharing data of published papers is the least problematic way of sharing data, and scientists perceived it as a duty: it allows other qualified scientists to verify or build on

the authors' published work. Similar ways of sharing "published" data include presenting at meetings, conferences, and workshops.

Hilgartner showed that, in the molecular biology lab, while published data would be shared easily, biologists would share their "unpublished data" mainly in the context of formalized "labs collaborations." Hilgartner suggests that labs collaborations enable scientists to conduct cutting edge research without loosing control over "unpublished" resources. Hilgartner defines "collaboration" as a "genre of knowledge-control regime," which is based on an agreement among specific agents to participate in some joint projects or activities (Hilgartner, 2017, p. 82). Collaboration often forms when two or more laboratories possessed potentially complementary resources. Prospective collaborations typically identify a domain-specific project and relevant resources that the parties possess. Typically, collaborations would end in co-authorships, and collaborators would be middle-authors. The setting in which collaborations take shape – phone calls, side conversations at meetings, were "sporadic and elusive" and that "matters of compatibility and trust figured into interlocutors' thinking when assessing prospective collaborations." Pre-existing animosities, Hilgartner points out, sometimes impede the formation of collaborations. In this collaborative framework, scientific credit typically concentrated in the "first author" (usually the postdoc who ran the project) and "last author" (usually the laboratory head). "Middle authors" were often the graduate students, and the collaborators who helped by providing a "service," for example other labs supplying biomaterials or unpublished research data, or other labs performing specialized analyses (Hilgartner, 2017, p. 53).

The "genomics revolution:" the rise of unpublished open data

In his latest book *Reordering Life: Knowledge and Control in the Genomics Revolution*, Hilgartner (2017) argued that the design, constitution, and implementation of the Human Genome Project (HGP) – an international consortium for large-scale sequencing – introduced a novel regime of "knowledge-control" in the biomedical community (late 80s – early 2000). To revise the features of this regime is of particular interest for this dissertation project because, as I will discuss later, the DataFace Consortium shared some similarities with the HGP. In many ways, DataFace reproduced the vision for data sharing and openness that emerge during the HGP – which I refer to as "radical openness" regime for data sharing. However, in the context of DataFace, this regime was applied to a research community fairly new to genomics' methodology and the ethos of openness.

In the US, the HGP leadership included members of the National Center for Human Genome Research (NCHGR) within the NIH, and members of the Office of Biological and Environmental Research (OBER) within the DOE. Several scholars studied the history and the politics of the HGP (Keller, 2002; M'Charek, 2005; Reardon, 2017; Stevens, 2013), which is often described as the epitome of the "genomics revolution." For the scope of this dissertation, it is enough to highlight the main traits that characterized what Hilgartner calls the "genomics vanguard." The HGP had three main goals: a) to map and sequence the genome of the human and several model organisms; b) to make all the sequence data freely and technically available in public databases; c) to develop tools and methods for gene hunting and whole genome analysis (Hilgartner, 2017, p. 31). While

discussing to what extent the practices of genomics actually constituted a methodological revolution in biology is beyond the scope of this dissertation project, it seems reasonable to agree with Hilgartner on its observation the HGP ultimately promoted a "paradigm shift" in biology that aimed at making the field more computational. Through the HGP, the genomics vanguard sought to center data analysis on massive quantities of sequence data, from many organisms and individuals.

Before the HGP, single laboratories were mainly focused on locating and analyzing one by one genes related to specific disease or to human variation. This kind of work was extremely time consuming. The HGP aimed at providing a set of "reference maps" of the whole human (and certain animals) genome that would simplify and speed up the process of gene hunting. In order to build these reference maps, scientists would need to first sequence, locate, and map the location of each protein-coding piece of DNA on the whole genome. In the US – Hilgartner pointed out – a strong leadership was overseeing the HGP program. This "informal" but powerful leadership, which included funding officers and few selected science advisors, was in charge of coordinating the mapping and sequencing process in the US. The HGP leadership distributed the work of sequencing and mapping genes to over 20 laboratories. The selected laboratories were named "genome centers" and they received funding to accomplish the HGP's goals, especially sequencing and mapping genes. Hilgartner observed that the HGP leadership particularly emphasized accountability, stressing that genome centers should be tightly focused on achieving HGP goals (p. 97). Rapid data release was a persistent priority from the

beginning of the project. Centers would indeed often release datasets "prior to publication," as a way to show compliance with the agencies' requirements.

This transformation of molecular labs into genome centers resulted in few challenges. As discussed by several commentators, genome centers resembled factory-like data production facilities (García-Sancho, 2015; November, 2012; Stevens, 2013). In these "mapping facilities" young researchers would be assigned to high-throughput data production, which involved hundreds of robots and machines in a highly automated manufacturing process. Already during the HGP, Hilgartner (2017, p. 55) reported to have heard "complains from some of the postdocs about spending time on "routine" mapping work rather than on career-enhancing projects and grumbling about doing the work of "mere technicians." People were concerned over how could young scientists working on long-term, collective projects such as genome mapping, stand out as individuals who had made notable contributions. In other words, young researchers wanted to dedicate their time to the analysis of the sequence data – to find specific variants that code for disease, for example – not only to the design of reference maps.

Another related challenge was the expectation, and eventually requirement, of depositing sequence data right after data collection in public databases, prior to publication. As I will discuss in the next section, after long debates, the HGP leadership decided to require genome centers to release sequences as soon as these were collected. While some scientists raised concerns over loosing merit and "being scooped," this requirement was not seen as particularly controversial by the leadership and by senior

scientists promoting open access. As discussed by Hilgartner, behind the sharing requirement lied the assumption that sequence data do not constitute complete "research data." From the HGP leadership point of view, sequences of ATCG, often shared with minimal annotations on biological functions, constitute "hypothesis free" resources from which knowledge can be extracted – somewhat how fuel is extracted from oil. Hilgartner called these data *Unpublished in Journals, but Available in Databases* (UJAD) research data. Differently from curated research data that are deposited – after publication – in community databases (i.e., "small facts"), and also differently from research data shared in between labs "upon request," sequence UJAD are intended to work as "fungible" resources: research commodities whose individual units are essentially interchangeable (Leonelli, 2016).

The HGP: competing regimes for data sharing and open data

In his account of the knowledge-control regimes before, during, and after the HGP, Hilgartner (2017, p. 91) argues: "The rise of a self-consciously revolutionary research program such as the Human Genome Project created a context in which the settlements that govern scientific knowledge and data became susceptible to renegotiation." For Hilgartner, the HGP sponsored and enacted a novel mode of exerting control over genomics knowledge and data. Interestingly, the author argues that the US and Europe promoted quite different control regimes. The US program intentionally opted for a distributed system in which independent genome centers were hold accountable for the collection and the distribution of genomics maps. Productivity metrics, data reporting, and release requirements, unusual in molecular biology at the time, were used to hold genome centers accountable. In the US, the HGP leadership aimed to establish strong

control over the data producing centers. The data sharing process consisted in a one-way flow of data: from the centers to everybody. In this context, "technologies," such as the "Sequence-tagged Site" (STS) standard, were used to free users from needing connections to the centers, or to each other, to interpret, compare, and reuse maps.

As pointed out by Hilgartner, the UK proposed a quite different model of knowledge distribution. Like Dayhoff few years earlier (see next chapter), and like the editors of the drosophila newsletter, the European leadership imagined a central laboratory as a vital hub that integrated individual laboratories' datasets into a collective knowledge-producing network. The central laboratory would house everyone's data centrally while providing data to the users "on request." Where the US system sought to speed the one way-flow of data, the UK system sought to compel outside laboratories to contribute with their own data to the project, and, in exchange, it would provide access to all data. In this sense, the UK regime resembles the open/closed systems of model organism communities (Kelty, 2012). While the central lab would demand authorship in exchange for reuse, it would also retain some control over who would get access to the sequence data and when. We can see how, once more, the scientific community was divided between a regime of collective semi-openness (UK) and one of individualistic full-openness (USA).

These two different regimes of "knowledge-control" originated in two different ways of imagining the scientific community. The scholar argues that "it would not be much of an exaggeration to say that the US regime was premised on an imaginary of the scientific community as a population of autonomous laboratories, disconnected agents interested in

maximizing their individual freedom of action. As opposed to the UK model, in which the scientific community was imagined as a group of laboratories bound together by common goals, technologies, and materials and interested in maximizing heir collective achievements" (Hilgartner, 2017, p. 122).

Sequence databases: open vs. proprietary

The debates leading to sequence databases creation and governance – especially in relation to issues of attribution of credit and authorship, and the proprietary nature of knowledge – illuminate the different moral economies at work in the life sciences during the genomics revolution. These stories offer perspective on the recent wide spread of policies for full and immediate access to research data.

Similar to paper newsletters, the governance of sequence databases – since their very early conception – have been debated between competing visions of data credit, knowledge production, and data ownership. From the perspective of attribution of credit for collection and reuse, some sequence database creators stressed the idea that sequence databases added value to "raw data" via data curation and annotation techniques. In some cases, database managers would also retain the right of mining collections of others' data, and publish the resulting analyses. Others believed that database managers had no rights over the data collections, stressing the "molecular lab" ideology of individualistic knowledge production. These normally pushed for the total open access of the data submitted by the scientists to the databases. From the perspective of data ownership and proprietary issues, some sequence database managers promoted the idea of organizing collection in proprietary databases that could become a sort of revenue. This could be

done by partnering with private companies and by patenting certain sequence data. Promoters of full open access strongly opposed the possibilities of creating proprietary databases out of publicly funded data collections. In this section, I briefly summarize the debates around the constitution of two distinct sequence databases. The first one is the famous controversy behind the creation of GenBank. The second one is the less famous but quite illuminating case of the dbEST database in the context of the Human Genome Project, which was recently brought to attention to the scholarly community by Hilgartner.

GenBank, a public database of nucleic acid sequences officially funded by NIH starting from 1982, is arguably the largest and most frequently accessed collection of experimental knowledge in the world. The creation of GenBank, as Strasser (2011) pointed out, represented a significant historical turning point in the organization of biological knowledge. In the early 80s, while the proposal for HGP was being discussed, the scientific community started to pressure for the creation of a centralized sequence database. A comprehensive database of DNA sequences seemed indispensable for making sense of the abundant new data that was being produced. When, toward mid 80s, the NIH finally opened a call for funding opportunity for the constitution of a centralized database, two institutions were particularly well positioned to take the lead in developing such a facility in the United States. On one hand, the National Biomedical Research Foundation (NBRF) led by Margaret O. Dayhoff, and, on the other hand, the Los Alamos Scientific Laboratory, led by Walter B. Goad. As discussed by Strasser, the two very different proposals submitted by the institutions revealed competing ways of thinking

about the value of research data for knowledge production, attribution of credit and authorship, and ownership of the data.

On one hand, Dayhoff, an experienced data collector, proposed a system by which researchers who had determined sequences would share them voluntarily with her in a computer-readable format (e.g., tapes) for inclusion in her database. In exchange, her database would provide access to all the data, but at the condition of reserving proprietary rights over the data once the contract with NIH would terminate. Dayhoff's proposal put great emphasis on verifying the data for accuracy and on having the sequences "certified" by several experts, including the original authors. She argued that a carefully verified collection was "more economical in the long run than a 'quick and dirty' collection," a clear allusion to other sequence collectors who didn't put the same effort into verifying the data. Dayhoff would also reserve the right to mine the data deposited in the database, mirroring, according to Kohler, a specific natural historian tradition in which creators of collections own the items that compose the collection.

On the other hand, the theoretical physicist Goad proposed that journals would require scientists to submit their data to a centralized database as a pre-condition for publishing. His plan was going to be implemented with the help of the journal editors. He also specified that data would be made available to all researchers with no restrictions for reuse. Unlike Dayhoff, Goad had no experience in collecting sequences, but at Los Alamos he had access to an increasingly globalized networks of computers (the Internet) that he proposed to use to distribute and collect data from labs around the world. On 30

June 1982, the NIH contracted Los Alamos for the creation of what became GenBank, providing $3.2 million over five years to set up and maintain a nucleic acid sequence database. As Strasser observed, one main factor that influenced this decision was the different ways in which Dayhoff and Goad thought about credit attribution for data reuse and data ownership. The scholar argued that Dayhoff's standards of knowledge ownership were unacceptable to many experimentalists, who considered the data they produced to be their own and therefore to be published, distributed, and used only with their agreement. Dayhoff had a history of data management that crashed with the open access ethos of experimentalists. Indeed, few years before she made available a preliminary nucleic acid sequence database for free over the telephone network, while requiring researchers accessing the data were requested to sign an agreement not to redistribute the data. Lacking funding support, she also tried to set up a partnership with a private company to finance her data collections, which didn't work out, and eventually opted for selling access to the collections through a subscription. On the contrary, Goad's open access ethos matched the essential values of the experimental sciences' moral economy: namely, that the production of knowledge deserves individual, not collective, credit. As discussed by Strasser: "The creation of GenBank did more than just reflect the current moral economy of the experimental sciences and the culture of computer scientists: it served as a model and as a resource to promote open access to scientific knowledge." The debates leading to its creation–about the collection and distribution of data, the attribution of credit and authorship, and the proprietary nature of knowledge–illuminate the challenges of making a natural historical practice compatible with the moral economy of the experimental sciences in the late twentieth century.

The constitution of the GenBank database is only one example of how biomedical data collections have ben contented between the private and the public domain. I will give a second example taken from Hilgartner's latest book (Hilgartner, 2017). During the 1980s, the congress implemented a series of policy measures that expand the scope of patentable subject matter to include genetically engineered that have been isolated and cloned (Michael Fortun, 2001). Since the end of the 1970s, new policies in the US aimed at increasing national economic competitiveness by capturing commercial advantages from research. The 1980 Bayh-Dole Act famously enabled universities to file for patents on inventions that their scientists produced. Consequently, universities started to establish technology-transfer offices dedicated to secure rights to and licensing inventions (Lessig, 2001; Mirowski, 2011; Ramello, 2005). One of challenge of the HGP was to establish which parts of the human and animal genome to sequence. At the outset of the HGP, some members of the genomic community asked for sequencing only those parts of the DNA that codes for proteins, the so called "cDNA" (i.e., "cloned" DNA, clones made from DNA that codes from proteins). In their view, it was unnecessary to code the entire human genome. In the United States, the HGP leadership rejected the cDNA strategy: no one knew with certainly that noncoding regions of no apparent significance were really "junk DNA." This may still have vital but unknown biological functions. In the early 90s, Craig Venter found a way to "package" cDNA fragments in a way that can be used to identify genes, creating what is know as the "Expressed Sequence Tags" (ESTs). Venter and NIH sustained that these tags contained enough information to search GenBank for matching genes that had already been found, and made the statement that the EST could

find all the human genes for a fraction of the costs of the Human Genome Project. With full support of NIH, Venter filed for patenting the EST sequences. The patenting process took few years and raised strong opposition by the HGP leadership. NIH was eventually able to obtain a partial patent on the EST sequences. The patent stimulated the spread of "genomic companies." The Human Genome Sciences (HGS) created a proprietary EST database in collaboration with Smithkline Beecham. University researchers were granted access to the data in exchange for granting the company exclusive rights in any discovery.

The Human Genome Sciences's strategy of using its EST database to leverage query-based collaborations in exchange for patents rights met some resistance form academic scientists, who argued that this model could slow down the discovery process. In response to the HGS business model, the Human Genome Project leadership opposed using the HGS database in genome project work and partnered with Merk, a Smithkline competitor, to independently generate vast number of ESTs and establish an open-access EST database. Like GenBank, dbEST made its information available to any and all with no strings attached.

It seems to be that while similar to some extent, these two cases are also quite different. Both Dayoff and Venter thought that it was right for them to retain some IP rights over their "creative work" of curating and packaging data. However, Dayoff – at least in Strasser's account – is motivated by a natural historian ethos of ownership over

curated collections, while Venter – at least in Hilgartner's account – is driven by the explicit intention of monetizing "his discovery."

The birth of "UJAD" open data

Hilgartner's (2017) discussed that the genomics revolution eventually led to the diffusion of what he defined as "unpublished in journal, available in databases" (UJAD) research data. The scholar examined the process that gave birth of UJAD open data from the 1970s, when DNA sequence databases were first established, to 2003, when the HGP was officially completed. As I will discuss later, the DataFace Consortium decided from the very beginning of the project to make all the produced data and resources immediately and freely available to the scientific community at large. The story narrated by Higartner is particularly relevant because it allows me to trace back the origin behind the idea of making all data immediately available right after data collection to the very constitution of the HGP.

Before the HGP: the "staff-driven collecting" regime

Before the HGP, scientists would sequence DNA in connection to specific biological research projects that featured sequencing as a step in a broader analysis of biological functions or mechanisms. In this specialized research framework, DNA sequence served as data for supporting knowledge claims (García-Sancho, 2015). At this point, very few laboratories were "sequencing for the sake of sequencing," and the volume of sequence data in the published literature was relatively small. As we have seen, few research groups – especially those operating in model organism communities – grew interested in gathering together the available sequences, and collect them from scientific journals.

These researchers started to develop software and database tools for managing, searching, and analyzing these data. I have explained how GenBank became the first nationally charted, federally funded database to provide an archive of all published sequenced data. Hilgartner called the GenBank initial model for data collection and curation as "staff-driven collecting:" databases' employees searched the scientific literature for papers containing nucleic acid sequences, identified the sequences and other relevant information (i.e., annotation from the literature regarding the known biological functions of each given sequence), and entered them manually in the database. Once inserted in the database, sequence data would be open to any and all, made available on magnetic tape and online, with no restrictions on use or redistribution. Hilgartner noted that this collection was initially conducted independently from journals and authors. Database professionals, in order to ensure sequence quality, would type each sequence twice and spend a long time consulting the literature to annotate the data with the most updated information.

Dealing with the HGP data deluge: the "direct data submission" regime

The exponential growth in the number of published sequences that started to occur in the mid-80s destabilized this knowledge-control regime. At this point, scientists heavily rely on GenBank's sequences for their research projects, and, as reported by Hilgartner, "they wanted access to their colleagues' data immediately after publications, not months late" (Hilgartner, 2017, p. 158). By the second half of the 80s, "a sense of crisis dominated workshops and advisory committee meeting were the future of GenBank was discussed" (Hilgartner, 2017, p. 159). In order to keep up with the data publishing rates, GenBank staff started to add new sequences without internal annotation, scarifying the

quality of data and the completeness of coverage of the literature. Toward the end of the 80s, GenBank started to develop a system through which scientists would submit sequences directly to the database and annotate their own data whenever they would submit a paper to a journal. Initially, "direct submission" was completely voluntary, and it resulted in low participation. However, the rate of data submission increased rapidly when databases started to equip the scientists with tools that would enable them to easily prepare data submissions and annotations. For example, Genbank introduced a "friendly if not seductive annotation software" called AUTHORIN, which was designed to easy submission and automatically check for errors (Hilgartner, 2017, p. 161). A big boost in data submissions arrived when journals started to make sequence data submission mandatory for publication. The first journal to do so was the Nucleic Acid Research (NAR) journal, in 1988. As reported by Hilgartner, by 1992, at least 36 natural science journals required sequences to be submitted to the databases as a pre-condition of publication. This mandatory direct submission of data would soon be called "electronic data publishing." Significantly, the phenomenon of electronic data publishing created for the first time the distinction between "sequence data" (all data sequenced by a lab) and "results data," coupled with the conviction that results goes into the journal, and sequence data in the database. Hilgartner observes that: "Since the very beginning, electronic data publishing did not have the same academic credit as a paper. […] Indeed GenBank defined its role as maintaining a bibliographic record of how had submitted what, not evaluating the significance of the contribution" (Hilgartner, 2017, p. 163).

A rationale for immediate release: sequencing data as "a service"

A main issue that emerged with mandatory data publishing was what Hilgartner refers to as "the when question:" the right timing for making the data accessible for reuse. Initially, Genbank offered to hold the data submitted in confidence until a related paper appeared in print. However, some communities vividly argued for making the sequence data open right after data collection. Significantly, those who argued that sequence data should be made available right after collection sustain that sequence data was not, indeed, experimental data, it was more like a "service" type of activity, rather than "real research." Defenders of the idea that sequence producers retained no residual rights suggested that sequencing centers were being funded to produce and submit information, not to analyze it. Particularly vocal were the scientists running the sequencing of *C. elegans* worms, whose moral economy resembled those of the *Drosophila* community (De Chadarevian, 2004; Kelty, 2012; Kohler, 1994). However, not everybody agreed on making all data available right after data collection. While laboratory heads wanted to submit the data quickly to show productivity, among younger PhDs in HGP laboratories questions about personal credit were acutely felt matters, they needed to make distinctive personal contributions in order to produce a viable scientific identity. Indeed, in the lab, young researchers wanted to have time to do a career-enhancing analysis of the data. For them, it was not just service work.

The Six-Month Rule and the "overview papers"

Because scientists perceived that some colleagues were strategically releasing data with delays, in the early 1990s the HGP community called for a standard rule aimed at preventing genome project laboratories from engaging in such practices. In the United States, the initial idea that gained traction was a rule specifying that data and materials

should be made publicly available within six months of being generated. According to Hilgartner's analysis, most of the sequencing groups would submit their sequence information to the databases in order to demonstrate productivity. The quantity of data submitted was for them the most visible indicator of a laboratory output. At the same time, these six months were used by the sequencing projects to publish broad "overview" papers in which they would describe what had been sequenced, outlined the method used, and provided a preliminary analysis of biological features such as the number and density of density of genes. Publishing a primary overview paper became what genome scientists still view as the traditional way for large-scale sequencers to get credit for their work.

The "Bermuda Principles" and the rise of "unpublished open data"

The six-month strategy worked for few years. However, as Hilgartner reported, once the HGP laboratories started to transition more and more from mapping to sequencing in the mid 1990s, the HGP leadership had to come up with a new rule to coordinate the request of data access and the competition among the sequencing centers. In February 1996, the HGP leadership met in Bermuda to develop a new plan for data release. The new Bermuda Principles imposed extremely stringent data-release requirements on large-scale human sequencing centers (F. S. Collins, Morgan, & Patrinos, 2003). Newly generated "preliminary" sequences would be released as soon as possible, on a daily basis. "Finished sequences" had to be submitted to replace the preliminary data, right after the cleaning process was over. Beyond specific scientific progress, the Bermuda principle served as a means to demonstrate that the large sequencing centers were behaving responsibly, as an accountability mechanism. Hilgartner called this data sharing regime the "rapid publication" regime, which brought the practice of depositing and

reusing “unpublished in journals and available in databases” (UJAD) data into biology research.

The diffusion of UJAD challenged the data producers in few ways. As we have seen, data producers had the expectation to publish overview papers by analyzing their own data. This expectation was embedded in the ongoing practices, but not codified in any rule. Hilgartner reported that, in several instances, “sequencers were shocked to find out that other research groups had downloaded the UJAD sequence information from a nearly finished project, analyzed it, and published overview papers” (Hilgartner, 2017, p. 176; Rowen, Wong, Lane, & Hood, 2000). People would engage in this practice are sometimes referred to as “data parasites” (Friedberg, 2016; Longo & Drazen, 2016). Secondary data analyses of *others* data is indeed a very controversial practice that sees scientists divided between sustainers and firm opponents of such practice. Sequencers who participated to the Human Genome Project would argue that yes, they are providing a service, but they are also running the service “for a reason,” namely that they were scientists and that as such were naturally interested n analyzing the data that they produced (Hilgartner, 2017, p. 177). The HGP leadership eventually decided to update their data –release policy, in a very significant way:

> “NHGRI believes that a reasonable approach is to recognize the opportunity and responsibility for sequence producers to published the sequence assembly and large-scale analyses, while not restricting the opportunities of other scientists to use the data freely as the basis for publication of all other analyses” (NHGRI, 2000)

It is very important to highlight that in this update the HGP stated that scientists should get access to the sequencers’ data, but at the condition that they would use the

data to run a "all other" analyses. This meant that scientists are asked not to use their colleagues' data in ways that could directly hurt their colleagues' careers.

The Fort Lauderdale Agreement

In January 2003, a group of 40 people including representatives from GenBank, from the sequencing centers, journal editors and computational biologists interested in the reuse of large-scale sequence datasets, met in Fort Laurerdale, Florida to discuss ways of extending practice of immediate release of sequence data beyond the HGP. The meeting resulted in the 1800 document called "The Fort Lauderdale Agreement." The Agreement draw a set of informal rules that aim at finding a balance between providing access to sequence data, and, at the same time, ensuring academic credit for the data producers, and the quality of the submitted data. As summarized by Hilgartner, the agreement stated that:

- *Funding agencies* should require, as a condition of funding, free and unrestricted data release from community resource projects. But, they should also "support the ability of the productions centers to analyze and publish their own data;"
- *Resource producers* should make data "immediately and freely available without restriction." They should also "recognize that even "recognize that even if the resource is occasionally used in ways that violate normal standards of scientific etiquette, this is a necessary risk set against the considerable benefits of immediate data release;"
- *Resource users* should cite and acknowledge the resource producers' work, and "recognize that the resource producers have a legitimate interest in publishing prominent peer-reviewed reports describing and analyzing the resource that they have produced" (Trust, 2003)

The stipulation of the Fort Lauderdale Agreement established three important precedents. First, it established that the practice of releasing data prepublication should be restricted to "community resource projects," defined as: "large-scale efforts devised and implemented to create a set of data, reagents or other material whose primary utility will be as a resource for the broad scientific community." As Hilgartner observed, the concept of community resource project built of the governing frame of the HGP, in which "genome projects" (mapping and sequencing) had been conceptualized as significantly different from "ordinary biology" projects (gene discovery). Second, the agreement would not rely on in a set of codified "right," but it would be grounded in "mutual respect and self-restrain." The agreement envisioned a "scientific etiquette" in which scientists would trust each other in reusing data without scooping each other. Third, as Hilgartner significantly notes: "The principle that large-scale infrastructure projects should ensure rapid and widespread availability of data became a recognized norm of what some observers referred to as the post-genomic era" (Contreras, 2011; Hilgartner, 2017, p. 181).

Open data, human subjects, and participatory paradigms

In the first section of the literature review, I have examined how emerging data sharing regimes symbolize competing visions of what constitute credit for data reuse, data authorship, and data ownership in biology research. So far, my analysis focused on providing an account of how science stakeholders (researchers, funding agencies, journal editors, database managers etc.) shaped, reacted to, and act across these changing data sharing regimes. I now turn my attention toward the relationship between data sharing regimes, open data policies, and public benefit. Openness of research data is often

promoted as a means of epistemological transparency: if data are freely accessible, anyone can potentially verify the accuracy of scientific results (Leonelli, 2016). In this sense, open data practices promote trust from the public toward science. Open data policies are also a matter of accountability and return on investment (Borgman, 2015). Like in the case of the HGP, research data are often collected with public funding, so it is in the interest of the public to make sure that usage of data is maximized and tax payers money are not wasted. Finally, openness of data promotes public participation in science, such as in citizen science projects and Do It Yourself (DIY) biology initiatives (Darch, 2014).

Recently, scholars in the social studies of science and related fields started to raise concerns in relation to the openness of biomedical research data, especially human subject data. Beyond obvious concerns about patients' re-identification and privacy, scholars call for an examination and evaluation of the historical circumstances that led to the collection of highly reused human subject research data (Radin, 2017b). Joanna Radin recently examined the history behind the collection of the "Pima dataset." "Pima" refers to the members of an Indigenous community who live in the southwestern region of the US. The dataset, which contains sensitive biomedical information about the members of the community, was collected during the 60s "in ways that do not respect the human dignity of human subjects and that do not recognize the legitimate interest of the Community in the integrity and preservation of its culture" (Radin, 2017, p. 58). Today, computer scientists reuse the Pima dataset – freely available online in a data repository managed by the University of California Irvine (UCI) – to train machine-learning

algorithms for all sorts of purposes. In her historical analysis, Radin wonders whether those who reuse the data should somehow take this controversial history into consideration. Most importantly, Radin suggests that this history should be made visible to the data reusers via data citations and provenance techniques.

Stevens (2016) pointed out that the accumulation of sequence data and individual health data has generated a growing stock of information that can be mined to look for the relationship between genes and diseases. One challenge to the analysis of such stock of information is that it is highly distributed. Sequence and health data about individuals exist in a multitude of databases located internationally. Most of these data collections are publicly funded, such as all the data deposited on the National Center for Biotechnology Information (NCBI). In this context, meta-analyses conducted with semi-automated analytical tools are regarded as particularly promising. Researchers are increasingly employing machine-learning algorithms – such as deep neural networks – to partially automate the analysis of this huge amount of distributed data (Mamoshina, Vieira, Putin, & Zhavoronkov, 2016). Genomics, health, and phenotypic data about individuals are used as "training datasets" to teach algorithms what to look for. Once the algorithm is trained, it can be reused to make predictions about novel data.

By the mid-2000s, some entrepreneurs saw a commercial opportunity for monetizing the predictions coming out of machine-learning analytical methods. The first so-called "personal genomics" services started to spread in early 2000. The company 23andMe was founded in 2006 and offered its first testing services to the public in November 2007.

Navigenics and DeCODE likewise started to sell their services around the same time. These services sell direct-to-consumer genome sequencing, and, by analyzing consumers' data, make predictions about their potential health risks. They provide information about an individual's risk for a wide variety of diseases and traits. Some of them also provide consumers with reports about their ancestry background, as we have already discussed. These companies argued that genomic information should not remain in the exclusive domain of scientists or biotech, but could rather be used to empower consumers. Like pharmaceutical, personal genomics services promise an increasing personalization of health care and an increasing ability to use biotechnology to tailor medical interventions to individuals' bodies (Nelson, 2016; Stevens, 2016).

Alternative, "participatory" versions of personal genomics have also emerged. For example, the "Open Humans" platform encourages individuals to upload and donate their personal health and genomics data to diverse research projects (Open Humans, 2016). At Harvard, the "Personal Genome Project" (PGP) – led by George Church – hopes to enroll 100.000 volunteers, their while-genome sequence, and their medical records (Buhr, 2018). In both cases, the hope is to increase knowledge of human disease, but also to genomically tailored and targeted treatments and pharmaceuticals.

The diffusion of all these different types of individuals' health and genomics information raises difficult questions about who has the right to own, access, interpret, and distribute genomic and health data. Genomic information is indeed identifiable – it can be directly associated with a particular individual. Advocates groups are worried that

individuals' genomics information can be sold to or accessed by insurance companies, and used to discriminate against individuals with pre-conditions. In the United States, the creation of the Genetic Information Nondiscriminary Act (GINA) – signed by President George W. Bush in 2008 – partially addressed this concern (Hudson, Holohan, & Collins, 2008). GINA makes it illegal for employers and insurers to discriminate on the basis of genetic of genomic information. However, the law does not apply to life, disability, or long-term care insurance.

Scholars also raised ethical concerns over the ownership of the research datasets obtained from the patients. Probably the most famous case of exploitation of a patient's data is the case of Henrietta Lacks (Skloot & Turpin, 2010). Ms. Lacks was a 31-year-old working-class African-American woman from Baltimore. In 1951 she was admitted at the John Hopkins Medical School where she was diagnosed with cervical cancer. Without the knowledge or permission of the patient and her family, some cells from Lacks' tumor were sent to the laboratory of Dr. George Gey, who used them to create the first ever "line" of human cells. Dr. Gey distributed the cells for free, never attempting to patent them or limit their distribution. However, the line was later reused to test the polio vaccine, and consequently marketed for sale by a biological supply company. This raised issue of data ownership: given that companies profited by using Lack's cells, why shouldn't Lack's descendent be able to sue for a share of the profit? Also, shouldn't doctors have asked Henrietta for permission to use her cells? Shouldn't her identity being kept private? More recently, in the 90s, John Moore, a leukemia patient, lost a historic property rights battle in which he claimed he deserved to share in the profits from an anti-

cancer drug derived from cells taken from his spleen (McLellan, 2001). Differently from Moore, T*ed Slavin* was a *hemophiliac* who successfully sold his antibodies and aided Dr. Baruch Blumberg in the discovery of the link between the hepatitis B virus and liver cancer. This eventually led to the first hepatitis B vaccine (Skloot, 2006). Some are of the opinion that all people should be able to sell their data to research. George Church at Harvard, for example, is encouraging individuals to sell their DNA to his PGP project (Buhr, 2018).

The use of informed consent helps to control the reuse of research data obtained by patients. However, scholars pointed out that informed consent has its limits as well. As I anticipated in the introduction to this section, Joanna Radin recently wrote about a well-known case in the realm of bioethics and medical history: the Pima Native American tribe in Arizona, which is known for unusually high rates of diabetes and obesity (Radin, 2017a). The Pima were the first Native American tribe to be granted a reservation in Arizona at the beginning of the California Gold Rush. In 1963, following nearly half a century of mass famine among the Pima, NIH conducted a survey for rheumatoid arthritis in the Pima tribe, instead discovering a frighteningly high frequency of diabetes. In 1965, the NIH initiated a long-term observational study of the Pima that continued for about 40 years, though it was meant to last no more than 10. The goal of the study was to learn about diabetes in the "natural laboratory" of sorts that the Pima reservation unwittingly provided. The data collected in this study came to be known as the Pima Indian Diabetes Data set (PIDD).

Machine learning enters the story around 1987, when David Aha and colleagues at the University of California, Irvine (UCI) created the UCI Machine Learning Repository, an archive containing thousands of data sets, databases and data generators. The repository is still active today, virtually a gold mine for researchers in machine learning to test their algorithms. The PIDD is one of the oldest data sets on file in the UCI archive, "a standard for testing data mining algorithms for accuracy in predicting diabetes," according to Radin. Generations' worth of data on the Pima tribe have been publicly accessible in the UCI archive for over two decades, creating ethical controversy around the accessibility of information as personal as blood pressure, body mass index (BMI) and number of pregnancies of Pima Native Americans. Though the PIDD can help refine machine-learning algorithms that could accurately predict–and prevent–diabetes, the privacy issues provoked by the "publicness" of the data are impossible to ignore. This is where "eternal" medical consent enters the equation: no researcher can realistically inform a study participant of what their medical data will be used for 40 years in the future.

Current controversies in human genomics research

To understand fully what is at stake when biomedical data collected from human subjects are made freely available for reuse, I review critical literature in the fields of social studies of science and anthropology of science that surfaces the ethical controversies embedded in certain genetics practices and epistemologies. In what follows, I report on the societal issues raised by several scholars in relation to the practices of precision medicine and population genetics. As I will discuss, craniofacial research is deeply informed by the practices of precision medicine and population genetics. I also

briefly summarize the long history that ties craniofacial research to the practice of classifying human beings into supposedly discrete ethnic groups.

Precision medicine, population genetics, and racial categories

The novel field of "personalized medicine" – also referred to as "precision medicine" – aims at providing diagnosis and treatments that are targeted to individuals' genetic profiles. Patients' DNA is collected and searched for indications (genes, genetic markers, SNPs etc.) of how an individual might be predisposed to a certain syndrome, or how the patient might react to a targeted drug (Frizzo-Barker, Chow-White, Charters, & Ha, 2016). Precision medicine is expanding rapidly, capturing both the imaginary and the funding priorities of the funding agencies (Ferryman & Pitcan, 2018). Like virtually all funding agencies in the biomedical domain, also the agency that funded the DataFace Consortium included "precision medicine" in its 2014-2019 strategic plan as a main outcome of craniofacial research. In 2015, Springer released a collection of papers from the craniofacial research field titled "Genomics, Personalized Medicine and Oral Disease" (Sonis, 2015).

Scholars have raised a set of concerns in relation to the practices of precision medicine (Ferryman & Pitcan, 2018). Commentators in the field of social studies of science noted that this novel approach, instead of tailoring drugs on individuals, it actually relies on grouping individuals in "risk populations" that mirror old racial classifications (F. S. Collins, 2011; Rabeharisoa et al., 2014). To be sure, biomedical research has held a controversial relationship with the concept of "race" for a long time. The supposed usefulness of racial categories to study human health is constantly debated

within and outside the academy (Chen, 2016; Hacking, 2005). For a short period of time, it seemed that biomedicine was ready to dismiss race as a factor to study human health. In 1972, Richard Lewontin found that genetic human variation *within* ethnic groups is higher (93%) than genetic variation *between* groups (7%) (Hacking, 2005; Lewontin, 1972; Stevens, 2016). However, after the completion of the HGP, researchers felt the need to go back to sampling people based on their ancestry background. As noted by Stevens (2016, p. 292), in many ways, "the HGP was a triumph of technological and collaborative scientific effort." But in other ways it was a disappointment: it turned out that it is harder than expected to find significant correlations between genes and human phenotypes – including genetic syndromes. Many genes might be involved in determining simple traits, genes seemed to work together in complex networks, epigenetics factors seemed to play a signification role, and environmental signals also proved to have a large impact on gene expression. This set of novel factors are now investigated in a series of disciplines commonly referred to as "post-genomic" biology (Richardson & Stevens, 2015). As a way to less the complexities of genome analysis, and at the same time increase the chances to find statistically significant correlations between genes and complexities, researchers turned their attention to the examination of how genetically distinct groups (i.e., populations) might react differently to certain drugs, or are differently predisposed to certain syndromes. Troy Duster coined the expression "the molecular re-inscription of race" to refer to this switch back to racial classification in biomedical research after the HGP (Duster, 2006).

While old-school racial categories grouped individuals based on their "externally visible characteristics" (EVCs), modern population genetics uses the frequency of certain genetic markers – Ancestry Informative Markers (AIMs) – across different cohorts of patients (Nelson, 2016). The idea of genetically distinct "racial" populations is rooted in the assumption that individuals leaving in proximity to each other tend to share similar DNA profiles. Because of this proximity (i.e., continental affiliation), these individuals have been exposed to similar evolutionary processes, including sexual selection and gene flow, which made them genetically close to each other. With this assumption in mind, individuals' DNA is examined for a panel of AIMs, which are then correlated with "geographically separate populations." As a result, individuals are grouped in distinctive and supposedly mutually exclusive "types." There are four main genomic ancestry types: African, European, Native American, and East Asian; a division which, some have argued, recapitulates the centuries-old racial categories of Caucasian, Mongoloid, and Negroid (Caspari, 2003; Dewey-Hagborg, 2017). Commentators pointed out that population genetics is just another arbitrary way of classifying human beings into racial categories that are not actually mutually exclusive (Chen, 2016; Duster, 2006).

Also commercial companies like 23andMe uses AIMs for their direct-to-consumer genetic ancestry testing, though commercial companies also examine mitochondrial (mtDNA) and Y chromosomes (Y-DNA) (Donovan, Pasquetto, & Pierre, 2018; Panofsky & Bliss, 2017). Because these types of DNA are inherited directly from mother to offspring (mtDNA) or father to male offspring (Y-DNA) without recombination, they can be used to trace maternal or paternal lineages along one ancestral branch. However, while

mtDNA and Y chromosome DNA can provide an interesting path through ancestral history, they do not provide an overall view of an individual's constitutive "racial" percentages (M'Charek, 2005; Nelson, 2016).

In precision medicine, AIMs and population genetics have been employed to study the effect of certain drugs on specific populations, or to predict birth defect – like in the case of craniofacial studies (Parker et al., 2010). In some cases, the use of population genetics raised serious concerns. These cases include, for example, the design and commercialization of the BiDil drug, which was supposed to help African-Americans at risk of heart attack, but whose population-specific benefits were never proven (Duster, 2005; Temple & Stockbridge, 2007). Another infamous case is the discovery of the MCPH1 gene of "intelligence," which some researchers controversially claimed to be more common in certain populations over others (Evans et al., 2005; Stevens, 2016, p. 321). Beyond precision medicine, biomedical researchers employ population categories in a variety of research contexts. Evolutionary geneticists, for example, study the factors that cause changes in allele frequencies in within populations over time, and these changes are understood as being at the heart of how and why evolution happens.

Craniofacial research, facial measurements, and identification systems

Populations can be used in isolation (e.g., a dataset of only "Caucasian" DNA), like in the case of the Genome Wide Association Studies (GWAS), or in "admixture" research studies, which use samples from "mixed" populations (e.g., a dataset with mixed Caucasian and African samples). The craniofacial researchers I interviewed for this study collected two distinct GWAS datasets, one from a Caucasian population and one from an

African population. I will get into the details of how GWAS studies are conducted by DataFace researchers in the findings' chapter. For now it will be enough to point out that GWAS aim at finding statistically significant associations between certain genetic markers (e.g., AIMs) and some pre-selected phenotypes of interest. In craniofacial research, scientists test thousand of genetic markers that could be potentially found in genes associated with the formation of the facial shape. A set of pre-determined facial traits, obtained from 3D images of human faces, are used as phenotypes. By examining associations between genetic markers and facial traits, DataFace researchers aim at finding the genes that cause the formation of facial traits during human development.

The researchers involved in the production of DataFace GWAS datasets envisioned as the primary use of these data research activities in the context of clinical research, and in particular for the investigation of the genetic causes involved in craniofacial syndromes. As anticipated, however, DataFace data have been reused in the context of facial reconstruction research, to develop computational models to predict human faces from DNA samples. A team of physical anthropologists and computer engineers led the research on these models. Historically, craniofacial research is a highly interdisciplinary field, which has branched out in many different directions, from biomedical research, to forensic anthropometrics studies. In order to quantify and map the human face, modern craniofacial researchers use metrics (e.g., facial landmarks and linear distances) and tools (e.g., calipers and tapes) that were originally designed in the nineteenth-century for measuring the level of development of different people, especially across races (Stevens, 2016, p. 315). For instance, measurements of the skull, including its volume, were taken

to indicate cognitive characteristics and adaptive patterns (Boas, 1922). In his latest book *Biotechnology and Society*, Hallam Stevens (Stevens, 2016, p. 315) explains how the idea of using "facial angles" was used in this period to characterize human types: "The facial angle was the angle between two imaginary lines drawn on a profile of a human face: the first line went from the middle of the nostril to the earhole, and the second from upper jawbone to the forehead. Europeans were supposed to measure in at around 80°, African at around 70°, and orangutans at 58°." This classification was used to show the progression from lower to higher types of human beings. The early nineteenth century also saw the practice of "phrenology" to become widely popular. Phrenology believed that the brain was made up of large number of different organs that controlled different behaviors. Phrenologists also thought that the function of these organs was directly related to their size, and that the size of the organs varied across races (Farkas, 1994; Teslow, 2014).

Facial measurements developed in the context of biomedicine and physical anthropology have been widely used to develop systems of identification. For example, around 1880, Frenchman Alphonse Bertillon introduced a system of identification – the "Bertillonage" system – that was based on several measures of physical features, including facial and head measurements – such as head length (crown to forehead), head width (temple to temple), width of cheeks, and "lengths" of the right ear (Ragas, 2018). Today, Apple's latest iPhone X security recognition system, called "Face ID," uses similar facial metrics to confirm the phone's owner identity. The iPhone X's Face ID uses its cameras to make 3-D scans of the users' faces, which then enable them to unlock

their phones by just holding the device in front of their faces. Modern facial recognition and analysis techniques are empowered by high-tech cameras and advanced machine learning algorithms, which hold out the promise of scientific objectivity. Tools for facial identification are also used by law enforcement agencies to search for suspects in criminal investigations. The Federal Bureau of Investigation (FBI) developed the "Next Generation Identification" system, a database it is described as "the world's largest and most efficient electronic repository of biometric and criminal history information" ("Next Generation Identification (NGI) | Biometrics," 2016). It includes an automated facial recognition search and response system for law enforcement agencies. Similar algorithms and metrics are used for the design of DNA-based facial reconstruction technologies, such as those developed using DataFace datasets. In the findings section, I will show that novel computerized systems for facial mapping and classification are highly contested among craniofacial researchers.

4. Research design

This study is part of a larger research project funded by the Alfred Sloan Foundation with a grant titled *If Data Sharing is the Answer, What is the Question?* The overarching goal of this grant is to unpack the promises and challenges of making available – in different forms – research data. Given the current public investments in open data policies and infrastructures, the Center for Knowledge Infrastructures (CKI) examines the underlying motivations that brought to their design and implementation. The CKI researchers investigate questions such as: What kinds of problems are research data sharing policies and practices addressing? What kinds of solutions are these policies and practices providing?

Informed by this larger research design, this dissertation focuses specifically on the challenges and implications of *reusing* research data, once these have been already released in open access. In particular, I look at the policies and practices of data sharing and reuse in the biomedical field. My main case study is the DataFace Consortium (DF), a consortium for data sharing funded by one of the agencies of the National Institutes of Health (NIH) in 2010. Participants in this study collect and make available for reuse large-scale biomedical datasets. The three overarching questions that guide this project are:

1. What motivates the design of policies and infrastructures for open research data?
2. How do researchers reuse open research data for knowledge production?

3. What are the societal implications of making available and reusing open biomedical data across contexts of production?

Introduction to case study

I use the fictional name "DataFace" to protect the confidentiality of the research participants. The consortium is currently in its second five-year grant phase, which started in 2015 and will end in 2020 (DataFace II Consortium). Data collection for this dissertation project started in January 2015 and ended in September 2017, which means that my investigation focused on phase-two of the DataFace funding cycle (DF II).

The DataFace Consortium was funded to meet two overarching goals, naming *collecting* and *making publicly available* novel craniofacial biomedical data. DataFace data are deposited and available for download and reuse by the larger research community on a digital open repository accessible online.

DataFace's initial mission (phase I) centered on the collection of biomedical research data obtained from human patients and mouse animal models. Partecipating scientists mainly collected research data to the goal of informing biomedical research on the development of craniofacial syndromes, in particular oral clefting. From the very beginning, the consortium aimed at promoting and proving funding for the collection of large-scale genomic data, such as data collected in the context of Genome Wide Association Studies (GWAS) and via whole-genome sequencing techniques.

DataFace phase II extended the mission of DataFace I by complementing the collection of research data from humans and mice with new data collections from chimpanzee and zebrafish. In addition, DataFace II expanded the consortium research focus shifting the attention from oral clefting to a variety of other craniofacial syndromes, and novel developmental and evolutionary questions. As I discuss in details in the findings section of this dissertation, DataFace II also invested in improving the organization, integration, and granular curation of data collected by DataFace I and DataFace II investigators, to the goal of promoting wider reuse of these data.

As per fall 2017 (end of my data collection), DataFace II counts 11 principal investigators (PIs), and around 70 researchers total. During phase-two, new laboratories and researchers became involved in the DataFace project. The following description of the DataFace project mirrors the state of the consortium during the years 2015 and late 2016. It is beyond the scope of this project to discuss how the consortium evolved over time, even though it is an interesting aspect of the project, and it might become the topic of future research work.

The 11 satellite teams

DF is organized in eleven "satellite teams" selected by the NIH. These include on engineering hub, and ten satellite spokes.

The hub

One team is called "the hub" or "the coordination center," whose goal is to "develop and

maintain a DataFace II Data Management and Integration Hub infrastructure that will properly store, represent, and serve these data to the research community, and in addition provide access to tools for visualizing, integrating, annotating, linking and analyzing the data." The hub's members are the database engineers responsible for the development of the DataFace open repository, plus few senior scientists who help the engineers to meet the scientific and domain-specific goal of the community.

The spokes

The remaining 10 satellite projects are called "the spokes." Members of the ten spokes are geographically distributed among nine academic laboratories, one national lab, and three international labs located in Europe, North America, and in the Middle East. Collectively, investigators span molecular and developmental biology, computational biology, genomics, human genetics, bioinformatics, medicine, dentistry. The ten spokes are formally grouped in two sub-categories based on their role in the DF consortium: scientific and technology spokes.

The eight scientific spokes are responsible for the collection of novel imaging and genomic data. Scientific spokes use different methods and tools for the collection and analysis of the data, such as DNA and RNA recombination, transgenic experiments, gene editing experimental techniques, genome-wide association studies, next generation sequencing for DNA, RNA and protein sequencing; and machine learning algorithms for the analysis of the data. Scientists submit the data in highly heterogeneous formats, including 3D images, gene expression data, ChIP-Seq, RNA-Seq, animal and human

tissues, etc.

The two technology spokes have distinct roles in the consortium. One is referred to as the "ontologies spoke," and the other one as the "software spoke." The ontology spoke is in charge of extending the Ontology of Craniofacial Development and Malformation (OCDM) to accommodate conditions of interest to DataFace II researchers, such as human and mouse facial, palatal, and cranial vault development, and dysmorphology such as craniosynostosis, midface hypoplasia, frontonasal dysplasia, craniofacial microsomia and microtia. The software development spoke aims at developing a software interface to enable DataFace website users to apply human genetics analysis software (e.g., PLINK) to human genomics data from craniofacial research, with access to this tool through the hub. Both the ontology and the software spokes collaborate with the hub to reach and coordinate their goals.

Sampling strategy

In this section, I define my units of analysis and I describe how I selected my sample population. Babbie defines units of analysis "the what or whom being studied" (Babbie, 2012, p. 97). In social science research, the most typical units of analysis are the individual people. The overarching goals of this dissertation are twofold. I examine how the participating scientists reuse others' data in their individual practices, but also as "teams" distributed in different labs reuse data that were not collected others' labs. For this reason, my units of analysis are both the individuals and the teams.

The individuals and teams were selected applying the techniques of purposive sampling. Babbie defines purposive or judgmental sampling as "a type of nonprobability sampling in which the units to be observed are selected on the basis of the researcher's judgments about which ones will be the most useful or representative" (Babbie, 2012, p. 190). I selected my sample to include the most diverse and representative population. The final sample consists of:

- 1 Engineering hub
- 2 Technology spokes
- 6 Scientific spokes

DataFace scientific practices vary by many factors. I selected the sample population for the scientific spokes (6) in a way to allow for variation across at least three recurrent variables: sub-disciplines, data types, and model organism specialty. Individuals are specialized in diverse disciplines, such as molecular biology, developmental biology, epigenetics, and quantitative/computational biology. Each scientific team submits different data, including images (microCT, TIFF), gene expression data and drawings, data analysis results (gene functions), RNA-seq, ChIP-seq. Data are collected from four animal models, namely zebrafish, mouse, chimps, and humans. In the findings section I detail the data creation practices of three spokes. I also selected participants for interviews to ensure diversity of career stages. For each of the 11 teams, I interviewed the leading scientist, a lab manager, and one or two doctoral students or post doctoral students. Most of the DataFace teams include no more than five individuals.

At each spoke, I conducted interviews, participated as an observant in team meetings, and spent recreational time with the participants. Overall, I collected a total of 50 interviews over an observation period of three years and 6 months, with approximately 150 days of presence in the sites. Over the last three years and six months, I also participated in four DataFace Annual Meetings, and I presented my preliminary observations to the engineering hub in February 2016. I have also collected and analyzed a number of documents relevant to this community, which include academic papers, white papers, research notes, research presentations, posters, grant documents, email exchanges, and GitHub conversations.

Research methods

In this dissertation research, I used ethnographic fieldwork as a method for data collection and analysis, which includes ethnographic observations, semi-structured interviews, and document analysis.

Ethnographic fieldwork

Ethnographic fieldwork is one of the most common methods of knowledge discovery among qualitative social researchers. Ethnographic research is a theory-generating research activity that is particularly appropriate when the researcher is interested in uncovering in-depth knowledge in relation to one or a limited set of case studies (Babbie, 2012, p. 324; Lofland, Snow, Anderson, & Lofland, 2005, p. 16). This methodology is normally carried out via direct observations of phenomena, in-depth interviews and document analysis (or archival research). The community studied is typically spatially-located and the observations occur over extended periods of time (Babbie, 2012, pp. 296–

298). By observing the everyday practices and beliefs of a community of actors in their "natural habitat," the ethnographer looks for patterns in the ways actors employ social interaction to understand reality and act in space (Tavory & Timmermans, 2014). Compared to surveys and experiments, ethnographic fieldwork demonstrated to be an efficient method for investigating scientific practices because it allows the researcher to study communities in their "natural settings" (Shankar, 2006).

My role as a researcher relies on the technique of "participant observation," which allows me to study phenomena directly at the scene of the action. According to Lofland et al., "participant observation refers to the process in which an investigator establishes and sustains a many-sided and relatively long-term relationship with the field studied" (Lofland et al., 2005, p. 18). Typically, participant observation involves activities such as looking, listening, watching, asking and taking notes about social interactions. I joined DataFace teams as a participant observer during meetings, research work, and recreational activities, such as during lunch or other extra-work activities. During my observations I took extensive notes about the ways in which scientists think about and describe their research, their data practices, and the impact that data, code and tools have on their work. I was particularly interested in observing how they go about formulating research questions before and after data collection. I also took notes about how scientists interact with each other, and with other DataFace members during online calls. As I compiled notes of observations, I applied the technique of "memoing," which consists of writing theoretical notes about overarching patterns that are emerging from my observations. The theoretical notes cover reflections on emerging dimensions, deeper

meanings of concepts, relationships among concepts, and theoretical propositions. I often integrated memos with conceptual maps, which allowed me to graphically display concepts and their interrelations. This ongoing process of data collection and analysis served as a basis for the design of this dissertation proposal. It also served as a record of how my understanding of DataFace data practices evolved over time (Babbie, 2012, pp. 399–401; Emerson, Fretz, & Shaw, 2011, p. 19; Lofland et al., 2005, p. 66).

Observations are paired with semi-structured interviews collected with individual team members. Interviews took place after a period of observation of at least one or two weeks. Semi-structured interviews followed an IRB-approved protocol (see ethics section), which includes a set of topics and research questions to be discussed with the subjects. The goal of semi-structured interviews is to collect relevant information while, at the same time, allowing conversations to naturally evolve in un-predictable directions. Semi-structured interviews demonstrated to be appropriate for exploratory studies that use inductive methods for data collections because they allow the research to collect highly heterogeneous and comprehensive data (Lofland et al., 2005, p. 123). Some of the topics to be discussed with the interviewees were decided before data collection started and are based on an interview protocol developed by the UCLA Center of Knowledge Infrastructure (CKI), of which I am a member. Some others emerged from my ongoing analysis of the data collected during preliminary field investigations. During the interviews I obtained highly detailed data that cannot be obtained during observations. For example, during interviews I explored each individual's data practice by asking how she/he reuses data, to what purpose, where she/he finds data, the processes for data

selection and appraisal, what databases she/he uses to find data for reuse, the relation between databases, and so on. I also used interviews to discuss the origins and evolutions of collaborative efforts between scientists by asking in which ways research collaborations are established, what kind of data are exchanged and reused during these collaborations, and to what purposes. Interview data integrated and contextualized observational notes. Interviews are recorded, transcribed, and stored in on a UCLA closed access server.

In ethnographic fieldwork, document analysis is often used as a procedure for reviewing and evaluating printed and electronic written material (Tavory & Timmermans, 2014). Documents are examined and interpreted to elicit meaning, gain understanding, and develop empirical knowledge (Bowen, 2009). For this dissertation research, documents analyzed include (1) grant documents; (2) academic publications, such as conference and journal papers; and (3) lab documents, such as lab notebooks and other forms of lab recordkeeping; (4) digital conversations between DataFace participants on online systems such as group emails, shared notes on Google doc, GitHub contributions, and chatting on Slack. Grant documents are of particular interest because these are records of how participants' understandings of DataFace goals changed over time. They also allow me to compare formal communication with funding agencies to other forms of communications, such as journal articles, and informal discussions during meetings.

Grounded theory

This research study follows the approach of grounded theory (Glaser & Strauss, 1967, p. 67). In social science, and especially in the Chicago School of sociology, grounded method of data collection and analysis tends to be associated to inductive qualitative approaches to knowledge discovery (Lofland & Lofland, 1995, pp. 123–132). Glaser and Strauss proposed that social scientists should build theory "from the ground up," privileging a positivist position while emphasizing an inductive methodology uncontaminated by preexisting theories. In anthropology, library, and archival studies, grounded theory has been employed in association to "interpretivist" approaches. While positivist ethnographers seek to discover timeless truths behind human actions, interpretivist ethnographers tend to acknowledge actors' own constructions of meanings and their subjective and partial interpretations of realities (Gilliland & Mckemmish, 2006, p. 182). Nevertheless, these different uses of grounded theory share a primary concern with discovering concepts, categories, variations and hypotheses directly from the data, where the ethnographer approaches knowledge discovery with few preconceptions about what he or she will encounter in the data (Gilliland & Mckemmish, 2006, p. 178). However, theory is not totally absent in this research method. In grounded theory, data collection and analysis are still guided by an overall theoretical framework, which is composed of all the preexisting studies that informed the topic under analysis (here exposed in the literature review section). It is essential for the researcher to be aware of the literature in order to develop that "theoretical sensitivity" that will guide the interpretation of the data during and after coding (Tavory & Timmermans, 2014, p. 17).

My research design does not rely on a detailed hypothesis of how the DataFace participants reuse data. However, the literature analysis and preliminary field investigation provided me with enough knowledge to elaborate an overarching theoretical approach. For example, I expected researchers to reuse others' data in more than one way, and I expected tools and data management strategies (e.g., data curation) to have an impact on the ways in which data are reused.

Data collection and analysis with "open coding"

In this research study I adopted the practice of qualitative coding as taught by Tavory and Timmermans in their manual Abductive Analysis: Theorizing Qualitative Research (Tavory & Timmermans, 2014). "Abductive analysis" is an iterative process that gradually guides the researchers to the interpretation of the data and eventually to theory building. Typically the researcher does not initiate the data analysis at the end of the data collection, but conducts data collection and analysis iteratively.

I started my initial coding procedure after a few weeks in the field, early in 2015. As I accumulated data from observations, interviews and documents, I labeled my notes by overarching categories, attributes and concepts. I used the "open coding" technique to code my sets of interviews. The "open coding" technique consists in the process of breaking down, examining, comparing, conceptualizing, and categorizing data (Strauss & Corbin, 1990, pp. 61–74). First, the researcher reads the notes collected from observations and the transcripts of the interviews and label the data by different main categories, or concepts. This can be done either on paper or via software, I use a mix of manual analysis and NVIVO IBM software for qualitative coding. For each set of notes,

the researcher asks: What phenomena are emerging from my notes? What these phenomena are instances of? The researcher names the phenomena by overarching categories and attributes. Second, the researcher identifies how the main categories can be subdivided in different instances and nuances. This time the researcher asks: What are the variations of each phenomenon? During the third and last step of open coding, the researcher looks for patterns and relations in the data. How these different phenomena relate to each other? How they do not?

Ethical statement

This dissertation research was conducted with the highest ethical regard. I completed the CITI training and I am a member of the UCLA CKI Institutional Review Board (IRB) approved protocol #10-000909. Dissertation interviews are conducted with the full consent of participating interviewees who were provided with consent information documents and signed IRB-approved consent forms. Consent materials informed the interviewee of the research scope and enabled the interviewee to make an educated decision as to whether or not to opt-in to the study. In addition to the consent materials, an IRB-approved Deed of Gift form was used for all recorded interviews. The Deed of Gift document is signed by the interviewee and ensures that the audio recording and transcription can be used and retained by author and the UCLA CKI research team into the future. Interviewees had the right to complete each form as they felt comfortable, and the right to end participation in the study at any time. The privileges of the individual continue to be respected. Interviewees are always asked, and never forced, to participate in an interview. Interviewees are not quoted by name in this document nor in subsequent publications. While full anonymity is impossible due to the nature of in-person interviews

(Babbie, 2012, pp. 64–65; Lofland et al., 2005, pp. 43–44), all efforts continue to be made to maintain the confidentiality of interviewees under approval of the UCLA CKI IRB-approved study (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1978; UCLA Office of Research Administration, 2015).

5. Findings

The DataFace leadership

Setting aside the official organization of the DataFace Consortium (see Introduction to case study), the participants in this study tend to identify themselves as members of one of the following groups: "the funders," "the engineers," or "the researchers." These informal affiliations mirror participants' understanding of their work practices and disciplinary configurations.

- The funders. The members of the funding body who strategized the overarching goals of the collaboration and allocated the financial resources for its constitution;
- The engineers. The members of the engineering team who developed the online digital infrastructure for data sharing;
- The researchers. The scientists, computational biologists, bioinformaticians, ontology experts, and clinicians who are responsible for the collection and submission of the data.

During informal conversations and interviews, the study participants often mentioned

"the DataFace leadership" as a major decisional force driving the Consortium. Based on my data analysis, the participants seem to use this expression to refer to the members of the funding agency, plus few Principal Investigators (PI) who oversee some of the science spokes. This "unofficial" sub-group of DataFace members holds a key role in taking strategic decisions about the design and the evolution of the DataFace project.

DataFace overarching workflow

While each participating lab employs their own workflow for data collection and analysis – see later in this chapter – the DataFace Consortium is also characterized by a overarching workflow that can be summarized as follows:

a) Scientists and clinicians collect novel data; meanwhile the engineers design and maintain a data model to organize and access the data;

b) The ontology spoke provides the hub with ontological terms and relations to name and classify the data;

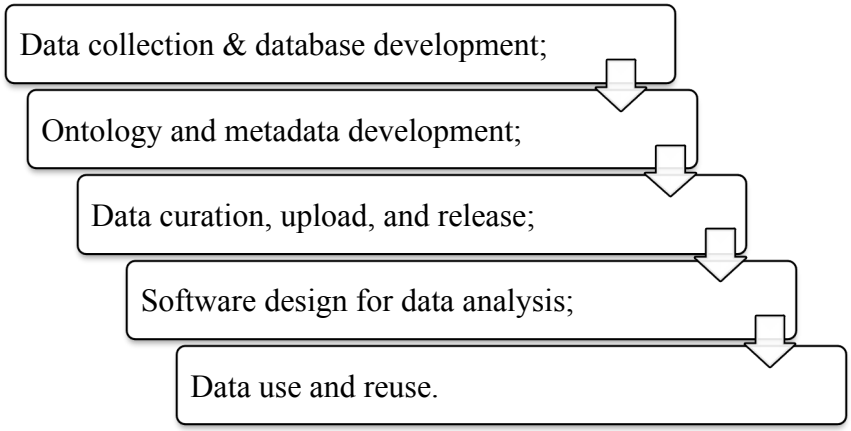
c) Engineers and bioinformaticians collaborate at the development of the metadata that are used to describe the conditions for data collections;

d) In each spoke, one bioinformatics expert is in charge of curating the collected data and submitting the data and the metadata to the engineering hub;

e) The hub uploads the data in the data repository and releases them for open access;

f) The software spoke designs and builds tools for data analysis to help scientists find and mine the data;

g) Scientists from all over the world access, analyze, and reuse data from the digital platform.



Why open data? Rationales for the DataFace Consortium

As discussed in the introduction to the case study, the DataFace mission is to collect and make available, in a secured and organized fashion, biomedical data related to craniofacial research, to the final goal of enabling the reuse of these data by the scientific community.

The datasets collected as part of DataFace II are highly heterogeneous in terms of data types and formats. Data are collected through a variety of experimental practices, from four animal models, and with distinct research designs. Traditionally, the craniofacial field is a very interdisciplinary discipline with relatively segregated laboratory practices (Mossey & Catilla, 2003). The DataFace participants, by centralizing the collection, storage, and access to many different datasets, hope to encourage the integration of a diverse set of skills and expertise that is currently distributed among

many different labs working in the craniofacial domain (Van Otterloo, Williams, & Artinger, 2016; Van Otterloo et al., 2016).

A second feature that characterizes the datasets collected as part of the DataFace Consortium is their relatively large scale. Like many other domains in biomedicine, also craniofacial researchers want to adopt high-throughput technologies for the collection of large-scale genomic data. Such collections represent for the scientists the opportunity of gaining a vast and detailed amount of information on complex biological entities and phenomena. As I will discuss in the next section, regular research grant are not designed to fund the collection of big genomics datasets. The DataFace Consortium provided ad-hoc funding for the collection of these datasets.

Integrating scarce and segregated knowledge

Craniofacial research is multifaceted discipline made of multiple expertise and research interests. Researchers participating in the DataFace consortium operate in a variety of sub-disciplines and possess a vast range of research interests. Not surprisingly, doctoral students and post–doctoral fellows tend to identify "craniofacial research" with the research work that they are personally conducting at this point in time in the labs where they are working. On the other hand, principal investigators (PIs) and agency officers hold a more systemic view of what constitutes craniofacial research. From an interview with a senior experimentalist at the DataFace Consortium:

Travis: […] So there are many aspects of general interest about craniofacial research to the general scientific community. And then there are very specific aspects to the biomedical community. […]

So, for example, in terms of evolutionary biology, think of facial development in most vertebrates, so probably even like in crocodiles, mice, rats, pigs, their facial shape is a very long extended snout, right? Whereas human face is basically taken out, and you have squashed it in flat. So you can see some of that happening between primates and many humans. So like a gorilla clearly has a more pulled-back face than a pig, for example. And then we're even more pulled back, right? So there's the evolutionary aspect of how did the human face get to be the shape it did, compared to all other vertebrates? And that's gonna involve a lot of issues dealing with the brain, because we have a much bigger brain, and the way we hold our brain, is really different. It's gonna have issues to deal with how we chew, and how our jaw works. So there are people out there who are interested in... Did you ever have braces yourself? So there's this idea that the evolution of the human jaw, and the way we eat food now just... There's some debate about it. But the fact that we eat a lot more soft food than probably our ancestors means that our jaws are more crowded. So our teeth are more crooked. So Neanderthals had less crooked teeth than we did because of their diet and their jaw shape. And our jaws, our teeth have not evolved yet to be in a good alignment. So there's that sort of evolutionary aspect of head development.

And then you can also think of it as an engineering problem that there's a lot of sensory systems to plug into your brain that come from your face. So your face is basically, is a way to integrate all these systems with your brain. So your taste, touch, smell, sight, they all come from this area. And it's very close to your brain. But all those nerves and everything, they have to get engineered with the blood vessels, and the bones, and the cartilages, and the ligaments to all fit into this really tight area. So how does that come to? So that's an evolutionary thing.

Now, from a human medical thing, there are... I should go back. So on the evolutionary point as well, there's the questions that Johanna will be asking, like why are different people's faces different shapes? That's also an evolutionary question. How much is heritable? How much is non-heritable? It's not really a biomedical question. It's again, more of a sort of an evolutionary question, a genetics question.

Now, on a biomedical question front, you could have things like birth defects, right? If you read the

general reviews on craniofacial development, it will say that 70% of all birth defects have a head component, right? Approximately 1.5 in 1000 children will have some sort of a facial cleft. So it's quite common. […] There's a type of craniofacial birth defects, it can affect not just the appearance of your face, which has a social aspect to it, but they can affect your hearing, they can affect your breathing, your swallowing, your speech. So there's a lot of problems that come out of human craniofacial birth defects. And correcting them... Even if you correct somebody's cleft as a child, they still might need multiple surgeries, and multiple interventions over time. It's a lot of burden on the patient, the family, with the multiple surgeries.

The extracts below are from a conversation with a NIH officer in relation to the key outcomes of craniofacial research:

Stuart: So obviously, in the distance there should be something that benefits people. So, it could be as immediate future as screening for new syndromes, the ability to recognize particular genes that we should now be screening for in genetic counseling. So, if a family has a syndrome that appears to be running in it, how do you test for that to know who is a carrier and who is not? That's one practical output. The other is that we wanna do regenerative medicine. So, someone has a facial dysmorphology or they have a facial injury, a bone injury; what pathways would you manipulate to regenerate that tissue properly? So, that's another outcome. It's also just the basic science of it. But obviously, for craniofacial research, somebody loses a tooth. Well, you can put an implant and it's totally artificial, or do you have ways to redevelop a tooth, by understanding the pathways that are involved? And people are doing that, so you could...

I identified three macro research-areas in which craniofacial researchers operate:

A. Developmental biology (How does the face systemically develop in the animal and the human embryo?),

B. Human genetics and clinical research (What are the genetic causes of craniofacial syndromes in animals and humans?),

C. Evolutionary biology (What role do heredity and genetics play in facial variation among humans and across species?).

While this heterogeneity of expertise is overall valued by the participants in this study, at the same time researchers working in different labs often expressed the need to integrate each other's knowledge in order to study the face "as a system." Developmental and clinical biologists particularly value systemic and integrated approaches to knowledge production. During lab observations, it was explained to me that the human face evolves *systemically* in the embryos. For the first seven weeks of gestation, all facial components (forehead, nose, jaw, etc.) are compressed in one single organ made of one tissue type (neural crest) that is situated at the top of the spinal cord (Adameyko & Fried, 2016). Starting from week eight, this one tissue starts to differentiate in multiple tissues, and each face components slowly grows out of it. The facial morphogenetic process requires precise coordination between multiple tissue types that behave differently during consecutive developmental stages. This delicate process is common among all vertebrate animals and it differentiates vertebrate from invertebrate animals. When something goes wrong in this process, a person develops a facial syndrome such as cleft lip and cleft palate. DataFace developmental biologists investigate how "normal" facial growth is carried out, while clinical geneticists focus on analyzing what happens when something goes wrong, and a craniofacial "defect" originates in the embryo. Because distinct tissues and cell types behave differently in the embryos, and genes are expressed at different

stages and locations, labs tend to specialize in observing and describing one tiny piece of the facial development puzzle. For this community, increasing knowledge integration is perceived as a way of integrating different pieces of this complicated puzzle. Knowledge integration is necessary to build a systemic understanding of how different tissues and cell types interact to each other to form a "normal" face, or to cause craniofacial syndromes.

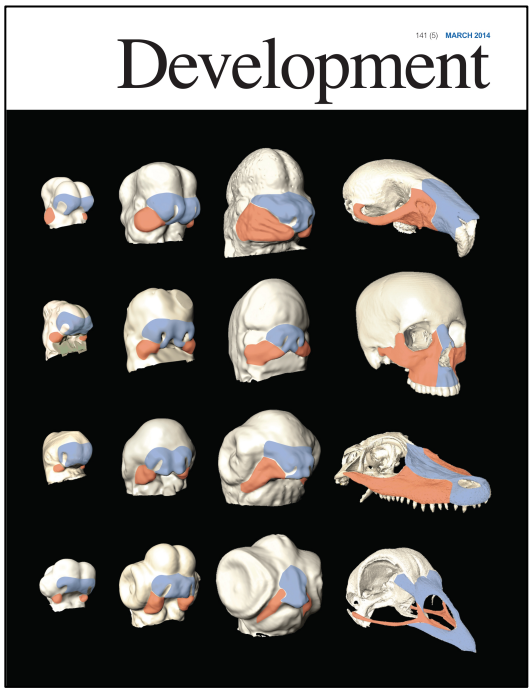


Figure 1: Cover page for Development, March 2014.

The facial morphogenesis in the human and the animal embryos (Young et al., 2014).

Overall, DataFace participants demonstrated interest in collaborating and developing complementary research trajectories between different labs. The challenge of increasing collaborative efforts between different labs operating in the craniofacial field is a recurrent discussion topic during DataFace official meetings and gatherings, during

online calls between the engineers and the scientists, and during DataFace annual meetings. The DataFace leadership is particularly vocal in stressing the importance of collaboration and knowledge integration. During annual meetings, representatives from the DataFace participating labs gather for two full days of discussion on DF advancements. During the first day of the meeting, representatives from each lab share their progresses on collecting, making available, and analyzing their data. The second day is dedicated to develop strategies for what the DataFace leadership refers to as the design of "shared research questions." Participants would divide up in multiple concurring sessions based on research specializations, research interests, or methodological approaches. During these hands-on workshops, representatives discuss potential common research threads. In this interview extract, a NIH officer explains how important it is for NIH to avoid research "silos:"

Stuart: The buzzword at NIH is silos, we do not wanna have silos. And I'm sure you've heard that before many times. So we have, for instance, people who do genetic epidemiology. We have a very strong community in the genetic epidemiology of oral clefting. So that would be like GWAS studies. Big genomic screens to look for associated loci. We also have lots of people who work on mouse models of syndromes. We have other people who work on genetic regulation. Those can all synergize, because when you've got a human candidate for something you wanna know, you have a region of the genome, you wanna be able to figure out which gene you wanna look at, and then you wanna have a mouse model. So by bringing those people together, you get the sort of soup to nuts, one end to the other way of looking at it. So it's much more complete.

For NIH officers, data integration is a means toward knowledge discovery. In this sense, one of the goals of the consortium is to pull together and make available

fragmented craniofacial data collected by multiple labs so they can be re-analyzed systemically. In the following extract, a NIH officer explains that she envisioned DataFace as the "Google Earth" of craniofacial research.

Rose: The biggest promise and the desired outcome is data integration. Many years ago, we talked about DataFace being the Google Earth equivalent of the face, of the craniofacial region. If you look at Google Earth, you can zoom in from the Earth to a single street. You can get information about the house and the neighborhoods and what's related. The vision was that, to be able to have data at different resolution, different types, images, maps, and traffic, if you think about signaling networks as traffic. So that was the vision.

[…] the main goal of DataFace is to have a systems biology approach to craniofacial development. So going back to the Google Earth analogy. So what we can do now with Google Maps and Google Earth is that when you're driving, you know what where the traffic is, and you can go around the traffic and find a different route, because it has that real-time monitoring and real-time predictive power. If you think of it in craniofacial development, if I can predict a birth defect and I can have a workaround to avoid the birth defect by eliciting some alternative signaling pathway so that it does not have to go into the default pathway, that's really what could be done. But it will require the entire knowledge of all the pathways, all the genes and the triggers and what are the roadblocks, what are gateways to a phenotype. So that's really the big, longer vision.

Collecting and accessing "hypothesis free" genomics data

A second rationale behind the constitution of the DataFace project is to fund the collection and sharing of whole genome, exome, and trascriptome large-scale datasets. Whole genome sequencing techniques are also referred to as "high throughput sequencing" or Next Generation Sequencing (NGS). Following a common trend in today's biomedical research, the craniofacial field is facing a methodological

transformation from gene-centric investigations, to the study of groups of genes at once, their regulatory processes, and the epigenetics and environmental factors involved in gene expression and protein making. As I have already discussed, after the completion of the HGP, the genetic community at large realized that single genes are not solo players in dictating life structures (Richardson & Stevens, 2015). It turned out that multiple variants are responsible for a single phenotype, and also that multiple phenotypes can be related to one gene variant. At the same time, new attention is being paid to regulatory processes and transcription mechanisms, whose role in shaping proteins is shaking the once-dominating "central dogma" of molecular biology (DNA to RNA to protein) (Keller, 1984). When high throughput sequencing became available around 2010, scientists employed it to find patterns and correlations across groups of genes expressed in the whole genome, exome, or transcriptome (instead of studying one gene or region at a time). Over the last few years, scientists started using whole genome sequencing to study the roles played by transcription mechanisms, regulatory processes, and the environmental in coding proteins. In the following extract, two NIH officers talk about how the craniofacial research field is changing:

Stuart: I think this is a recognition that is not unique to craniofacial, but it's hitting craniofacial. There was a time, not so long ago, when everybody assumed that all dysmorphologies that had genetic basis were going to be based on mutations in genes themselves, protein coding regions. And that was true of all GWAS, and now it's becoming very... It is very clear that a lot of these changes are actually in regulatory sequences. And so, that emphasizes the collaboration, let's say, between M's type of research, and V, where he studies enhancers, or J. That a lot of these things are regulatory. And I think that's where it's going, is studying the combinatorial effects of all these regulatory mutations.

Rose: Certainly, craniofacial birth defects as a group is one of the largest birth defects, and it does not necessarily only involve the craniofacial region. It actually presents from other syndromic problems. I mean, there are craniofacial manifestations of other systemic birth defects. So like heart defect could have a craniofacial presentation and so on and so forth. And so there are over 1,000 of these syndromic or non-syndromic defects and many of these... I would say most of these are rare diseases, and so chasing after genes at the beginning was very hard. People would want to chase after a single gene, and it was clear that many of these disorders could not be explained by single genes. And so people started looking for gene-gene interaction, gene-environmental interaction. And so this has always been a very big field that attracts state of the art science. And so because cleft lip and palate as a group is the most common type of craniofacial birth defect, a lot of scientists go in that area because you can find the families, you can clone the genes, and that's basically a big focus. Less on the rare diseases because it's just harder to find families and develop the animal models and have the resources to do that.

Emanuela is a young scientist who was recently appointed as tenure-track assistant professor in a prestigious craniofacial lab in the US. She pointed out:

*Emanuela**:** The (craniofacial) field is changing. The human genetics side of it is changing because of technology. The technology of doing human sequencing has just exploded. So what took 10 years to do the first human genome, you can now do in a matter of months. And so we're dealing with huge amounts of sequence data on thousands of people, and that is increasingly being used clinically. And so I think that's where the field is going.*

Before the constitution of the DataFace Consortium, the craniofacial community had hard time collecting and accessing large-scale sequence datasets, for two main reasons. First, funding for the collection of these datasets was not easily accessible. While sequencing the genome of one individual is a relatively small financial investment,

sequencing multiple animals and patients, multiple times, for specific mutations, and at different grow stages, it requires considerable financial and human resources investments. The sequencing itself is a quite automated process, and it is normally outsourced as a third-party service (e.g., to the Broad Institute for the Boston area). However, the work of skilled bioinformaticians is necessary to analyze the sequence libraries for quality control, and filter and annotate sequences to find relevant variants (see Research Workflows*)*. As explained to me by a NIH officer, regular research grants (R01) do not fund the collection of large sequence datasets. The main reason seems to be because "genomics" datasets are regarded as "hypothesis free" resources.

From an interview with a NIH officer:

Stuart: So the idea was this, as I understand it, that Rose (another NIH officer) had a longstanding interest in doing this and... So that was 2000. I came here in 2009, in fact it's now eight years since I got here. And that was at the beginning of a number of these collaborative projects around NIH where people were starting to talk about sharing data, and collecting data that you couldn't generate in a standard grant. So the standard type of our grant is an RO1, that's our standard investigator-initiated funding mechanism. And because of the way they're peer-reviewed and the attitudes of the reviewers, if you just said, "I wanna do a big data generation project," they do not get funded. It's like it's not hypothesis-driven, it's not mechanistic. So there were certain types of data that are very hard to fund, because the applications do not do well. So the idea was, I think to use DataFace to generate those data, and allow people to collect the data and use it for their own mechanistic experiments.

During interviews and observations, participants often pointed out that DataFace data collections are "hypothesis free" data. With this expression, DataFace participants refer

to the idea that large sequence datasets are not collected in a traditional experimental setting where scientists aim to collect one dataset in order to investigate a pre-defined hypothesis. On the contrary, whole genome sequencing data are understood as large pools of data that, after collection, can be coded in different ways in order to investigate many different questions. As I discuss in the theoretical framework of this dissertation, the most attractive quality of big datasets is their capacity to be "used" to investigate multiple patterns at the same time, which is related to the idea of context-independent "fungible" data. Genomic datasets represent a great example of this feature of big data. Ideally, in this perspective, genomic data can be mine by a potentially infinite number of scientists at the same time, each of them investigating their own research questions.

The second reason why craniofacial researchers did not have access to large collections of whole genomes is because they operate in a domain characterized by scarce data. Most craniofacial syndromes, including atypical variations of oral clefting, are considered rare syndromes. This means that there are not many patients that hold these phenotypes. Human geneticists researching rare syndromes need to find as many "matches" as possible between patients with similar phenotypes and related similar mutations in their genotypes. Usually, these few patients are distributed in multiple clinical institutes in the world, and, typically, those researchers who operate in collaboration with these institutes have priority access to the patients' images and genotypes. Some of these data are shared with the whole community. In the US, researchers funded by the NIH are increasingly required to make patients' "raw" genotypic and sequence data available prior to publication in secure databases such as

DbGaP. Processed data and analyses' results are made available after publication, normally by submitting them to sequence dababases such as to the Database of Genotypes and Phenotypes (dbGaP) or to the Gene Expression Ontology (GEO) database. While there are all these regulations for data sharing in place, DataFace researchers still feel like they do not have access to enough data, or to "all the data." In this perspective, the DataFace Consortium is a means for craniofacial researchers to gain funding for the collection and sharing of large-scale "hypothesis-free" sequence data.

The sum up, the constitution of the DataFace consortium was motivated by the willingness to stimulate systemic approaches to knowledge discovery in craniofacial research, the desire to fund the collection of whole genome sequence data related to craniofacial development and syndromes, and the urge to integrated specialized data in a convergent knowledge representation system.

Funding strategy: the U01 and R03 grants

The National Institute for Craniofacial and Dental Research (NIH) funded the DataFace collaboration as a cooperative agreement grant "U01," which differs from a traditional NIH research grant "R01." The Research Project Grant (R01) is the original and historically oldest grant mechanism used by NIH to provide support for health-related research and development. It is defined as "an award made to support a discrete, specified, circumscribed project to be performed by the named investigator(s) in an area representing the investigator's specific interest and competencies, based on the mission of the NIH" (National Institutes of Health, 2017b). A NIH Research Project Cooperative

Agreement (U01) grant's scope is similarly to the one of a R01, however, in a U01 "a substantial programmatic involvement is anticipated between the awarding Institute and Center (i.e., the researchers who receive the grant)" (National Institutes of Health, 2017c). In short, whenever a NIH Institute funds a U01, the Institute does not simply oversee the advancement of the research, but it is deeply involved in setting priorities, designing strategies, and delivering results. Another substantial difference between the two grants is that traditional R01 grants require investigators to demonstrate that their research project was design to investigate and test specific hypotheses, while U01 does not have such requirement. As we have seen, the craniofacial community was eager to collect "hypothesis free" genomics data, and the U01 grant represented an ideal way to fund the collection of this kind of common pool resources.

In 2016, NIH introduced supplementary small grants to the DataFace funding mechanism, called R03s, with the intent of encouraging the reuse of DataFace data by new investigators external to the DataFace collaboration. A R03 is defined as a "grant mechanism that supports small research projects that can be carried out in a short period of time with limited resources" (National Institutes of Health, 2017c). R03 are released with the explicit goal of encouraging secondary analyses of open data. Labs that demonstrated the willingness to reuse DataFace data in their own research projects received the R03 grants. These are labs that are not members of the DataFace Consortium I or II. In the interview extract below, Rose, one of the NIH officers, explains why they chose U01 and R03 grants to fund DataFace:

Rose: Yeah, so we have funding opportunity announcement out there for secondary data analysis,

meaning that these are R03s. People outside of the data producers can go and download data sets and... In mind that those data sets for new information that's being supported as well. So the DataFace hub and spoke projects are supported by cooperative agreements. So they are a little different from grants because we, the government, are substantially involved in terms of managing the direction, the progress and deliverables, such as data uploads. We hold them to a certain frequency and integrity of the data.

Rose: Yeah, the grant is primarily driven by the PI, and they have more flexibility in re-directing. In a cooperative agreement, prior approval is often needed for any redirections. So within NIH, we have a program team to manage these projects. So Stuart is the lead program officer, but there are three other program officers, including myself and a health specialist. So we have regular meetings internally, we make decisions, we look ahead, and that's how even internally we have to collaborate to manage such a big consortium.

Curating and making available data "before the fact"

The DataFace leadership insisted that all data types would be released publicly "prior to publication." From a science policy point of view, this is a quite innovative element, since scientists more often publicly share their datasets (when they share) after they conduced their analyses, and they have published at least one major academic publication based on the data analysis results (called primary analysis, or overview paper). Some datasets are shared years after they have been collected. Holding data until publication is a common practice among developmental biologists working on validating a limited set of genes or regulatory processes. Large-scale sequence data, as we have seen, are more often shared few months (usually 6 months) after collection.

For the funders, openness of research data is also a matter of civic duty. Making all DataFace data open to the public is motivated by the fact that the "tax payers" paid for the collection of the data in the first place. Stuart explains:

Stuart: I think the NIH policy is gonna continue to be to force people to share because the taxpayer has paid for those data to be generated. It's the same thing with PubMed Central, we're sharing our papers. The public's already paid for the generation of those papers. The public shouldn't have to pay again to allow other researchers to access it. So data sharing will continue to be encouraged and enforced. But how to weigh whether keeping all that data, the cost of keeping it all versus the cost of generating new stuff, it's hard to say. Science policy is gonna change so that you do not wanna encourage people to be generating the same new data all the time.

As already discussed, in an ideal situation, publicly sharing data prior to publication creates the opportunity for other researchers to exploit a given dataset at the same time that the scientist who collected the data is doing so. Based on this reasoning, gaining immediate access to research data would make the science process more efficient, and faster. Stuart pointed out that "holding data until publication" is used as an excuse not to share data. Overall, funding bodies look at this practice as highly problematic.

Rose: Our expectation is prepublication sharing. I think we have been very clear about the DataFace. We're not talking about a direct data dump from machine to the hub. We're talking about sets of data, a fairly complete set, maybe to ensure quality of the data before it goes to the hub. But publication is not a hold for not sharing the data. We've always had the expectation of prepublication sharing.

Irene: And why do you think it's better to share the data before publication?

Rose: Because data should be out there [chuckle] to share. I do not think holding data is good for anybody else other than the person who wants to publish it. I think data being shared would be acknowledged so that the producers would still get the credit even without being a first or last author. So I really do not see any downside of data sharing right away, as long as it has the quality.

Sharing data "prior to publication" is also seen as an opportunity to curate the data "before the fact," where the word "fact" literally refers to scientific facts. I have discussed how, in the tradition of model organism communities, bio-curators are in charge of harvesting biology data from publications and institutional repositories and organized them in data structures. This practice heavily relies on the ways in which scientists share their data in the first place, which can vary by journal to journal. Richard (a database engineer) points out that, in his opinion, curating data "upstream" (e.g., right after collection) is better than "after the fact" (e.g., after publication) because it allows for better standardization of metadata and ontological terms:

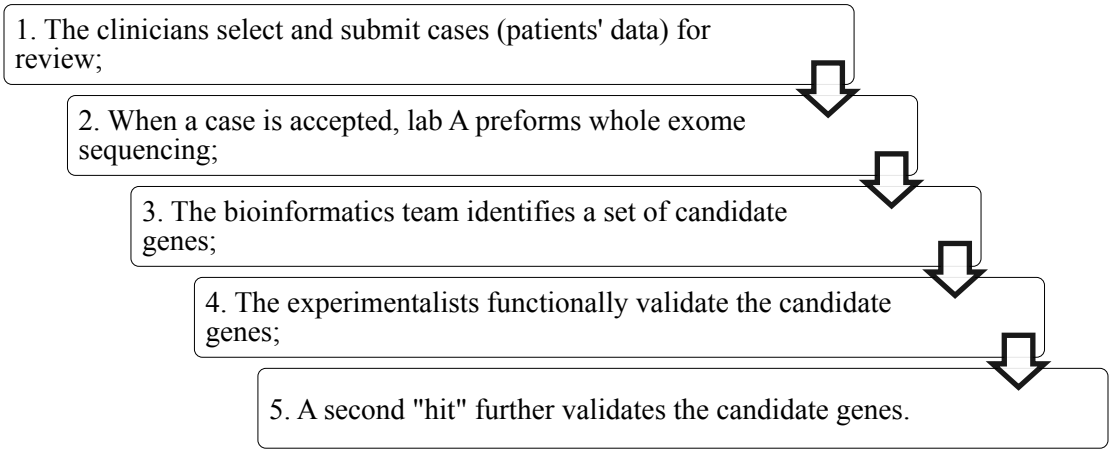
Richard: What a lot of projects are doing is very different from our site. So what a lot of projects have to do, is they go to the literature and what they call curation is basically getting a published journal article, and then they have to go through that and... A team of humans has to go through that and extract. […] That's a lot of times what a bio-curator is called. And they have to go through there and extract things, and then it depends on whether the publisher required a certain terminology to be used. So some publishers will say, "You have to name your genes... " If you're gonna mention a gene, you have to use the go term or something like that. I would suppose the same holds true for other parts of the nomenclature that they might, if required. So we're not bio-curators like that, we run a resource. When we say we do curation, it's a very different sense. We do not do any curating of literature. It's they submit data in a specific format, it's already been agreed on, so we do not have to curate the data after the fact, or literature after the fact. We get it upstream, and they produce it.

Research workflows: data collection, analysis, and release

In this section, I describe the research settings in which selected DataFace datasets are collected. I describe how the DataFace researchers collect and make sense of the datasets that are made available on the DataFace repository. The DataFace researchers collect a wide range of data types, employing multiple research methods and experiments. As previously discussed, the consortium funded the collection of what the participants refer to as "high-throughput data collections." Among DataFace members, the expression "high-throughput" is used to indicate the collection of large-scale datasets via automated technologies. These include whole genome, exome, and transcriptome sequencing and gene expression data, genotypic and phenotypic data from genome-wide association studies (GWAS), and atlases of animal and human images collected via computer tomography (CT scans) and magnetic resonance (MRI).

My analysis focuses on three "sample" research workflows designed by three different participating teams, or spokes. I refer to the three spokes as the Blue spoke, the Green spoke, and the Pink spoke. These three workflows embody research methodologies from several sub-fields of biomedical research including human genetics, clinical research, developmental biology, and evolutionary biology.

Research workflow #1: The Blue Spoke



Domain:	Clinical research, human genetics, functional validation studies
Syndrome:	Cleft and lip palate
Sequencing technique:	Whole Exome Sequencing
Model organisms:	Humans, zebrafish, mouse
Research design:	Candidate genes
Experiments:	Gene editing (ZFN, CRISPR-Cas)
Goal:	Gene discovery and validation
Data submitted to the hub:	Sequence and phenotypic data from humans, zebrafish, mouse

The members of the Blue spoke are two biology labs (lab A and lab B) and one bioinformatics lab. The scientists working at the Blue spoke received funding for the collection and analysis of biomedical data related to the identification and functional validation of ~24 genes involved in craniofacial developmental disorders in humans, in particular cleft lip, cleft lip and palate, and oblique facial clefts. Cleft lip is the second most common birth defect in the United States, affecting one in every 940 births and resulting in 4,437 cases every year (Parker et al., 2010). Worldwide, oral clefts occur in

about one in every 700 live births. However, the co-occurrence of cleft lip and cleft palate, or of an isolated cleft palate, are generally considered rare syndromes (Mossey & Catilla, 2003). For example, cleft palate is present in one in every 1574 births (Parker et al., 2010).

The research on clef lip and palate is a quite specialized – but well-established – research "niche" of medicine, especially compared to cancer research. Few specialized labs have been conducting research on oral clefting for many years now, and a set of genes are thought to be associated with oral clefting in humans. Clefting is also classified as a "complex trait" because it seems to be related not only to a set of genes, but also to a variety of epigenetic and environmental factors. For example, research studies suggest a strong correlation between smoking during pregnancy and increased changes of developing oral clefting in the embryo (Derijcke, Eerens, & Carels, 1996). Building upon this knowledge on the occurrence of oral clefting, researchers at the Blue spoke aim at functionally validate it. This is done in two main steps: first, by identifying and sequencing new human cases, and, second, by creating genetically modified animal models that present the clefting observed in the humans.

Selecting patients, exons capturing, and DNA sequencing

Data collection at the Blue spoke starts at the hospital. The clinicians identify patients – referred to as "cases" – who present facial traits potentially related to one or more genotypes. Once they think a case could be of interest to the spoke, they collect the genetic pedigree and some phenotypic images of the patients, and send this information

to the research labs. The members of the two research labs meet with the clinicians and, together, they evaluate the cases. If a positive decision is made, lab A proceeds with the collection and sequencing of a blood sample of the patient and of few (from one to three) members of the patient's family.

The sequencing technique chosen for this study is Whole Exome Sequencing (WES). WES is a technique for the sequencing of all the protein-coding genes in a given whole genome (US National Library of Medicine, 2017a). These regions are known as the exome. WES technique consists of two steps. The first step is to isolate and select the exons from the sample. Exons correspond to a subset of DNA that encodes proteins. This action is commonly referred to as the process of "capturing" the exons. Humans have about 180,000 exons, constituting about 1% of the human genome, or approximately 30 million base pairs (Ng et al., 2009). The "capturing" step is conducted at lab A, typically by a bench biologist. During step one, the researcher employs what are defined as "target-enrichment strategies" to selectively capture genomic regions of interest from a DNA sample prior to sequencing. It is during this process that Polymerase Chain Reaction (PCR) is carried out ("Nature Chemical Biology," 2005).

The second step is to sequence the exonic DNA using high-throughput DNA sequencing technology, also referred to as Next Generation Sequencing (NGS). NGS for the Blue spoke's data are conducted at an external sequencing facility (e.g., the Broad Institute). During the NGS process, DNA is divided in "short reads" that are particularly well suited to analyze many relatively short stretches of DNA sequence, as for human

exons (Rung & Brazma, 2012). Each read is sequenced and transcribed. Reads are eventually re-assembled using "reference DNAs."

When the sequenced exons come back from the sequencing facility, the bioinformatics lab takes over. The members of the bioinformatics lab for the Blue spoke include Daniela, who is also a principal investigator for the spoke, one software developer, and one computational biologist. The team receives the files from the sequencing facility in "raw" formats, such as in FASTA or FASTQ formats. Raw sequence files contain "quality scores" that represent the probability (p-values) that each read was not sequenced properly (i.e., the base calling process is incorrect). At this point, the bioinformatics lab is in charge of checking the quality of the DNA reads. Hank, one of the DataFace bioinformaticians, refers to this practice as "bioinformatics as a service." In short, the goal of the quality control is to convert the raw, machine-generated sequence data into a useable and useful data resource.

```
@HWI-ST880:63:B01A6ACXX:1:1101:5627:25582 1:N:0:CTCGTA
AAGAACGTCAGGGTTTCCTGCGCGTACACGCAAGGTAAACGCGAACAATTCAGCGGCTTTAACCGGACGCTCGACGCCATTAATAATGTTTTCCGTAAATT
+
@@@ADDDDFFD+<EGEFFCHF1@E8@D@BDFIAA?)=FEFIIFEC>BBB@AAA::@B8BBBABBB87;7@@BBBBBBB<8>>@ADB@>:::<3:<:<&2>A
@HWI-ST880:63:B01A6ACXX:1:1101:5519:25586 1:N:0:CTTGTA
TTTGTTGTTTTACAGAACTCCACAGGAACAACTTCGTACCATGCTACCAAATACATTCACACATCCACATCAAGCTACTGCAGAGGCACAGTGCACTCAGA
+
CCCFFDFFHFFHHJGGIJIJGIIIGGIGIGIIFIIJAGGHIJIIJICHIFBFHBHIIIGGGIJIFIJIJFEECHGDFFFFFECCCBBBBDD>A:A@CCDAC
```

Figure 2: Example of FASTQ file format. It begins with a @ character and is followed by a sequence identifier. The second row contains the raw sequence letters. The format also ecnodes the quality score for the piece of sequence. Source: NCBI website.

Getting "the list": gene mapping and annotation

After the reads passed the quality check, the bioinformatics team needs to find the

patients' disrupted genes (if any genes is disrupted). The disrupted genes are identified by mapping them on the sequences of the known genome, a process referred to as "gene mapping" (Griffiths, Gelbart, Miller, & Lewontin, 1999). Obviously, the genome is too long for the researchers to map the genes manually, so gene mapping is done statistically using a technique called recombination analysis, and a metric called recombination frequency (RF). Sequence alignments and their mapping coordinates are stored in Sequence Alignment Map (SAM) or Binary Alignment Map (BAM) file formats.

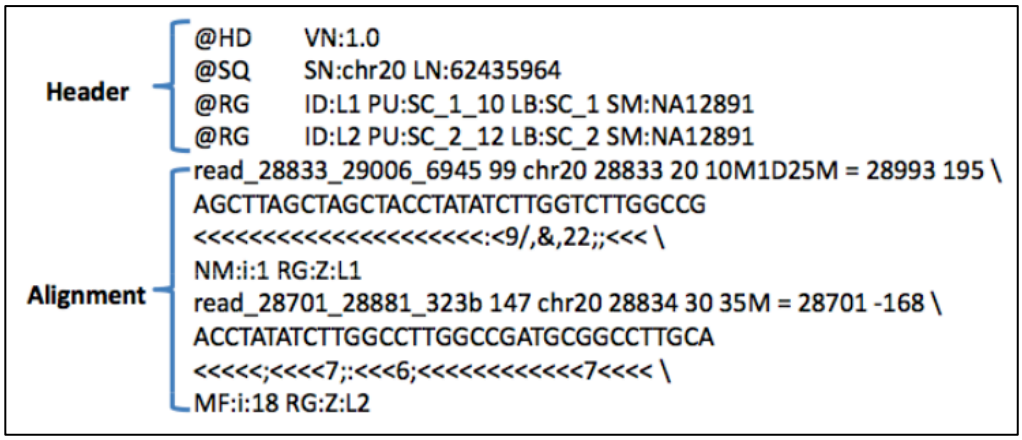


Figure 3: Example of SAM file format. Headings begin with a @ symbol, which distinguishes them from alignment sections. Alignment sections have 11 mandatory fields, as well as a variable number of optional fields. Source: NCBI website.

The "raw" data are now finally ready for analysis. The bioinformatics team can now analyze the SAM and BAM files in order to identify those genes that present potentially interesting mutations in relation to the patients' phenotypic profile ("most likely causal variants"). Data analysis will result in a list of mutated genes (or variants) that the bioinformatics team refers to as the "candidate genes." The shorter the list, the better. This process of data analysis is also referred to as "gene discovery" or "gene prediction."

Once a gene is probabilistically mapped on the genome, scientists need to learn as much as possible about the gene. The practice of "annotation" helps scientists to fulfill this task, and also confirm that the location is actually accurate (US National Library of Medicine, 2017b). Scientists annotate their sequence data by accessing and combining thrives of information – to use Leonelli's expression, "small facts" - about a gene function and related regulatory processes that have been collected, accumulated, and integrated over time by geneticists all around the world. Many of these bits of information about genes are accessible through the website of the National Center for Biotechnology Information (NCBI) (US National Library of Medicine, 2017b). NCBI has links to many repositories of human and animal sequence data, including Genbank itself. The researchers use these repositories, and the related annotations, to get a detailed knowledge of all that is known about that particular gene. Every time a researcher discovers something new about a gene function – through experimental practice – publishes it, and submits it to GenBank, this information is added to the annotated database. NCBI annotated sequences provide up-to-date knowledge about gene locations and functions, and are used daily by researchers as references to annotate their own raw sequences.

The NCBI annotation tools are only a subset of the variety of tools that are available to researchers to annotate sequences. To guide her annotation process, Daniela's team uses a mix of open software and proprietary tools. For example, Daniela's team employs, among many other tools, the Exome Aggregation Consortium (ExAC) browser, which

was created by a coalition of investigators "seeking to aggregate and Hankonize exome sequencing data from a wide variety of large-scale sequencing projects, and to make summary data available for the wider scientific community" (The Broad Institute, 2017). The ExAC Browser hosts sequence data from 60,706 individuals collected as part of various disease-specific and population genetic studies. The dataset allows scientists to type the name of a gene and obtained all the information that is known and available for that gene. Overall, annotation practices vary greatly between areas of research, and even between labs conducting similar research. The team also developed their own "in-house" pipelines, which are tailored on the kinds of research questions they are interested in.

A very popular tool for gene discovery and annotation is the UCSC Genome Browser, which is hosted by the University of California, Santa Cruz (UCSC). Most DataFace's participants use this tools for gene discovery, in a way or another. The interactive website offers access to genome sequence data from a variety of vertebrate and invertebrate species and major model organisms, integrated with a large collection of sequence annotations (Marx, 2013). It hosts a browser, commonly referred to as "The Genome Browser," which is open-sourced and functions on top of a MySQL database. The Genome Browser presents a diverse collection of annotated sequences known as "tracks." Tracks are presented visually on a horizontal axis. Blocks of color along the coordinate axis show the locations of the alignments of the various data types. The ability to show this large variety of data types on a single coordinate axis makes the browser a very precious tool for knowledge integration and annotation practices. To find a specific gene or genomic region, the user may type in the gene name (e.g., CAPZB), an accession

number for an RNA, the name of a genomic cytological band (e.g., 20p13 for band 13 on the short arm of chr20) or a chromosomal position (chr17:38,450,000-38,531,000 for the region around the gene BRCA1). By clicking on the annotations, the researchers can access further details on each note.

Some of the links provide access to other data resources, such as to the Online Mendelian Inheritance in Man (OMIM), which is also highly used by DataFace researchers. OMIM is catalog of human genes and genetic disorders and traits, with a particular focus on the gene-phenotype relationships. Other annotation tools used by Daniela to identify the candidate genes are ClinVar, which is a public archive of reports on the relationships among human variations and phenotypes, and 1000 Genomes, a catalogue of human variation between ethnic populations.

Experimental practice and functional validation

Once Daniela's team identifies a list of candidate genes – or gene variants – this information goes to lab B, which is specialized in designing animal models for function validation studies. Ideally, the job of lab B would be to conduct an animal model experiment that verifies that the candidate genes identified "in silicon" by Daniela and her team are indeed responsible for the phenotype observed in the patient. Developmental biologists at lab B have been working on genes discovery related to craniofacial syndromes for quite some time. They use zebrafish as their main animal model. Kristina is the lab biologist in charge of functionally validating some of the DataFace candidate genes. Kristina is a postdoc with a background in molecular biology and she is specialized in the investigation of the CAPZB gene. She started working on CAPZB in

2014, when a first patient with cleft palate and a disrupted CAPZB gene was identified. Since then, Kristina's goal has been to make a zebrafish model where CAPZB gene is disrupted and the cleft palate phenotype is present in the fish (i.e., to functionally validate the CAPZB gene). To make her zebrafish model, Kristina used multiple gene editing techniques over the years, from Zinc Finger Nucleases (ZFN) to CRISPR-Cas technology. The process of making a CAPZB mutant fish is done in several steps. First, a guide RNA needs to be created and insert in the mother's embryo. There, the guide disrupts (knocks out) the CAPZB gene. The resulting "fish babies" – as Kristina calls it – is referred to as "first generation." If a mutation occurs, and the phenotype of interest is observed (in this case cleft lip or cleft palate), the researcher would cross wild types mouse with the mutant mice until the mutation is stabilized and a new mouse with consistent mutation on CAPZB is created. Karla, a Ph.D. student and Kristina's assistant, pointed out that it might take up to two years to have a stable mutation in a zebrafish. Once a mutation is stabilized, the researcher needs to invert the process and bring back the mouse to a wild-type. Still using a guide RNA, this time the gene (in this example CAPZB) is "knocked in" from the DNA sequence. If the fish does not present the sick phenotype at this point, the experiment for the functional validation of the CAPZB gene in the zebrafish animal model can be finally considered successfully concluded.

After interviewing several postdocs involved in the DataFace collaboration, it is my understanding that a postdoctoral fellow in a developmental biology lab conducts research on an average of one or two genes at a time, and each validation study takes from three to five years to be completed. At this pace, in order to functionally validate 24

genes in zebrafish models in five years (the duration of the DataFace II grant), the Blue spoke would need a dozen of postdocs working on different genes at the same time. However, as the reader may recall, the "real" goal of the DataFace collaboration is not to functionally validate the genes. It is to collect and make available all the sequence data collected from the patients and analyzed by Daniela, and also the experimental data resulting from the preliminary work of Kristina and Karla on the zebrafish model. Data from the patients and from the animal models will be obviously made available before Daniela, Kristina, and Karla publishes their results on the 24 genes.

For Kristina, the fact that her gene CAPZB was included in the DataFace list constituted an incredible opportunity. As she explained to me, it is necessary for her to access as many patients' genotypic and phenotypic profiles as possible in order to compare her experimental results with others' result and further validate her hypotheses. However, as Kristina pointed out during an interview, gaining access to human data is often troublesome because researchers need to be part of pre-approved IRBs. Now that CAPZB is part of the DataFace project, Kristina will have access to all the patients' data that will be obtained as part of this effort by other labs. Not by chance, from Kristina's point of view, the main goal of the DataFace project is to provide access to patients' data:

Kristina: From my understanding and I'm not sure if it's correct, but I believe that DataFace recruits all these patients or at least has information about patients who have craniofacial anomalies. Which is why our lab has been a part of DataFace, is because we're interested in getting into those kind of gene databases where we can actually look at candidate genes. And I think that's what DataFace is for me to give us candidate genes for craniofacial anomalies.

Irene: So, how do you think the craniofacial research community would benefit from the data collected by the DF consortium beyond your specific team?

Kristina: I think that there's a huge opportunity here if the DataFace consortium opens up, because for me, at this point, I found it a little bit hard to get to the data. But I think this is a huge mine of data, which is waiting for the scientist to go into and get all these genes and work on them. Have animal models, some of them mice, zebrafish and then kind of develop on those genes and actually get to what is causing these craniofacial anomalies. So a lot of research-based opportunity is waiting. I, myself have not been able to get to that point, but I think it is highly beneficial.

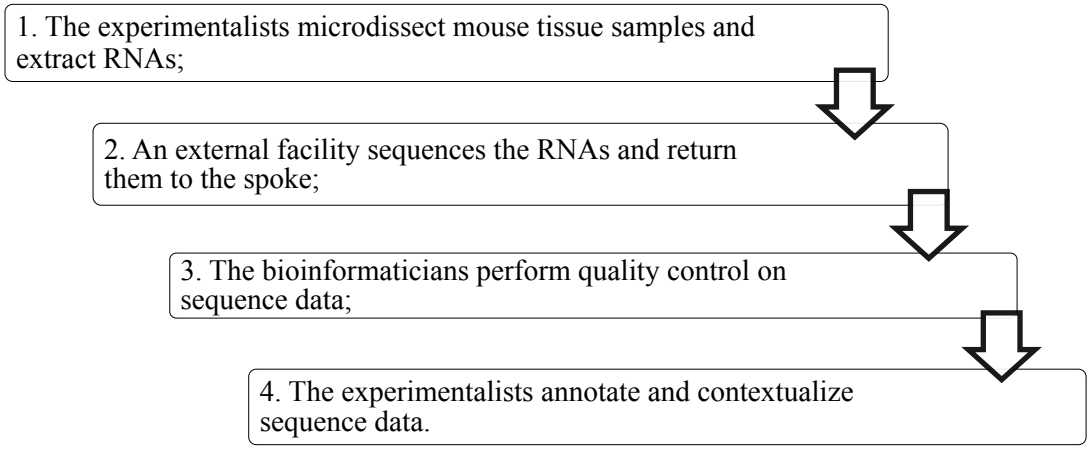
[…] I know there are some whole patient databases where there are patients who have craniofacial anomalies and I can see that those genes have been published or those genes are somewhere there, except that I can't access them. But if they're open then we have this whole list of genes which we... We are not dealing with one patient or two patients, we actually have patients, multiple of those that makes it... We have a strong background to go form our story on.

Since these interviews were taken, Kristina and Karla published a paper in which they functionally validate, for the first time, the role played by the gene CAPZB in craniofacial development in zebrafish mutant model. Next step would be to conduct the experiment in mouse. In this case, lab A, which is specialized in mouse animal models, would take over, and conduct a second experimental study. Ideally, the gene list will be fully validated when both animal models will consistently present the given phenotype (oral clefting), and the team will find a second "hit" for the CAPZB gene in a new patient presenting the same phenotype. Daniela explained to me how hard it is to find a second hit when dealing with rare syndromes, such as the co-occurrence of clef lip and cleft

palate:

Daniela: So, collectively craniofacial diseases is a burden, is a serious burden in the society, but individually each private disease is very rare, so it's very hard to find a second case to see that indeed the gene of your interest might be involved in all type or portion of this type of phenotypes.

Research workflow #2: The Green Spoke



Domain:	Developmental biology
Syndrome:	Developmental syndromes
Sequencing technique:	Whole Transcriptome Sequencing (RNA)
Model organisms:	Mouse
Research design:	RNA expression
Experiments:	RNA-seq
Goal:	Identify and describe spatio-temporal gene expression profiles
Data submitted to the hub:	RNA-seq, microarray RNA expression data

The members of the Green spoke are one biology labs (lab C) and one bioinformatics lab. The goal of the Green spoke for the DataFace project is to collect and share data related to the identification and description of RNA dynamics in the developing mouse face. During the development of the embryo, in mouse like in humans, RNA signaling processes direct the interactions of different tissue and cells that eventually come together to form a newborn face. Facial areas are made up of a layer of ectoderm and a large core of mesenchymal cells derived from the neural crest and mesoderm. Neural crest cells are

a temporary group of cells unique to vertebrates that arise from the embryonic ectoderm cell layer, and in turn give rise to a diverse cell lineage, including craniofacial cartilage and bone (Adameyko & Fried, 2016).

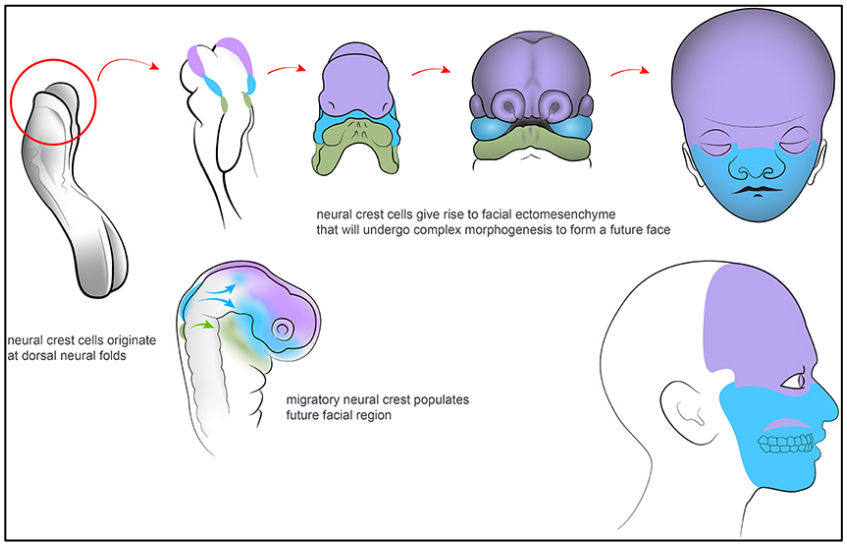


Figure 4: The formation of the face in the human embryo. Source: private slide donated by a study participant.

The manipulations or altering of the RNA signaling processes and tissue interactions have grave consequences for facial development, resulting in various types of medically important dysmorphology, including orofacial clefting. Thus, the Green spoke aims at collecting detailed data on the RNA dynamics and ectoderm/mesenchyme interactions that contribute to the description of facial development. The data collection workflow for the Green spoke can be divided in two main stages: experimental phase, and data analysis phase.

Micro-dissection and sequence of facial tissues in the mouse embryos

The experimental process initiates when Hazel, a postdoctoral fellow trained in

developmental biology, conducts a micro-dissection to remove nasal, maxillary and mandibular components from mice pre-natal fetuses (wild-type). Components are removed at three stages considered to be crucial for facial development, naming at day 10.5, day 11.5 and day 12.5. During the period between day 10.5 and day 12.5, distinct facial parts start to fuse to form the newborn complete face. These distinct facial components are separated and analyzed to determine how their unique expression profiles correlate with their eventual fates. After the components have been removed, the ectoderm is separated from the mesenchyme (citation redacted for anonymization purposes). At this point, the biologist would extract different parts of the RNA, including mRNAs and a variety of small RNAs, from the separated tissues. The experiment is conducted using RNA microarray for data collection. Once the RNA parts are extracted, these are sent to an external sequence facility to be sequenced.

Data analysis: bioinformatics meets molecular biology

The Green team deeply values a multi-interdisciplinary and systemic approach to knowledge discovery. Team members have mixed expertise in craniofacial biology, mouse molecular genetics, bioinformatics and computational biology. Two senior experimentalists, Jane and Travis, recently re-trained themselves in bioinformatics and computational analysis. Jane explains:

Jane: I am, on Mondays and Wednesdays and Friday a bioinformatician, and on Tuesdays, and Thursdays and Saturdays, I'm a development biologist. (…) So, classically the way I was trained we think about a single gene and what it does in the system. But this [e.g., what they are doing] is a much more systems biology kind of approach, where we try and think about all of the information in the entire system and how it integrates. [single-gene and whole-genome sequences approaches] are very complimentary.

They are the yin and the yang. And neither one means much or you can get a little bit out of either one but you do not even begin to embrace the whole thing until you use both.

In this team, as opposed to the Blue spoke, the same individuals who conducted the experiments carry out the core analysis of the sequence data. When the sequence dataset comes back from the sequencing facility, it first goes to a bioinformatician, Kaine. Kaine performs quality control over the sequence libraries, and then converts these into BAM files. At this point, BAM files are taken in custody by Jane and Travis. BAM files contain sequence alignments data in binary formats. Multiple statistical tools and programs – especially R packages – allow researchers to visualize and analyze BAM files in order to identify variance in the expression patterns. Jane and Travis investigate the significance of those variants by analyzing and annotating the RNA expression data, similarly to what Daniela does with gene sequences for the Blue spoke. Like Daniela, also Jane uses a diverse set of databases and tools for annotating her data, including the Gene Ontology, the Gene Expression Omnibus database, the Reactome Pathway Database, and the KEGG Pathway Database. Jane explains how she uses these databases for annotation:

Jane: The gene expression type of database that we reuse, that would be to provide more depth and context for our gene expression data. And the databases that have more functional annotations, we use them to interpret our transcriptome data. Oh, a database that we use extensively, that I did not mention, is miRBase. And the various others are in a classes of RNA databases. There's the snoRNA database, and there's... I can't remember the name… but there's a number of databases that annotate, essentially annotate, all these other classes of RNAs besides messenger RNAs.

Jane: A major thing that we want to do [by annotating the data] is to take our 25,000 of genes or so

that we have, and reduce the complexity by binning them, into either pathways like, A makes B makes C, or functional groupings. They're all involved in doing the same thing. So all this annotation information either for pathways, or for function is fundamental to doing that, reduction of complexity that we need to do to get the system's level analysis to work.

For Jane, biology and bioinformatics approaches to knowledge discovery are complementary, but at the same time fundamentally different. In her daily research, Jane works hard to find ways to integrate them.

Jane: I've spent the last five years learning how to communicate with, and think like a bioinformatician, and what the resources are and what the questions are. And they think in very different way about the problems than the biologists do. They are fundamentally interested in different questions. And as a biologist I'm interested in how things work, and an informatician is interested in how to transform data into information (...) And it's not black and white, but it's where the most emphasis is. And as biologists, we get involved in the details and the exceptions to the rules and the informaticians throw those out as noise. And as biologist we think about it one gene, and one protein, and one molecule at at time, and build our picture of the world from the bottom up. And the informaticians think about it on the systems level and their picture of the world comes from the top down.

For Jane statistical analyses and sequence annotation are a starting point, an indication that something interesting might be happening in those tissues. Once she is done employing databases and tools to annotate sequences and find significant variants, Jane goes back to her knowledge of molecular biology and uses it to contextualize and make sense of the analysis results. As she explained to me, Jane makes sense of RNAs expression patterns by looking at molecular changes in "time and space." For instance, Jane's team found that the mesenchymal samples are very different from the ectodermal

samples. By micro-dissecting and observing gene expression profiles at three different grow stages, she was able to locate how individual genes differently activate in distinct tissues (space) and at distinct grow stages (time).

Jane: I start with the group of genes that are in this particular cell type, at this particular age, in this particular prominence that are different from the ones here and here and here. And then, I say, "So, what is different between, for instance, what's going on at E10 and a half in this cell, and what's going on at E12 and a half in this cell? How do things change over time?"

And so, you have to find in this group of 1000 genes the ones that are interesting in this particular way. So, what's changing as we age is the question. And then, you see, okay, these are the ones that we see a lot of things that are involved in cellular structure. A lot more of those in the old ones, where the young ones, we all know embryos are soft and squishy and so, in these ones, we see much more of things involved in communication between the cells.

And so, we would then interpret this to say that and the young cells are primarily involved in communication between the cells, and maybe establishing the transcriptional patterns that have to do with cell fate. Whereas by the time we get to the older cells, we're involved in the structure of a cell. So, maybe a cell that has a lot of structure is gonna be making cartilage or something like that.

In the slides below, Jane visualized the results of this research work. The content of these slides is pretty unique, it allows Jane to visualize the results from the whole transcriptome analysis at a molecular level at two different developmental stages (Figure 11 and 12). *De facto*, these busy images enable Jane to translate results obtained via informatics approaches into a "developmental biology" perspective: "in time and space." Bioinformaticians generally represent gene expression differences during development by means of diagrams and bar graphs, where genes are represented at a summary level (see Figure 10). Developmental biologists visualize biology processes by means of drawings

in which they visually map processes on to the physical cell. In the slide represented in Figure 11, Jane mapped multiple genes obtained from whole transcriptome sequencing one by one on the physical cell and on all its functional components, including the nucleus, the secretory apparatus, and the cytoskeleton. And then she did it again for the following developmental stage (Figure 12). In this way, Jane can explain and visualize which gene is responsible for which action in the cell, individually, and at different times.

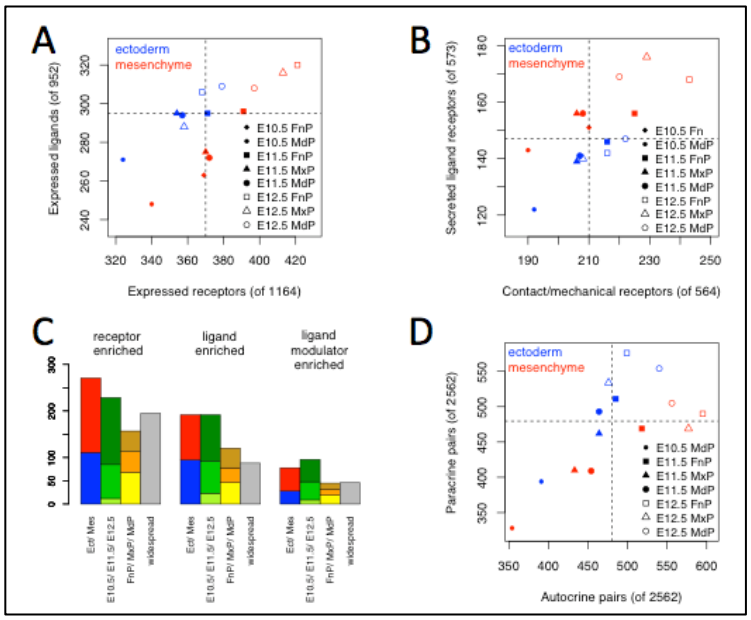


Figure 5: A bioinformatics view of cell-to-genes interactions. Source: private slide donated by a study participant.

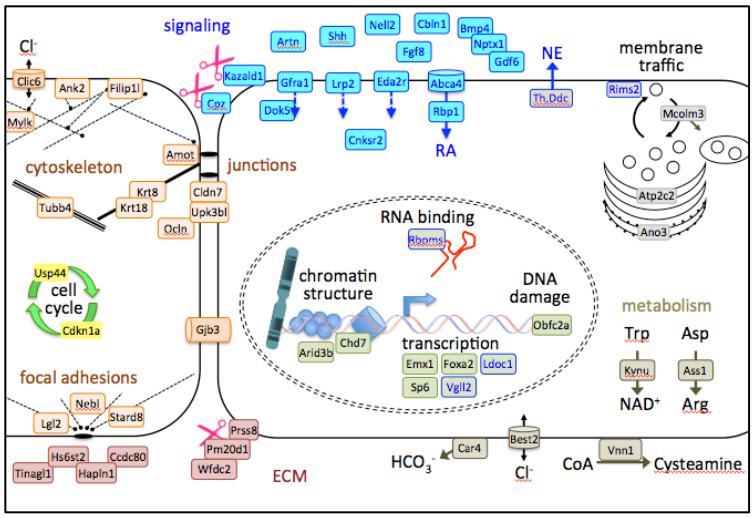


Figure 6: A molecular biologist's view of cell-genes interactions (early developmental stage). Source: private slide donated by a study participant.

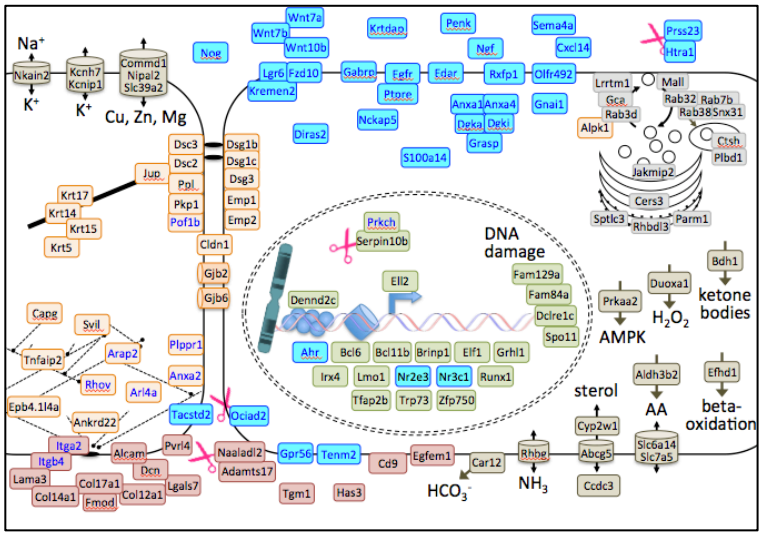
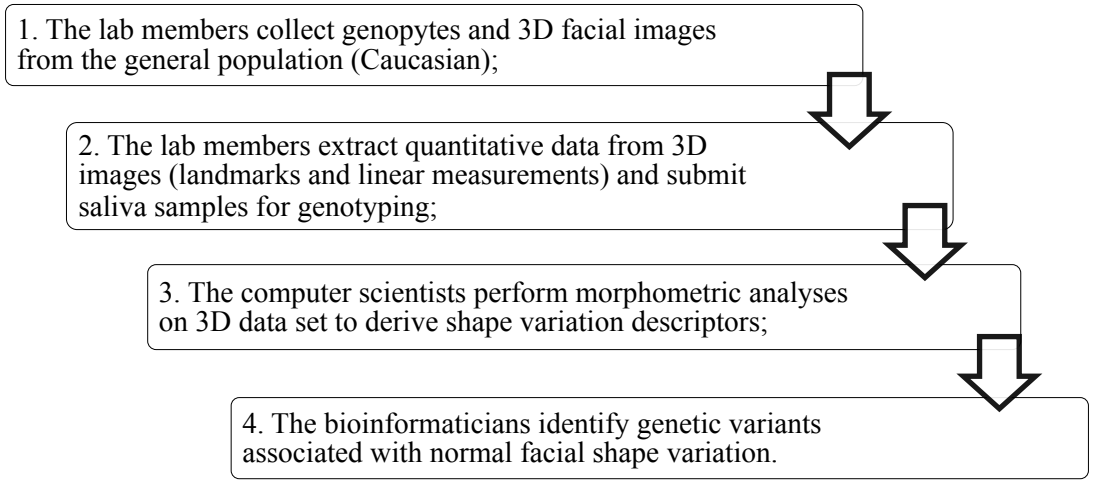


Figure 7: A molecular biologist's view of cell-genes interactions (later developmental stage). Source: private slide donated by a study participant.

Research workflow #3: The Pink spoke



Domain:	Computational Biology, anthropometry, human genetics
Syndrome:	Craniofacial complex traits
Sequencing technique:	Genotyping
Model organisms:	Humans
Research design:	GWAS, SNPs analysis
Experiments:	N/A
Goal:	Quantify normal facial variation, identify associated genes
Data submitted to the hub:	Facial images, genotypes, facial landmarks and measurements

Members of the Pink spoke include one computational biology lab (lab D), and a set of external collaborators including human geneticists, statisticians, and computer scientists. For the Pink spoke, data collection started during phase I of the DataFace funding cycle (2010). The Pink spoke is one of the few spokes that received continuous funding from the DataFace project from phase I through phase II (until 2020). Currently, the human genomic and imaging data collected by the Pink spoke during phase I are the

most requested and reused data among all data types collected by the DataFace collaboration. For DataFace I, the spoke built a repository of 3D facial images and measurements called "3D Facial Norms Database," which is accessible from the DataFace website. For DataFace II, the spoke developed a software package for data discovery that can be used to visualize summary level p-values of the data that they collected in phase I (and potentially of others' data). In DataFace II's official documentation, the Pink spoke is referred to as "the software spoke."

The making of a genome-wide association study (GWAS) for normal variation

The overarching goal of the scientists operating at the Pink spoke is to address the current "dearth of information regarding how variation in specific genes relates to the diversity of facial forms evident in our species." The research design used for data collection is called genome-wide association study (GWAS), which is defined as:

> A genome-wide association study (GWAS) is an approach that involves rapidly scanning markers across the complete sets of DNA, or genomes, of many people to find genetic variations associated with a particular disease. Once new genetic associations are identified, researchers can use the information to develop better strategies to detect, treat and prevent the disease. Such studies are particularly useful in finding genetic variations that contribute to common, complex diseases, such as asthma, cancer, diabetes, heart disease and mental illnesses" (National Human Genome Research Institute, 2015).

As per definition, a GWAS consists in the collection of genotypic and phenotypic data from a large population, it is typically used to identify genes involved in a particular disease, and it is performed on affected patients. The Pink spoke collected data from a "normal population" that is not affected by any syndrome. The spoke collected "normal

variation data" in order to use them as control data in successive studies on oral clefting. 3D facial surface images and DNA samples from 3500 healthy Caucasian individuals (age 5-40) were drawn from the general populations. Saliva samples and facial images were collected by graduate students working at the spoke computational lab, and in other partner labs, at four different locations on the West Coast and in the Midwest.

Quantitative facial measures were extracted from the 3D facial images (Figure 13). Images and measurements were organized and made accessible in the DataFace 3D Facial Norms database. Images were analyzed by performing morphometric analysis of mid-facial shape differences in the general population, this analysis resulted in the identification of multiple shape variation descriptors. The morphometric analysis of the imaging data was conducted at a partner laboratory located in Europe. I provide further details about the GWAS datasets in the section where I describe the ways in which these datasets have been reused.

The DNA extracted from the saliva was sent to an external facility to be genotyped. Once the data came back from the facility, these were analyzed by a group of computational biologists that includes two human genetics experts, one statistician, one computational biologist, and one computer scientist. The genotyping data are hosted on the NIH Database for Human Genotypes and Phenotypes (dbGaP). Genotypes data are analyzed to identify genes potentially associated with normal "non-syndromic" facial shape variations. For example, the team identifies few genes that seem to be related to the attachment of the lobe to the ear, which is present only in certain individuals, or to the

length of the nose.

At the Pink spoke, the postdoctoral fellows with computational biology training conduct the data analyses. They do so via gene mapping and data annotation techniques similar to those I described in the Blue spoke's workflow. However, the Pink spoke "genotype" the saliva samples, as opposed to "sequence" them. The term "genotyping" is used to refer to the process of determining which genetic variants an individual possesses. "Sequencing" is used to refer to the method used to determine the exact sequence of a certain length of DNA. Researchers can sequence a short piece, the whole genome, or parts of the genome, as in the workflows of the Blue spoke. Depending on the region, a given stretch of sequence may include some DNA that varies between individuals, in addition to regions that are constant. Thus, sequencing can be used to genotype someone for pre-selected genetic markers, as well as to identify variants that may be unique to that person. Genotyping can be performed through a variety of different methods, depending on the variants of interest and the resources available. Genotyping is commonly used in GWAS studies, where whole genome sequencing is more expensive and time consuming. Sequencing is slowly replacing genotyping techniques (Roetzer et al., 2013; Yang et al., 2016).

Because of the computational nature of this lab, most data analysis for genotypes data is performed in house. The principle behind a GWAS study is to associate phenotypes with genotypes. As we have seen, GWAS are more often conducted with syndromes' data, where "sick" phenotypes/genotypes are compared to control phenotypes/genotypes.

Because the spoke collected data on normal variation, the researchers first had to define what their phenotypes of interest would be. I will get into the details of how facial phenotypes for normal variation are established in the chapter about data reuse. Once phenotypes are established (e.g., facial images with a broader nose), the researchers would look into the associated genotypes to verify whether certain variants, or "SNPs," are common within individuals who present the phenotypes. A single nucleotide polymorphism, or SNP, is a variation at a single position in a DNA sequence among individuals. The concept of SNP is quite important to understand the functioning of a GWAS study. The DNA sequence is formed from a chain of four nucleotide bases: A, C, G, and T. If more than 1% of a population does not carry the same nucleotide at a specific position in the DNA sequence, then this variation is classified as a SNP. If a SNP occurs within a gene, then the gene is described as having more than one allele (Nature, 2014). Although a particular SNP may cause a disorder, some SNPs are associated with non-syndromic traits, like in this particular GWAS study. These associations allow scientists to look for SNPs in order to evaluate an individual's genetic predisposition to develop (in the embryo) or to possess a certain facial trait.

Once SNPs are identified, researchers need to dig into the literature to integrate their findings with everything else that is known about these genetic markers and their biological functions. Julieta conducts most of this "interpretation work" for the lab. Julieta is a postdoctoral researcher from Belgium. For example, in order to find the SNPs that are associated with the attachment of the ear lobes in humans, Julieta analyzed and annotated the literature about over 358 genes (Shaffer et al., 2016). For her annotation

work, Jasmine consults a wide range of integrated resources. Julieta uses the Mouse Genome Informatics (MGI), and the Online Mendelian Inheritance in Man (OMIM) database to investigate whether certain genes or SNPs variants have been previously associated to syndromes that affect the formation of the face, either in mice or humans. She also uses the VISTA Enhancer browser to check whether genes are located in "enhancers." Enhancers are short regions of DNA. Whenever a protein, or activators, is bound to an enhancer this increases the likelihood that transcription of a particular gene will occur (Nature, 2017). These proteins are also referred to as "transcription factors". Finally, she looks for clues in the DECIPHER database, which is an online clinical database that contains human genome variants and phenotypes of thousands of patients worldwide.

DataFace workflows' summary

Research workflows at the DataFace Consortium span a variety of data types, research questions, research interests, and methods for data collection and analysis. In the examples provided in here, genomic data are collected and employed for the diagnosis of rare syndromes (Blue spoke), to study how RNA expression varies in time and space during mouse development (Green spoke), and to identify how certain genetic markers possibly relate to certain facial traits (Pink spoke). Researchers use a variety of sequencing techniques to collect genomic data, namely exome sequencing, transcriptome sequencing, and genotyping. Sequencing and genotyping are performed on humans, zebrafish, and mice. Sometimes humans and animal models are sequenced as part of the same research setting, and sometimes in completely independent experimental settings.

The processes of gene mapping and "data annotation" play a central role in the interpretation of the sequenced "raw datasets" as they come back from the sequencing facility. Annotation can be done in many different ways, and it tends to be idiosyncratic to each lab. Annotation can be conducted by using in-house pipelines, or online data analysis tools, or a combination of both. The annotation work is distributed among team members based on computational skills and domain knowledge. Sometimes it is distributed between developmental biologists and computational biologists, for example at the Green spoke, and other times it is mostly outsourced to a computational biologist lab, such as at the Blue spoke. Some teams, like the Pink spoke, are composed of highly technically skilled geneticists and tend to do all the annotation work in house.

Data are released right after data collection, while the presented research workflows were still in progress. The sequence data are released when they come back from sequencing facilities, – after quality control – in raw formats (e.g., SAM, BAM files). Phenotypes data such as images and scans were released after these were taken. As discussed earlier, releasing data as soon as possible, prior of publication, is a core mission of the DataFace collaboration.

Building tools for data search, browsing, and discovery

In this section, I describe how the DataFace participants designed and implemented infrastructural solutions for data sharing, retrieval, integration, and reuse. Building tools to find and access data revealed to be a complex and multifaceted activity that requires multiple specialized skills in database design and management, data curation, data search, data visualization, statistical analysis, biology and genomics.

The database engineers at the engineering hub are in charge of creating the database schema, uploading the data into the database, and designing a user interface to filter and retrieve the data. The DataFace leadership was conscious of the fact that database engineers needed support from domain scientists to properly understand the data and curate the data with the proper metadata. For this reason, the leadership asked each science spoke to appoint "go-betweens" individuals that could help the engineers to better understand how to organize the DataFace data. Each spoke appointed one person with bioinformatics skills. This person, referred to as "the bioinformatician," is responsible for discussing data curation strategies with the database engineers (i.e., pick the right metadata), curating the data (i.e., apply metadata), and sending the data over to the engineering hub. The hub scheduled weekly meetings between the appointed bioinformaticians and the database engineers, during which they discussed best uses of metadata schemas, ontologies, vocabularies, and so on. In addition to the weekly bioinformatics meetings, the collaboration also set up monthly meetings with all the DataFace PIs. As I have already discussed, the process of releasing and curating the data was conducted right after data collection.

From an interview with an NIH officer:

Stuart: One of the things I redid between, when I rewrote the RFA, is I remember that the informatics people in DataFace I had a very hard time communicating with the Spoke projects. They didn't speak the same languages at all. And that was a determent to the Hub, that made their life much more difficult. So we required the second time that every project have an informatics scientist on their personnel list. And one of

the things the reviewers judged was, is this spoke project capable of speaking informatics? So that C, you do not have the two sides to start jabbering at one another and not understanding anything that they're talking about. Because C is not a biologist, and the Spoke people aren't generally informaticists, so we needed a translator.

Stuart: Everybody had to have a go-between, and there was actually a committee early on, a working group of all the informaticists from all the projects and they would work with the hub. So how do you curate? So it's lessons learned the hard way through DataFace I, how do we get the data from you? So they've made these spreadsheets, you will use the spreadsheets. That was a problem for DataFace One because people didn't listen much as we pushed, they didn't understand the need for structured curated data.

From an interview with a database engineer:

Richard: I think the ultimate challenge for us is working with the spokes to help identify what will be the high-value datasets for the community 'cause ultimately, I think the technology will be important and could be a road block, but ultimately having the right data that satisfies the needs for the craniofacial research community is gonna be the biggest challenge. And that's something that we as the hub and the computer scientist in the group can't independently solve that, but we can work with the domain experts to help them identify that.

Richard: So, we would kind of start by trying to get a rough sketch of what the problems are, do some brainstorming internal to the hub. We meet weekly on Mondays to kinda discuss some general ideas. But then, we have a regularly scheduled meeting with the informatics team. And so, the informatics team is comprised of a minimum of one representative from every spoke. And so, we go there to basically present ideas, discuss, collect input. We sort of kicked things off by having weekly meetings with them, and we went through kind of a more rapid ramp-up to get a handle on the data and such. But in general, so we hold these weekly meetings then with all of the spoke representatives in the informatics team and discuss issues

that can be cleaning up the vocabulary, talking about the data structures down to more mundane issues of who has data to be submitted this month or coming up, so tracking data submissions and stuff.

Data modeling, searching, and browsing

The first step of creating tools for data access and retrieval is to organize highly diverse sets of data in a proper database model, so that a user can query the data collection. The DataFace search engine, which is called "the data browser," was developed to work as an intermediary between a set of specialized users and the datasets themselves. The platform hosts over 700 individual datasets. In the following extracts, Andrea (informatician), Hank (computational biologist), and Rose (NIH officer) point out that the DataFace database was envisioned as a cross-searchable repository of craniofacial large-scale datasets that could be downloaded and re-analyzed by the community.

Andrea: (DF's platform) It's more an organization of the data. I do not know if... It enables the analysis of the data but it's not... It is analyzing the data in terms of characterizing the data but it's not the type of analysis that will help you decide whether this gene has or does not have the right place or the right role into certain type of phenotypes or certain problems. The way we organize it helps make those analyses.

Hank: ...we're just basically translating the data that we get and we upload it to DataFace hub, so that the data is available for others to do data analysis. So even though we do that data analysis ourselves as well, for the grant, it's just getting the data and processing it, mapping it and making it available in the hub so that the wider craniofacial research community can start working. [...] they can do other types of analysis that we do not do or integrative analysis with other DataFace projects or look for their favorite genes or whatever they want to use the data for.

[...]

(DF's goal) It's kind of comparable to I think what the ENCODE effort has been doing. It basically creates

a rich resource of different types of data relating to craniofacial development that can then be mined by different research groups with different research interests to get an added value out of those data that will help describe craniofacial development, regulatory mechanisms, gene expression profiles, regulatory elements and regulatory factors, new candidate genes.

Rose: Short term, I think I would be happy, and that's where we're at as a repository of data, as a reference set, so that we would know what a normal face would look like, so that it could be compared with any deviations for diagnostic purposes. It can be a reference set to collect animal models that would be reflective of craniofacial birth defects, and serve as a reference that data is reproducible, if others want to reproduce any particular data set, so that's fine. Some sort of correlation of, say, a mouse, embryonic day 10 is equivalent to a human day 60 of pregnancy or day 50 of pregnancy. That's fine in a short term, if there's some of this correlation. I think there's knowledge to be discovered further. It's just that the longer vision seems quite challenging. [chuckle]

In the following extract, Stuart points out that enabling effective data discovery is a core part of the DataFace mission. He envisions the DataFace online platform to work as the "Google" of craniofacial research data.

Stuart: [the DataFace mission] *it's encouraging data science, so it's encouraging... Yes, some of it's reuse, some of it is like the data commons where stuff will be moved to the cloud, and instead of having to download 12 terabytes of data, you move – what's the jargon? – you move the compute to the data; you do everything on the cloud. Other parts of it are, for instance, the data discovery index. There are tons and tons of data out there and they're in repositories, but you find them; [chuckle] it's a treasure hunt. So there is a project going on to try to index all the data and be able to find it, so it's searchable, like a Google for data basically. And that's a potential goal for DataFace in the end as well. What's out there that other people have generated outside of DataFace that DataFace can now point investigators to where it becomes a hub, a real informatics hub where it's like, "Here's all the data. You wanna find something, you do a search, we give you URLs."*

During interviews and informal chatting, the DataFace leadership often referred to the DataFace search engine as "the Google" or "the Amazon" of craniofacial research. Database engineers, like Richard in the extract below, portray Google and Amazon platforms as the "standards" for best practices in data management and retrieval.

Richard: what I do for my work (…) so (consists in) exploring the techniques that have kind of have been popularized in the enterprise and professional media and personal or prosumer kind of spaces for managing content and then applying those to scientific data management.

The DataFace database was programmed to search and access data at different levels of granularity. Scientists can search for a specific gene mutation, or they can visualize a list of all the available gene expression datasets. Database engineers looked for ways to make the retrieval of specialized data "easy" for all scientists, not only for those specialized in the sub-discipline in which the data were collected in the first place. In order to do so, the DataFace search engine enables cross-search between datasets belonging to different categories, which means that datasets can be retrieved and classified from multiple "points of view" (i.e., expertise). Engineers refer to these modes of data seeking as "data search" and "data browsing." Engineers also distinguish data search and data browsing from "data discovery."

In the following interview, Christopher (one of the engineers) explains how he needs to build a data model that can satisfy different types of data seeking behaviors.

Christopher: Well, yeah, I mean, internally, it's always a resource constraint thing that you're trying to figure out what you can do without throwing a huge set of different dedicated specialists at different tasks. So it's kind of balancing different roles. Externally, it's more about the domain model and the fact that a lot of these groups, what they... They have specialized domain concepts, but they do not necessarily have information modelers and highly fluent informaticians or database administrators or anyone who's really fluent and can put an interface between them and us. So we're kind of... The database administrator or the data modeler is one of the roles that we sort of manufacture out of all of our shared input without having an actual person dedicated to it. So I think that's one of the challenges and that probably makes... One of the biggest things that makes the project work or not is whether you find data models that work well with the computer and work well with the users.

It was essential for the engineers to avoid what they refer to as "the data dump" effect. In a data dump, datasets are not explicitly related to each other because that have been indexed separately, and, as a consequence, they are difficult to locate and retrieve by non-specialized users. As it emerged from my interviews, database engineers often refer to the first iteration of DataFace database as a data dump.

Christopher: Well, so the new system we've been rolling out, I pretty much am the architect of it and the primary developer of a lot of the actual back end data services catalogs. So I'm pretty happy about that, that it's being put to use. Also, we had to take on the Legacy system from the previous iteration of the project and basically receive a big data dump from another institution and try to piece it back together and get it up and running so that we could serve users in the interim. So all of that was sort of a challenge specific to this project and it worked out alright, and we've been migrating the users to, I think, a more useful way of getting their accounts managed through public Google-based services, basically, instead of having to have our own local team administering accounts, and that was kind of an important thing, I think, for us.

Christopher: Right, I mean, a big shift was that... The previous system, I would say, was very ad hoc, which is typical of the kind of web platforms that it was based on and the decade that it came out of, but it's essentially, you ended up having... It was really like a website, a publishing system that people just kept piling more things into, but there wasn't much consistency to it. So you'd have different sections of data that were only really reachable through a different set of pages and a very different navigation model, and different organizations. So there wasn't really any consistency to walk in from the outside and say, "What's everything here," and, "What's it about?" And so by moving it into our much more structured or explicit database management system, we have a much more specific idea that all the data has a certain kind of relational representation and there's a standard search interface where you can browse through it all using the different facets of it and find about authorship or kinds of organisms or projects and dig down in a much more consistent way across all the different contributors, as opposed to feeling it's almost a collection of little mini websites, each of which had a very different authorship and model, so...

To avoid the data dump problem, one of the solutions is to provide granular access to the DataFace data, as explained by Richard, one of the hub engineers:

Richard: We do wanna take one big task of the database and we're exploring a big change to the structure of the database, which would be to expose... So DataFace's data, since the beginning of DataFace, one has always collected data in these big bundles, like a big ZIP file, basically. Large archives that have, maybe data on three, four, five, six specimens as opposed to... And so it might be age stages of a particular specimen, mouse or... Typically mouse data. And what we'd like to do is split those up into individual elements and so that people can see more easily what's in DataFace. 'Cause there's actually a ton more data in DataFace than what one might realize by looking at these larger bundles of data. So we have a smaller number of large bundles, there's actually a lot of files in them, there's actually thousands of individual data points that have been collected in DataFace. And it would be more usable for some users to be able to pick and choose things at a more granular, rather than there's a big multi-gigabyte bundle, they might wanna be able to get individual... And it would also allow them to search across for maybe more specific sub-sets of the data they're interested in. So that's another thing that we'd like to do in this year.

In order to provide granular access to different datasets and make them cross-searchable, engineers need to organize the data in a relational database, which is a collections of tables (formally known as *relations*) containing data belonging to different overarching "kinds" of things. Tables contain rows (*tuples*) and columns (*attributes*), where rows represent individual data points, and columns the properties of the data points. Different tables are linked to each other by different sets of relations that are expressed in data definition language, a process called *decomposition* or *normalization*. In this way, each data item has a specific place in a data model.

One of the engineers described as follows what he thinks the function of the DataFace data model:

Andrea: Data model is... It will be kind of like all the different elements of the data and the relationship between the different elements of the data. So, how you organize the data. Let's say, I wanna have one piece of information that tells me everything about the user, one piece of information that tells me everything about the projects that the user work, one piece of information that tell me everything about the files that this user works for this project. So, the data model is how you organize these inside of the database and the relationship between these different pieces of information.

Data organized in a data model can be searched from different points of view:

Andrea: We have data in mouse, silverfish and humans. Then the data can also be looked... So that's if we look at the data by organism. Then the data can also be looked by different type of assays that are performed for samples, in each of these organism. For example, we have genotype, phenotype association type of data. We have data for doing expression profiling, enhanced identification. We have... And then we can also look at the data by the type of actually, experiment that they're conducting. In that we can separate the data in... Or sometimes we kinda tend to see at the data whether it's an imaging type of data or

whether it's a kind of by informatics type of data when they do sequence analysis.

Metadata and ontology work

In a data model, datasets are described and linked together using metadata and ontologies. Metadata and ontologies play quiet distinct roles in the organization of biomedical knowledge in database structures. Biologists use metadata to assess the quality and relevance of data collected by others (P. Edwards, Mayernik, Batcheller, Bowker, & Borgman, 2011). DataFace metadata provide information about the experimental settings of the data collection (location, conditions, durations, grow stages, etc.), the type of data (images, sequence data, etc.), the file formats of the data (CT scans, BAM files, etc.), and the instruments used to collect the data (e.g., lasers, scanners, tomography imaging). As discussed in the literature review section, Leonelli argued that, in biology, metadata enable data to move from context to context, to became to some extent "context independent" (Leonelli, 2016). In the previous sections of this chapter, I have shown how, in the context of DataFace collaboration, database engineers and bioinformaticians collaborated to select the "proper" metadata that can be used to describe the DataFace datasets. In the following extract, Andrea, a database engineer, talks about his collaboration with the bioinformaticians to compile the "metadata spreadsheet:"

Andrea: They [the bioinformaticians] had to give me all the information that would... Again, that would make this... That would characterize this data. That would describe how this data, what this data represents, what is contained, what type of data set, what kind of data types, what kind of genes are they studying, what kind of type of study it is. Everything that we make.

[…]

My main interaction with the bioinformatician was to try to, in the first stage, was to try to understand the data that we're trying to submit, and that help us define the list of metadata fields that we needed to collect for that type of data, and the result of that has been the metadata spreadsheets that generated in order to collect their data sets, which there are... We already have it in certain form, and they're evolving because sometimes we find that we need to add this or remove that. That's been the main information, so the main interaction with them. And then on a normal basis, I interact with them when they need to submit the data. If they need to submit a new data set, I direct them to the latest version of the metadata spreadsheets, they send me, I look at them, I review them, I tell them... I ask for information, "Why is this field not complete? Why do you need, not putting this?"

In biology, data ontologies play an essential role in advancing research. Ontologies are composed of two main parts: terminologies and relations. Standardized terminologies are used to specify, in a standardized way, information about the data, such as the species involved in the experiments (humans, mice, chimpanzees, zebrafish), and the anatomical regions that have been tested. Like metadata, and ontological terminologies, relations are also "data about data," but their function is to link a particular piece of data semantically and logically to other pieces of data that were already catalogued. Ontologies are means of positioning new datasets in pre-existing knowledge organizational schemas. It is by ontology work that "experimental data" gain the status of "accepted knowledge." This classification work is essential for knowledge validation and transfer in biology research. The process of sequence annotation previously described, which is the core of data analysis in genomics research, completely relies on the possibility of locating information about a piece of biological data, in relation to all other known biological data. The integrated databases of bio-data used for annotation are built on ontological relationships.

The bioinformaticians also helped with the selection of the proper ontological terminologies to name DataFace data. The process of "naming the data" with ontologies was more challenging than choosing the right metadata. Biologists use many kinds of ontologies in their work, as this ontology developer points out:

Mario: The BFO, is the Basic formal ontology, and is the parent ontology of all ontologies. Then, there are ontologies for different domains, you have anatomy ontology, you have physiology ontology, you have pathology ontology. Those are different kinds of ontologies. Ours is anatomy. So there are species-specific ontologies, there's the human anatomy ontology, there's mouse anatomy ontology, there's zebrafish, there's frog, there's worm, there's fly, all kinds of species related ontologies. There are ontologies on drugs, there are ontologies for protein, then you have the gene ontology. So, all kinds of ontologies you will find out there that's related to the different biomedical domains, different species, different procedures, even laboratory procedures, laboratory results procedures, there are ontologies being developed for those.

For example, the community had to decide how to name similar genes that appear in multiple model organisms. Multiple naming structures exist for these similar genes; specialized thesauri can be used to name these genes in different ways, depending on the species. At the same time, overarching ontologies (also called "bridging ontologies") can be used to convert the specialized names in a common umbrella term that can be used to identify a group of similar genes from different species under one single name. In this process, the database engineers were very careful not to develop new naming systems, they relied instead on the bioinformaticians to get guidance on using existing, established, and authoritative sources.

Richard, the database engineer, explains that the best ontological term is the most

widely accepted:

Richard: […] our method for picking which ontology to use, is essentially to look for external, established ones that are considered the most authoritative. Picking one in the end is probably somewhat subjective, but you look for the one that is gonna have the most adoption and be the one that's most broadly accepted. And so that's what we try to do. And so that has meant things like Jackson Lab, ZFIN, for Zebrafish, OMIM codes for human...

Naming new data

One of the challenges that emerged with describing phenotypes with known ontologies (e.g., OMIM codes for human phenotypes, which provides codes for diagnosis), is that some phenotypes studied by the DataFace community were not included in these established ontologies. These anatomical structures or developmental processes were not present in these standardized ontologies because they were just being discovered by the DataFace scientists. In order to solve this issue, the database engineers consult "ontologies experts" to develop new names for these novel phenotypes. In the following extract, Richard explains that whenever the database engineers find a piece of unknown data, they need to "push that back upstream" to the ontology experts.

Richard: So, Jackson Lab may have anatomical terms... Or I think they're the one that runs the Mouse Atlas and they have two different versions of this Mouse Atlas, or an anatomical ontology. So while we might wanna use one, they may not have terms to cover the many bones and sutures of formations and structures of really specialized craniofacial, especially abnormal development. They may not have that term. And so we are always running into situations where there's a term that's not covered. And that came up at our meeting […] we run into situations where there is a term that does not exist. And then what we wanna do, is push that back upstream to the teams that are running those ontologies.

Examples of well-established teams of "ontology experts" include the information professionals at the Ontology for Biomedical Investigations (OBI) initiative (for animal images data), the Gene Ontology (GO) (for gene expression data), and the Protein Ontology (for proteins). In particular, the OBI ontology defines more than 2500 terms for research assays, devices, and objectives. Richard describes further how the process of going upstream to name new pieces of data works in practices.

Richard: Each of them will have forms to say "submit a new term." I mean in general, I think that's a general practice that most of them have, 'cause they wanna encourage people to submit terms to them. So there will be either somebody you email, somebody you know or generally maybe something on their website, that let's you go and contact them or a form that you can fill out to say, "I have this new term or this new code, and here's a description." They have something to fill out and you submit it to them and they fit it into their ontology or maybe come back to you. So basically, that's what I would call going back upstream. We're the downstream users of their ontologies, but if we had a term that's... If we had something that needs to be named and there's no name for it, at least in their ontology, then we would go back to them and submit it back upstream to them. And then, it gets included in the ontology and then we can get an official term from them and an official identifier for that item and include it in our database. That's kind of the process there for working with those other ontologies, and how we deal with naming something.

However, it turned out that the scientists who collected the data in the first place sometimes disapproved the terminologies suggested by the ontology experts. As Richard points out in the interview extract below, there seems to be a disconnect between how the ontology experts suggest to name the data, and how the scientists would name them.

Richard: Building up these ontologies is itself an output of the scientific process. So some of the tension that happens too, is that these ontologies are developed by very well-meaning and often very bright informaticists or computer scientists, but they're not necessarily the bench scientists. And you'll have somebody who's the biologist, who's actually a geneticist or anatomist or whatever they may be, and they'll say, "But this term does not really describe it right." Or, "This is kind of a weird term." Or, "I have no idea what this means." And so despite the many terms that they have to choose from, they do not necessarily always find one that they think is right. And so that's an interesting problem 'cause we do not necessarily have those biologists building up those terms.

Richard: And I think that's one of the things that we've observed from running the DataFace consortium, is that you kinda see in practice how at times there's really this disconnect between the ontologist and the actual, what I would call really, the domain scientist, or let's just call 'em the biologists. And a lot of people who do those, have come from a biology background. So, we just try to do the best we can with accommodating them and encourage them to use the terms that are in the control vocabularies and then the process is, if we find something where we do not have a term... Because craniofacial research is a specialized area.

The "ontology spoke" and the human anatomy data

The DataFace ontology spoke is in charge of supporting the development and selection of ontologies of imaging data related to anatomical parts in humans. Specifically, the spoke provided consultation for:

- Identifying standardized terminology for human phenotypes;
- Mapping species-specific ontologies (translational research);
- Annotate human malformation genotypes and phenotypes.

Mario is the principal investigator of the ontology spoke. Mario is a senior figure in

the human-anatomy ontology community. In collaboration with his colleagues, Mario designed in the late 90s the first version of the "Foundational Model Anatomy" (FMA), which organized for the first time human anatomy images in semantic relations. The FMA contains over 180.000 terms organized in over 104 over-arching classes and 150 different types of relationships, "from the whole body to the smallest structure, the macromolecules." Before the FMA was developed, human anatomical images data were named using controlled vocabularies provided by the National Library of Medicine (NLM). Mario and colleagues had the idea of organizing anatomical terminologies used to describe images semantically, linking terms to each other. The FMA was also different from the previous systems of classifying anatomical images because it organized ontological relations of anatomical parts not based on shared attributes (as it was done before), but bases on structures and physical properties. So, for example, if before hearth would have been classified as "a cavitated organ," in Mario's ontology would be classified as a "pumping organ." Mario and the team called themselves the "structural informatics group" (SIG).

During an interview, Mario highlighted that is very important to design ontology relations based on one coherent principle, and not more than one. He refers to this process as "putting data in proper boxes."

Mario: We were able to put them [terms] in their what we call proper boxes. So it's easy to now understand what the heart is, what the cavity of the heart is, what is the surface of the heart. They're not all the same thing, they are different things. But there's a relationship between them. The surface of the heart bounds the heart, and the lines bound surfaces and points bound lines. So those are the relationships. So,

we have now the ontology built based only on structure, nothing else. We do not base it on physiology or function. In building an ontology, it's necessary to be able to have a good ontology that is logical, that is clean, that is organized. You have to base it only on a single context. In this case, just structural, physical properties. No function, no pathology or any other context.

The SIG team developed over 150 hierarchical relationships based on biological structures:

Mario: I will send you the publication on that. Okay. So the 150 relationship meaning, by structural relationship, I meant part, regional part, constitutional part, membership, systemic part, adjacency relationship, connectivity relationship, developmental relationship, the heart developed from this and this developed from that or this gave rise to that. So those are examples and there's more than 150 relationships. Now, the 104 classes are related to each other, by more than 150 relationships or properties resulting in about 2 million relationships between... There are more than 2 million relationships between the different classes, in the last 20 years that we have worked on the FMA.

To create the FMA ontology, Mario and colleagues followed the best practices for building ontology promoted by the Open Biological Ontology (OBO), which advice to build ontologies using a single inheritance principle. Mario is convinced that the fact that the FMA was organized around one single principle (structure) contributed to its wide scale adoption.

Mario: […] And so, while we were developing the Foundational Model of Anatomy there were also other ontologies being built in anatomy, but they did not follow the best practices in building proper ontology. This is based on that group, the OBO. And so, over time we became the only big anatomy ontology in the world.

Another challenge that the FMA designers had to deal with was the normalization of redundancy among terms.

Mario: So I give you example, the thumb. Okay? In some institutions, they refer to it as the first digit of the hand. This is the first digit. So let's say, in one study, in one group, they're studying injury of the thumb. They would describe it as injury of the thumb but another institution would record it more formally as injury of the first digit of the hand. If you do not associate the two, you would not know that they're one and the same. So that's what happens in anatomy. This is very common in anatomy. People either come up with their own terms in describing, especially if it's a newly discovered segmentation of a structure. What I mean by that is that the more people develop different modalities, say either in imaging or in diagnosing, they started subdividing parts and the newly subdivided parts do not have new terms and they create their own terms. And what happens is, one institution may call it one thing, the other institution may call it another thing. Or even worse, one may call it one thing to refer another thing, use the same term but refer to a different structure. So those are the kinds of things that we want to, what we call accommodate and reconcile in the FMA.

For the DataFace project, the SIG updated a section of the FMA dedicated to craniofacial regions. They called it the Ontology of Craniofacial Development and Malformation (OCDM). The new OCDM not only covered human phenotypes, but it also expanded to map human phenotypical ontological terms to mouse and zebrafish phenotypes. The Jackson lab provided terminologies for the mouse phenotypes, while the ones for the zebrafish were downloaded from the ZFIN platform.

Below, Mario explains the process of mapping ontologies from different species

(identification of evolutionary homology).

Mario: So the mapping, a lot of it is based on what you call the species evolution approach, meaning, for the development species, this particular structure became the upper limb for the mouse and then it became also the upper limb for the human. So in that case, we would say that, that's a homology between the upper limb of the mouse and the upper limb of the human. And what about with the fish, is it the fin? Because the fin is actually the homology of the arm for both the mouse and the human. So, the mappings that we have for the mouse and the human is more extensive than the human and the zebrafish and that's because of the information that is available. It's actually very difficult to find information, and the other thing is we need to have verification from the domain experts, and sometimes it's also hard to get the verification from domain experts.

The SGI group first mapped human ontologies (A) to mouse ontologies (B), and human ontologies (A) to zebrafish ontologies (C), separately. The team is now in the process of evaluating whether it is doable to map mouse (B) and zebrafish (C) ontologies by inference (if A = B and A = C, then B = C). One of the problem of this translation work is that certain human phenotypes do not exist in other species, such as the mouse lacks the uvula and the philtrum, and vice versa, certain anatomical features in the mouse are missing in humans.

Mario describes how different expertise from multiple teams are involved in the process of ontology mapping:

Mario: So the framework, I'm using a template. A template that we have developed from the human. And then that template is being used for the mouse and then being used for the zebrafish but also being adjusted based on certain requirements that are different for mouse and different for zebrafish. And for that part, I can't do that. I have to go to the mouse and the zebrafish expert to give me the proper guidance in building that. And many of the contents will come from them because I'm not... I do not know... I'm not a

mouse or a zebrafish expert.

The ontology spoke also helped with annotating genotypes and phenotypes related to malformation. As Mario explained, the community is very much interested in this activity, however, it is also the most challenging one. The construction of ontologies for malformation data is a complex process. For example, the anatomical ontology for the Apert syndrome would be based on a set of "observed abnormalities," to use Mario espression, in humans, such as wide nose, small head, small ears, protruding jaw. This set of features that characterize the Apert syndrome need to be related to specific genotypes that explain how these structures came to be. However, each feature can be related to multiple genotypes, and one genotype can be related to multiple features. For each feature, the ontology needs to specify which genes are mutated, and also at what developmental stage they mutated. The same need to be done for protein mutations. Ideally, the goal of the OCDM is be to map all the malformation phenotypes with all the genes and proteins mutations across species in a single knowledge representation schema. Most of the information needed to develop such a schema already exist, but it is scattered in multiple databases. Mario refers to "data discovery" the process of finding links across different pieces of information and organize them in an ontology of relations. This is what labs members referred to as data annotation process.

Mario: […] data discovery is... You're now able to access, be able to get to the information out there that already exist, and then you come to discover. You discover a new mechanism or you discover a new abnormality, or you discover a new pathway while you're navigating information and while you're looking for some data integration. That's the part that's very exciting for the DataFace consortium. And that's the part that we're working on in a very early stage.

Below, Mario explains in details how the OCDM relational ontology can be used to annotate data for the Apert syndrome.

Mario: Apert has component phenotype, broad nose, high palette, and so forth and so on. You will see. Has genotypic abnormality.

Mario: Okay. So, the phenotype are the things that you can see. Broad nose, high palette, hypoplastic midface and so forth. Genotypic is the abnormalities in the genes, so that means the Fibroblast Growth Factor Receptor 2 gene is abnormal. If I go, say, to broad nose, if I go to broad nose, it takes me to, "Broad nose is a pathological nose."

[...]

Okay, so now when I go to the genotype, Fibroblast Growth Factor Receptor 2 abnormality, it's otherwise known as FGFR2. So the Fibroblast Growth Factor Receptor 2 abnormality is a fibroblast abnormality, is a autosomal dominant abnormality, is a genotypic abnormality. It tells you what type it is, but that's not it. So if I go to this class, it would say that it is a genotypic abnormality observed in Apert syndrome. It would also say that it's a genotypic abnormality observed in Beare-Stevenson cutis gyrata syndrome.

Mario: It would also genotypic abnormality in Crouzon syndrome and the protein involved is and I give the URL. This [1:02:04] ____ an external ontology which is the protein ontology, and the protein ontology will now tell you what are the things associated with that. See how you can navigate information from one to another to the protein? These are the kinds of abnormality, there's a mutation in the FGFR2. So if I go to Beare-Stevenson syndrome... So there's three right now they've identified associated with that. At least that's important for the DataFace group. If I go to Crouzon syndrome, now it would also tell me what are the phenotypic abnormalities for Crouzon syndrome. And I say mandibular prognathia.

Mario: If I go to prognathia, it would say it's a prognathism, it is abnormality of the jaw and that it is observed in Crouzon syndrome. And it's a type of prognathism. So there's a lot of things that can be

developed, all this relationship. For example, when we go now to this gene and its URL, and if they have used this as the annotation for some other disease in another institution, another study, then we can access that through that relationship.

Mario started working on the annotation of malformation data in July 2017 and he will be presenting the result to the community in the next year or so. A list of the non-DataFace related projects that adopted the generic FMA ontology developed by Mario and his colleagues is available online.

Mario: So once we have finished this, we will present this to the DataFace experts and say, "Okay, this is what we have now, how do you want to use it and what else do you need us to put in there?" You have pictures, associate those pictures with any of these terms. If you label your pictures with this, then this will already help you navigate all other attributes or properties associated with that particular label.

By analyzing how DataFace participants built tools for data access, a diversity of practices for data curation and integration emerged. First, database engineers, in collaboration with the bioinformaticians, developed metadata schemas to curate the data right after data collection (curation "before the fact"). This first level of curation helps scientists to gain information on the quality and relevance of others' data. Database engineers also collaborated with the ontology experts to properly "name the data." By doing so they were able to bring multiple terms that refer to the same biology entity together, and standardize the naming for novel biological entities. Overall, the use of metadata and ontology terms enabled database engineers to build a cross-search data browser that access data at different levels of granularity, and avoid the "data dump" effect. DataFace data are integrated at the search and browsing level. A next step would

be to organize the data using ontological relations such as those envisioned by Mario and his team.

Tools for data discovery and visualization

Scientists cannot possibly be aware of all the datasets hosted in the DataFace platform. The collaboration developed some tools to help scientists to gain an understanding of what is inside the DataFace's repository, without having to access or download datasets one by one. These are referred to as "tools for data discovery." The engineering hub developed some of these tools, while others were developed by the science spokes. For example, the "Mouse Matrix" was developed by the engineering hub and is available on the homepage of DataFace's website. The Mouse Matrix allows scientists to get a glance of all the experimental data related to mouse that have been made available by the DF participants and are hosted on the DataFace repository. The datasets are organized along two axes: mouse age (y) and anatomical region (y). Intersections are color-coded in relation to the type of experiments conducted. By clicking on an intersection, the researcher can visualize a list of datasets available for download. For example, if a researcher is interested in knowing what datasets are available for the mandible at the E12.5 stage, a pop-up window will show all datasets available for download, in this case five datasets.

Participating spokes developed multiple tools for data discovery, in collaboration or independently from the DataFace engineering hub. Some of these tools are hosted on individual labs' servers, and accessed through the DataFace main platform. Some others

are hosted on the DataFace's servers. Some tools combine anatomical images and scan of humans or animal models, so that the user can explore known facial features in an integrated way. An example is the "CranioGui Virtual Machine," which is available for download on the DataFace website. A list of all the data discovery tools developed by the DataFace Consortium spokes is available online on the DataFace website (DataFace, 2017).

In the interview extract below, Christopher (a database engineer) explains that data discovery tools developed by the spokes independently from the engineering hub were treated like "black boxes." Engineers linked to the main DataFace platform, but they did not operate or modified them.

Christopher: There are some things that DataFace considers applications. I do not have much familiarity with it, there's some little calculators and things that are out there, and there might be some client side applications that I think we still distribute, you know, it's developed by one group and then we allow that. We try to provide a place where they can post it for download, but it's essentially not a website function. It's just... It's like a local imagery or tool that you could... You could download the dataset that's in DataFace and then run this tool on it, so it's kind of a black box from our point of view.

One of these tools, the "Human Genetic Analysis Interface," was developed by the Pink spoke (also referred to as the "software spoke"), and it is currently released in beta version. The interface allows researchers to visualize theirs' and others' genomics data at a summary-level (i.e., p-values) in a semi-automated way. In craniofacial GWAS, p-values represent an estimate of the probability of certain facial traits to be associated with certain genes (Qu, Tien, & Polychronakos, 2010). Normally, p-values for GWAS studies are generated by using a command-line interface called PLINK. By functioning as a

graphic user interface (GUI) to the command-line software PLINK, the Human Genomics Interface automates the visualization of p-values.

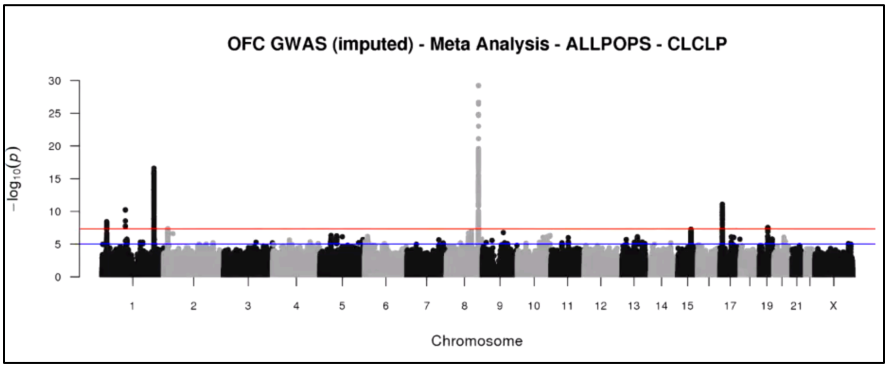


Figure 8: An example of summary-level visualization of genomics data, in the "Manhattan Plot" format. Source: the DataFace website.

Visualizing summary-level data is a way to gain important information about others' data without having to directly access, download, and process others' "raw" datasets. Automating the visualization of p-values enables researchers who have low computational skills to access this type of information independently, as opposed to send the data to a computation lab. But it also saves some time to the computational biologists and statisticians who want to run secondary analyses. By using the interface, computational biologists can visualize p-values to verify whether the datasets are worth downloading and re-analyzing. In the interview extract below, Emanuela, who is one of the human geneticists involved in the design of the platform, describes the goal of the interface:

Emanuela: The tool serves a couple of purposes. For the people who did the study [i.e., who collected the data], it's an easy way to look up the results without having to write a couple of lines of code to pull data out of a text file. You can just put the region in that you're thinking of and show the results. For people who are more animal model collaborators, they often will contact us and say, "I study this gene. I think it

causes clefting. Is it associated with clefting?" And then we have to then pull the results out of the database and what this tool allows them to do is to look that up themselves and then see if there's a result.

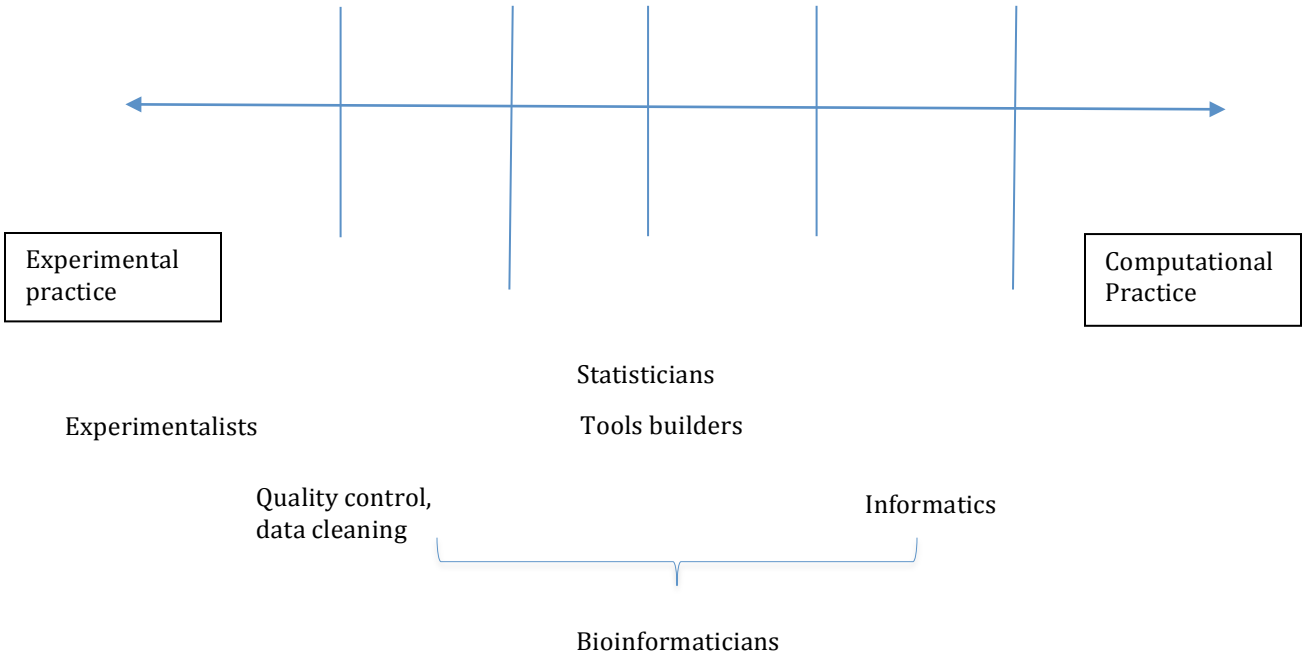
Emanuela: And if they still need help interpreting that p-value, then we can help. But it lets them do it for themselves without constantly having to go to somebody else and then wait for us to do it, it's immediately available. And then it also lets people who might be statisticians and statistical geneticists who do not wanna download the data or go through that process, see the results, and decide if they want to go through that process or see if the question that they're interested in could be answered by getting the data or collaborating with the group that has the data.

Emanuela: It's mostly a visualization tool. It also saves you a step. […] It saves that step of providing that visualization that someone would do anyway. […] PLINK (software) runs on the command-line and it operates with a series of flags. You load your file and say what parameters you want to use for the analysis and it can do most large-scale GWAS analyses in a straightforward and fairly user-friendly way to do it. We envisioned the tool being basically an interface to PLINK so that people can run their own analyses. As easy is PLINK is to use, it's just as easy to use it incorrectly, so PLINK will not tell you, "You have trios. You shouldn't run a case-control analysis." It'll let you run a case-control analysis on trios. That's not the analysis you should be doing."

The Human Genomics Analysis Interface enables scientists to gain relevant knowledge about others' data without having to directly access, download, and manually process individual-level datasets. Accessing summary level data, instead of downloading and re-processing raw data, is particularly useful in the context of GWAS studies. Accessing information about genotypes and phenotypes data from multiple GWAS studies help craniofacial researchers to establish meaningful relations between genes and facial shapes. Some GWAS are conducted by comparing two or more populations, some within one population, some on affected patients, and some, like the one conducted by

the Pink /software spoke, on general population. By visualizing p-values of genotypic datasets collected during any GWAS study, the Human Genomics Analysis Interface allows the possibility of integrating of knowledge across multiple GWAS studies.

Disciplinary configurations: craniofacial research as team science



DataFace's teams are highly interdisciplinary. I identified three overarching disciplinary configurations that represent the DataFace participants' backgrounds and specializations. These are (1) the experimentalists, (2) the bioinformaticians, and (3) the informaticians. The disciplinary boundaries between the experimentalists, the bioinformaticians, and the informaticians are highly unstable. Some individuals belong to one configuration, while others have mixed educational backgrounds and training that allow them to practice in more than one configuration. At the same time, each configuration holds a defined set of skills and research objectives that shape the way in which individuals belonging to a specific configuration work with data, and think about data reuse.

I use the expression "the experimentalists" to refer to those individuals who are primary responsible for carrying out the lab experiments. Experimentalists have an educational background in developmental biology, molecular biology, or human genetics. For example, Kristina at the Blue Spoke, and Jane and Hazel at the Green spoke conduct the experiments at these specific sites. In the labs, the experimentalists are heavily involved in the collection of the sample data (e.g., tissues or blood from patients and animals), in the data production (e.g., DNA extraction, collection of imaging data), and also in the contextualization of the preliminary findings obtained by the statistical analysis (literature review and data annotation). These individuals are normally highly specialized in the investigation of a restrict set of biological phenomena, like a sub-set of genes involved in oral clefting, in the case of Kristina, or the developmental process of an embryonic tissue, in the case of Jane. Kristina and Jane are the experts of these biological entities and processes. Traditionally, "wet lab" biologists are portrayed as those who manage bench experiments and handle biological material "in vivo" and "in vitro," as opposed to computer analysts and statisticians who conduct data analysis "in silico." Far beyond simply handling biological material, the experimentalists I interviewed and observed actively contribute to the knowledge production process with their specialized knowledge. For the experimentalists, the results obtained from statistical analyses are indications that something interesting might be happening, but they cannot be sure until they review all the relevant literature on the subject. Experimentalists are in charge of separating the noise from the signal by functionally validating statistical analyses and find proper explanations of why biological entities behave the way they behave.

The experimentalists who work in clinical settings can collect and access human sequence and imaging data. Those specialized in model organism research cannot directly access or use human data. This is because experimentalists who work in clinical settings normally have also a Medical Doctorate, which allows them to have access to patients and their data. When animal model experimentalists need to access human data, they can do so by obtaining an IRB approval, or by collaborating with a human geneticist who is already clear to obtain the data. For this reason, DataFace's mix teams of animal and human geneticists gave the opportunity to animal models specialists like Kristina to have access to human data. As I have showed, for Kristina this was the main advantage of being part of the DataFace collaboration.

Among the DataFace participants, a distinct group of individuals self-identify as "the bioinformaticians." These individuals have a background in computational biology and bioinformatics, sometimes in addition to a traditional biology degree, but not necessary. Daniela at the Blue spoke, Jane at the Green spoke, Emanuela and Julieta at the Pink spoke, serve as bioinformaticians in their labs, although in very different ways. I identified at least three different ways, or sub-configurations, in which bioinformaticians practically contribute to knowledge production in the DataFace collaboration.

The first sub-configuration is "bioinformatics as a service." In the Green spoke, Kaine provides a bioinformatics service to the lab researchers when he obtains the sequence data from the sequence facility, processes it for quality control, and returns it to the lab

for data analysis. Each DataFace spoke appointed a person to conduct this type of work. In few cases, this person would be a stable member of the lab, such as for Julieta at the Pink spoke. Often this person would be someone working in a near-by facility or lab, like in the case of Kaine at the Green spoke and Daniela at the Blue spoke. By talking with the bioinformaticians involved in the DataFace collaboration (N=9), it emerged that is quite common for an individual with bioinformatics expertise to be asked to perform sequence data quality checks for researchers working on different labs. Sometimes credit for this kind of work is given by including the bioinformaticians as authors in the related paper, or by sharing grant resources.

The second sub-configuration is "bioinformaticians as statisticians." These individuals possess the statistical knowledge to run the analysis (e.g., gene mapping) of the "raw" sequence data. As I showed in section 3 and 4, this analysis is conducted using ad-hoc software pipelines that rely on the access to a set of in-house, as well as publicly available, databases (e.g., OMIM, GEO), analytical and visualization tools (e.g., the Genome Browser), as software packages (e.g., Python, R). These analysts possess statistical skills and are well versed in the biology domain. This mixed background allows them to run statistical analysis "in silico," and to come up with hypotheses of gene/protein functions and regulatory processes (e.g., Daniela's gene list).

Finally, bioinformaticians can also be "software and tool builders." Builders develop all those pipeline components that are used by the analysts to perform statistical analyses. These include algorithms for the analysis of different data types (e.g., RNA-seq, Chip-

Seq, and GWAS), software packages for data analysis (e.g., Package Gene Set Analysis, GSAR, in R), and tools for data visualization (e.g., the Genome Browser). The Human Genomics Analysis Interface developed by the Pink spoke is also one example of these tools. As we have seen, the builders working at the Pink 's spoke designed the tool under the supervision of Emanuela, who has a mixed background of computational and experimental biology, but she is not a builder herself. The Pink spoke was the only spoke with builders in their team. As a matter of fact, the consortium didn't explicitly fund the development of tools for data analysis, which, as we have already seen and I will soon discuss further, are central in the process of data reuse and analysis.

As it should be clear by now, individual researchers belong to more than one disciplinary configuration. Daniela runs statistical analysis for the Blue spoke, while Jane does it for the Green spoke, and Julieta for the Pink spoke. However, Daniela also takes care of the data quality for the Blue spoke, while Kaine does the quality control for the Greens poke. Jane at the Green spoke conducts experiments and does also the analysis, but she does not do quality control. Julieta is an analyst, but does not do quality control either. Emanuela helps with setting up the experiments and performs data analysis, and she also supervises the development of tools for data analysis.

The last disciplinary configuration is "the informaticians," or the informatics experts. The informaticians are experts in knowledge organization, which means in data modeling, management, and curation. Both database engineers and ontology developers re informatics experts: they do not participate in the experimental procedures, or in the

design of the research studies, and they do not develop or use tools for statistical data analysis. Database engineers contribute to the organization of DataFace data in database structures that enable researchers to archive, retrieve, and cross-search different datasets. They use metadata to de-contextualize datasets, make them as much as possible "context dependent." Ontology experts apply standardized terminologies and ontological relations to the datasets to allow scientists to re-contextualize the data in the larger schema of biology knowledge representation. One significant aspect is that, even if their roles are complementary, in the DataFace project database engineers and ontology experts hold different educational backgrounds and possess different sets of skills. Database engineers are trained in computer science and are specialized in data modeling, they do not have a biology or life science background. On the contrary, ontology experts do not have a traditional engineering background, but they hold biology and medical degrees, and they learned "ontology work" during their careers as biomedical professionals and educators. Merino, for example, developed the "Foundational Model Anatomy" mainly as a tool for teaching anatomy to medical students.

Among all my interviewees, I did not find one individual that was at the same time an experimentalist, a bioinformatician (in any sub-configuration), and an informatics expert. However, I did find that most participants had some combination of the above expertise. In particular, most researchers conduct a mix of experimental and data analysis practices. I also observed that, among the DataFace participants, the database engineering and the ontology work are the most isolated and specialized practices. In other words, while most experimentalists conduct data analysis (either statistical analysis or data annotation), I did

not find experimentalists practicing data management or curation beyond the consortium explicit requirements. This is true also the other way around. During our conversations, DF's database engineers and ontology experts stressed the fact that they are not involved in experimental practices or data analyses.

Craniofacial researchers and data reuse

At the time of my data collection, the DataFace participants either just deposited – or were in the process of depositing – the data that they collected. For this reason, our conversations on reusing *others'* data for knowledge production did not solely focused on the datasets collected in the context of the DataFace project. I instead centered my analysis on their "routine" data reuse practices. I was particularly interested in identifying a typology of ways in which scientists reuse open data that they did not collect themselves. Findings in this section are based on the examination of interview and observation data from six participating spokes, and of the papers published by the participants in the study.

My conversations with the scientists really centered on one simple question: *Can you show me how you reuse open research data that you did not collect yourself in your research process?* And I would clarify: *I'm particularly interested in understanding how you reuse data that you did not collect yourself – or your lab – to produce novel knowledge.*

Some scientists would immediately answer something on the line of "I have never reused others' data." Others would say that they reuse others' data daily. After talking to

few scientists, I realized that when I was asking about "reusing others' data," the scientists would understand "reuse" in different ways. Based on this initial observation, I started to ask more detailed questions and, soon, a quite clear pattern begin to emerge. Scientists would think about reuse as two distinct but also intertwined practices. When asked about reusing others' data, some researchers thought I was asking about accessing others' data at a summary level and through data visualizations tools. Others would think of accessing others' data at a "raw" level in order to run secondary statistical analyses. The more I would ask questions about the challenges of reusing data in either ways, the more it would emerge that such reuse practices come with specific and distinct sets of socio-technical concerns. In what follows, I characterize how and why – from a socio-technical point of view – re-using others' data by accessing them at a summary and integrated level is different from re-using others' data by accessing these as a "raw" level to run a secondary statistical analysis.

The fact that participants reported very different ideas of how data could be reused for, and by whom, should not come as a surprise. Science data are often reused in unpredictable ways (Baker et al., 2015). Given the fast-evolving and deeply interdisciplinary nature of contemporary scientific research, it is often hard to predict future reuses of scientific datasets. As we have seen, engineers worked hard to accommodate the needs of different types of users by enabling cross-searches and developing data discovery tools. However, as I will show in this section and further argue in the discussion, different data reuse practices are only partially influenced by the technical skills of the data reusers, or by the granularity of the data curation. Issues of

trust and the need of collaborative analysis have a much bigger impact on data reuse practices. In other words, you can have the best database system and the most curated data, but nobody is going to reuse the datasets if the right socio-technical conditions are not in place.

Accessing and reusing others' data at a summary level

A daily practice. Most researchers reported to reuse others' data daily. Others' data are reused daily to set up experiments, to annotate raw sequences, and to interpret preliminary findings obtained from experimental practices and statistical analyses. I have already showed how Daniela, Jane, and Julieta extensively use databases and other tools for annotating preliminary results from sequence and genotypes analyses. Participants access others' data through a variety of digital tools and databases, most of which are freely available online. The Genome Browser, the OMIM database, the GEO database, and the Jackson lab mouse database are among the most popular databases used by the participants in their research routines. Most of these databases provide access to others' data at an integrated or summary level, and through easy-to-read visualizations, such as in the case of the Genome Browser (see research workflow #1). In the following interview extracts, Akiko, Daria, and Halo, who are also collecting genomics data for the DataFace project, discuss the tools that they use to access and visualize others' data for background purposes.

Akiko: I use this Gene Expression Omnibus. For example, what I am thinking right now is we found some interesting gene expression changed from our study, but I do not know if it's specific for the disease or it just happened in our sample. But if I find the others' data doing the similar study, I can search if this gene also differentially expressed in this type of, in other study.

Daria: It's absolutely fundamentally necessary to all research that we do to have the Genome Browser. We also will visualize our data this way too. The Genome Browser allows you to upload custom tracks and custom sessions and so we do a lot of that to be able to... When we map our data back on to the human genome or mouse genome being able to visualize it in this type of portal is really valuable. For NCBI I use a lot... PubMed just for literature, search. I also use OMIM a lot, and that is just trying to figure out gene function or genetic basis of various human diseases. We also use things... Some of these are integrated into the genome browser to like dbSNP and things like that, so we use them primarily through UCSC.

Halo: I think database for me is a place where if I'm looking for specific question like, let's say, I'm working with a margin of interest, we found mutation in the gene and we wanna look at the expression. And if there is any animal model for the gene, so the database basically makes it easier for you to see if there is any things which was done before, so you do not want to repeat what was done, basically. We use human genetic databases, databases to look at phenotype like OMIM. We use expression database in different animal models, where you look at the variation in the expression of your gene of interest so you will look at through this database. So, of course, there is database for mRNA, there is database for protein across different species. And we have database for the animal model so either mice or zebrafish, you can look at the data of any of this phenotype which was described before.

Databases as maps of small-facts. To use Halo's expression, researchers use these tools to look at what "was done before." On the Genome Browser, OMIM, and other NCBI resources, others' data – accessed through summaries and visualizations – perform as maps of "established knowledge," or as "small facts." Researchers use these maps of established "science facts" to navigate through their knowledge production journeys. These maps provide the researchers with hints on how to causally explain the function and behavior of certain biological entities. Researchers use data summaries and

visualizations from online databases as reference points, as means of guiding the design of their hypotheses, and the interpretation of their findings. Without these resources, the knowledge production process would not be possible.

Easy-to-use. Scientists also value online resources like the Genome Browser, OMIM, and GEO because of their practicality and their trustworthiness. For example, the GEO gene-level database allows users to search and visualize gene expression profiles relevant to the researchers' interests by simply entering appropriate keywords and phrases into the search string. Especially for the experimentalists, these platforms represent, first of all, an easy way to integrate others' knowledge, and validate their own knowledge. Whenever I asked the experimentalists how and to what goals they would reuse others' data, our conversations focused on the usefulness of these platforms, and on their desire to access more of them. Especially for the experimentalists, visualizations of summary-level data represent the main way in which others' data are useful to them. Experimentalists stressed the need to have platforms for data integration that are as easy as possible to use. Some experimentalists mentioned their desire to have access to a "dream database." On the dream database, one can visualize all the knowledge on a specific biological entity simply by typing the name of the entity (e.g., the name of a gene) in a search string.

Reusing others' data when data are "raw"

Most biology databases, including the Genome Browser, OMIM, and GEO, offer the option of downloading datasets at a "raw" level. Some databases, such as dbGaP, are primary composed of collections of "raw" genomics datasets. I use the expression "raw data" to refer to data released at minimal levels of processing. These include, for

example, lists of sequences in FASTA, FATSQ, SAM/BAM and SRA formats, lists of gene expressions in CVS format, and images and scans in OBJ formats (see table 1 below).

Table 1: File formats for low processed data. Source: NCBI.

FastA are text files containing multiple DNA* seqs each with some text, some part of the text might be a name.
FastQ files are like FastA, but they also have quality scores for each base of each seq, making them appropriate for reads from an Illumina machine (or other brands)
SAM holds an alignment of seqs w/qual scores against a template.
BAM is a compressed binary format for SAM, however it can also be unaligned in which case it's more like a compressed version of fastq.
SRA files are a common format used by the NCBI, EBI, and others for storing reads and read alignments.
OBJ is a geometry definition file format first developed by Wavefront Technologies for its Advanced Visualizer animation package. The file format is open and has been adopted by other 3D graphics application vendors.

As we have seen, the DataFace Consortium intentionally made available the datasets at a low level of processing. During my interviews and observations, I discussed with the scientists when, why, and under what conditions they decide to re-analyze someone else's raw datasets. This practice is not very common among the participants in this study, but I provide an example of a secondary analysis of a DataFace's imaging and sequencing dataset in the next section. As I will discuss, a team of physical anthropologists and computer engineers re-analyzed the Pink spoke's Caucasian GWAS dataset in a study

that aimed at prediction facial shapes from DNA samples. In what follows, I instead discuss why the participating scientists do not re-analyze others' raw data that often. When they do, it is in the context of a collaboration with the lab that created the data. None of the scientists I interviewed seemed to have downloaded someone else raw dataset to run a secondary data analysis autonomously. It emerged that participants tend to be skeptical of "big data" that were collected by labs they do not know personally, and that were made available prior to publication.

Big datasets and concerns over quality

Skepticism toward high-throughput "hypothesis free" datasets. Researchers reported to be fairly skeptical toward large-scale survey datasets collected by "strangers," such as the one deposited in dbGaP, or the raw gene expression data on GEO. As I discussed earlier, this community is at the verge of an important methodological transition. While craniofacial researchers have been trained to investigate one or a small set of genes at a time for a long time, they are now facing the challenge of collecting and analyzing thousands of genes and related phenotypical variations at once. Among participants, the practices of genomics research are not standardized yet, and researchers are in the process of figuring out where and how to find value in these large datasets. Several researchers seem to perceive "hypothesis free" data obtained from high-throughout technologies as inherently flawed, or as big collections of data that could include a signal, but could also be all noise. Below I report some conversations on this topic with three researchers working in three distinct DataFace participating labs.

Lisa: With high throughput technologies, instead of studying a gene vertically, like for example we do

on these projects where we look at PBX mutations and we made all these mouse models that carry various PBX mutant alleles. You look at all of the genes of the genome at once. […] Unfortunately, sometimes there are imperfections and errors and it's also the nature of the investigations. You cannot expect everything to be perfect when you analyze 3,500 mega bases of information in a genome, right, and you look at it all at once. Correct?

Sheryl: I think (the field) it's definitely becoming more computational. It probably depends what type of craniofacial research you're looking at. So, the human genetics for craniofacial research, I think, is going more and more towards whole genome sequencing. And so the challenge there is that you find lots and lots of rare variants, and so how do you interpret them? So yeah, just having all of this genomic data and what do you do with it? How do you figure out what is the true cause of X syndrome that you're studying or X disease or whatever it is.

[…] *So, five years ago when everyone was doing exome sequencing, and kind of grabbing the low-hanging fruit…looking at the easiest things to find. So, if you sequence a bunch of people with cleft palate or something like that, you'll find mutations that are within genes. So now, they're going back in, sequencing the entire genome of everyone who still was not solved by that initial survey. And so, our research will help in trying to interpret their results of these kinds of things. And so, further down the line, you might see eventually, like, "This is a regulatory region that affects this gene, that can be targeted with this drug or something." But that's super far away.*

[…] *In general, we're not looking at one specific gene, or one specific pathway. All of the assays that I do are unbiased. So just looking at active regions across the entire genome, or genes that are transcribed across the entire genome... So we end up sequencing the entire genome or the entire transcriptome and then using that data to go back and do some analyses in an unbiased way to see what interesting things pop out.*

Kristina: […] Sounds a little bit bad, but a lot of people like DataFace are generating data without a particular question and then they go, "Oh. Look at all this data we have. Let's ask a question." That's retrospective rather than prospective. So most of the studies that I do, you knock down a gene and you go

look at what happened. That is a prospective thing, you have a question, "What happens if I knock this gene down?" You knock it out and you go see. Versus what we're doing now is we are simply defining what the genetic construct is. I know that we have questions that are generating this but we're generating a database full of data that someone can then retrospectively ask questions about.

Concerns about reusing data shared prior to publication. Several data creators reported to be "not sure" about how other researchers could generate hypotheses from their "unpublished" data. In the following conversation, Brianna, who is a postdoc in a DataFace participant lab, brings up a set of concerns specifically related to the reuse of big datasets of raw sequence data in the context of the DataFace collaboration. These data, as the reader may recall, are released prior to publication and curated "before the fact." Brianna works in a traditional developmental biology lab, where most researchers are pure experimentalists and very few have computation biology skills. Brianna wonders how will other people be able to trust their data, given that these are not associated with peer-review publications. She wonders whether quantity will compensate for quality, as it was suggested to her by some colleagues. Brianna hypothesizes that maybe others could access their data and try to replicate them to verify their quality, but immediately reflects on the fact that this would constitute a waste of public money, in direct opposition with NIH motivations for open data (avoid duplication of efforts).

Irene: What impact do you see DataFace datasets having on craniofacial research or on your research, on your work?

Brianna: Well, this is difficult. We were talking, our group was talking yesterday about this that was bit hard for us to see how we even will be able to reuse our data or how it will be useful to us in the future, but

the hope would be, that some one could look at our data and say, "Oh, I'm interested in this gene and what this gene does and here are some people who have posted data related to this." And maybe I can get an idea for future research or something like this.

[…] I think it's interesting to think about what kind of relationship this sort of data has to more hypothesis driven projects which are the other things that we work on. Usually, we have a scientific question or an idea of how we think something works and then we go and try to test it. This is more generating data for the sake of having the data and it's a bit difficult for us to see how it could be useful because we are so used to thinking the other way.

I mean one thing we were talking about yesterday after this meeting about the XY was, "Well, why we should trust this data because it's not peer-reviewed, and we are just putting it on this website." So if you are some other researcher coming on the DataFace website, why would you trust the Lab unpublished data. And one thing that one of our researchers said is, "Well, if it's n equals three, maybe you do not trust it so much but if it's n equals 100, maybe then you trust it more. But we are not funded for n equals 100."

I do not know really where that leaves us, maybe trust but verify for yourself although I think the goal is really or the vision from the NIH is that you can really use this data in your publications not that you would say, "Oh, this is what they found now let's go and replicate it." And then use the new replicated data, the goal would be that you can actually use the DataFace data, I think. But then that's difficult because you really do have to trust what other researchers did.

Data creator and data reuser collaboration

Several scientists I talked with believed that the data creators themselves are those in the best position to re-analyze "raw" data – compared to scientists who were not involved in the data production process. For this reason, they reported to re-analyze others' raw data solely in the context of a collaboration with the data creators. Collaboration between

data creators seems to help the scientists to overcome skepticism toward large and unpublished raw data.

Data creators know the specialized literature. The experimentalists believe that their "raw" data need very specialized knowledge to be made sense of. While anyone can find patterns in the data, specialized and up-to-date knowledge of the literature related to that specific biology phenomenon being investigated is required to interpret the causal explanations behind statistical patterns. I have shown how Jane's specialized knowledge in developmental biology guides her through the interpretation of her transcriptome data (see the Green spoke's research workflow). Jane's data, without Jane's specialized knowledge of how tissues develop in "time and space" in the embryo, would be much harder to interpret. In the following conversation extract, Travis, Jane's colleague, evaluates the possibility for others to download and mine their RNA-seq data. As Travis points out, these RNA-seq datasets, which are collected at three precise developmental stages and in relation to specific embryonic tissues, need a lot of context to be re-purposed.

Travis: I do not see somebody going in there and... If people want to go into our data and do the type of analysis that we want to do, well, they can do that too. I think they would be at a disadvantage though, because they do not know exactly how it was... They do not know all the meta-analysis that's associated with it. They weren't involved necessarily in the design and the execution. So I think it's more difficult for people to get in there and make sense of this…

Skills integration. By collaborating with the data creators, specialized and tacit knowledge can be exchanged and integrated between craniofacial researchers working in different, but related, biological phenomena. Collaboration also allows for work re-distribution, and skills integration. Computational biologists access knowledge about the biological functions behind statistical results and low-hanging fruits, which helps them to design thorough hypotheses to be tested in the data. At the same time, experimentalists access statistical and computational skills necessary to analyze large-scale datasets.

Hank: I think collaboration is better because then it makes sure that everybody's on the same page and they know what's going on with the data and you can... The groups might have complimentary skills. So for example, for human data, we have our own GWAS and we participated in a different GWAS. That data is on dbGaP and we could just download the data from dbGaP but our preferred method is just to collaborate with the group that did that GWAS and work with them and then we're all co-authors on the paper and we share the results and we help each other on different analyses and we're always talking about who's doing which things and where are the priorities for different groups. So I prefer collaboration to competition or doing things independently.

Hank: Today's science is team science. We didn't necessarily answer one of your questions fully, is that, people who use big data and people who do not use big data and people that are going to use big data, they do not know how to use big data. So a lot of these guys (people interested in reusing big data from biomedical research) are MDs, they're surgeons, they do not know how to do what I do (data analysis). So, more and more, the projects are becoming team science where you have a bioinformatician of some sort, an informaticians of some sort, and a biologist, a bit scientist, a statistician... So data analysis and the generation of publishable work is getting more and more difficult as technology gets far and advanced, and to even all of these databases. So I know that the PIs could possibly extract data out of DataFace. Could they do it in an efficient amount of time? I do not know. Even if DataFace was well-documented, well-interfaced...there would still need to collaborate with someone like me […]

Data creators have access to all and up-to-date annotations. Data creators posess an "advantage" to re-analyze their own raw data – as opposed to external reusers –also because when they deposit raw data they do not share all the variables, annotation, and phenotypes related to the dataset. When researchers deposit genomics raw datasets, researchers would normally share a certain amount of information (i.e., annotations) that is needed to make sense of the raw data. They would share the "annotations" that are known at the time of the data collection. Let's take as an example a dataset of genotypes data from a craniofacial GWAS deposited on the dbGaP database. Let's supposed that a team finds indications in their data of a set of variables to be related to a certainly genotypic profile. When depositing a genotype dataset, the team would provide annotations and summary level data for those variables (e.g., total number of cavities). However, after some time analyzing the data, the data creators could find a better way to summarize or present this information, which could make the interpretation of the data more meaningful (e.g., surface area affected by cavities). Also, because data creators are not required to share all the annotation data related to a certain dataset, they sometimes retain the information that seem to be more valuable for a publication.

Data creators have special permission for human subject data. Another reason why secondary analyses tend to occur in collaboration with the data creators is related to the privacy and safety regulations on human genomics data. The data creators already own the permissions that are necessary to analyze the data. For example, because of privacy regulations, in order to access raw genotypes data on dbGaP, researchers need to be part

of a pre-approved IRB protocol, or to obtain a new one, which can take months. As I discussed earlier, gaining an IRB approval to access human data is particularly difficult, if not impossible, for researchers specialized in animal models research who do not have a degree as doctor of medicine (MD), and do not have access to computational infrastructures set-up for sensitive data.

In the interview extract below, Ella, a computational biologist and statisticians at the Pink spoke, talks about some of the challenges related to re-analyzing others' raw genotypes data from dbGaP that I mentioned so far:

Ella: Most genetic studies these days are done on big epidemiological studies that might collect hundreds of variables. The genotypes will come directly from the genotyping center into dbGaP, and they make you send your phenotypes at the same time. And usually you would sort of negotiate with NIH about which phenotypes you're gonna deposit. And it has to include the ones that are central to the study. If you're studying cleft lip and palate, you obviously put your clefting phenotypes in there. But if you also checked these people's height, and weight, and dental caries, and ear lobe shape, and hair length, and 100 other things...

Ella: You do not necessarily have to put everything in. And the other part of that is that, again; since you put those data in at the very beginning, you typically haven't done all the work yet to even figure out which of those are good variables or which ones you might want to transform or create a calculated variable with.

Ella: For example, when we do dental traits we'll look at... The dentists measure every surface of every tooth to see whether there's decay, so that's 120 of these little decay variables for each person's mouth. We've never put all 120 of those variables into dbGaP. We come up with some summary, like "total number of cavities in your mouth", and that's what we put in. But then two years later after we've been working with the data we might decide that a different summary is actually better, like "total number of surfaces

that have decay on them" or something, and then we're doing our analysis with that one but this older one is still in dbGaP

Ella: They encourage you to, but it's so hard that you have to be really committed to it... to initiate and complete a dbGaP submission on your own. I know it sounds crazy that it's that hard but it involves things like descriptions of the variables and documentation that you have the right consents for every sample that you've submitted. There's a lot of paperwork. It's for the bureaucracy that makes it so…

Ella: That's what makes it hard for, for example, mouse people. I have collaborators at Jackson Labs which is major, major, major mouse facility. They have a really hard time getting dbGaP data because they're not set up for the kind of security you need for human data; they do not have the right computer security setups, or even the right people to describe the computer security the way the dbGaP wants it. So they're very frustrated that... And because [36:15] ____ that they're not...

Attributing credit to the data creators

The choice of re-analyzing others' data in collaboration with the data creators is also related to issues of credit attribution and invisible labor. When collaborating to a secondary analysis, data creator and data reuser tend to co-author the resulting paper. All DataFace's datasets were physically collected by early career academics, namely master students, doctoral students, and post-doctoral students. Obviously, the principal investigators direct the research agenda, and are often responsible for choosing the overarching research questions, and to supervise the research design and experimental setting. However, the early career academics are those who invested time in collecting the data, and they are those who are in charge of conducting the analysis of the data, and publish academic papers out of it. I let Kristina speaks for herself about her relation with the data she collects:

Kristina: Any data is very precious for me because I spend much time to do these all experiments. Every data. If I have a good data, I'm super happy every time, which means if I do not have good data I'm so sad, and I'll be disappointed for a day.

Irene: Can you give an example what's good data for you?

Kristina: The good data is data what I expected or I have this expectation to get this data. So, when I inject it, I want to have mutation, but sometime if it does not work, it will not show the mutation.

Kristina has been a postdoctoral researcher for over seven years now. Her goal is to work as a researcher in the academia, but she says that she has no idea whether it's an actual option for her. Like many of the post-doctoral researchers I interviewed, Kristina is on "soft money" – her job depends on grant money - and she could potentially stay on soft money for an indefinite period of time. Getting "good data" represents her only chance to publish novel results on the genetic causes of craniofacial syndromes, and hopefully obtain a tenure track position. Kristina told me that she felt uncomfortable releasing sequence data right after data collection because she was not sure these data were worth depositing in the first place. She reported that her experiment was still in progress, and she wasn't sure that the experiment based on these data would show what it was supposed to show until her research design would be completed.

Irene: What kind of data would you feel uncomfortable to show to the public?

Kristina: If I'm working on it, it is still in the progress...the data, I want it (the data) to be complete when I show them to the public.

Irene: Okay. You want it to be complete. Why?

Kristina: Why? Because I will not be... There's no 100%. So we do a lot of troubleshoot, and so that's why I just wanted to make everything happens well. So for one experiment, we do... If this experiment is the first time we do, we have to do test experiment, and then we're gonna do real experiment and it will not...

Most of the time, it will not success the first time. That's why I wanted to complete everything and show it to people because there's no hundred percent.

For Kristina, releasing her data before completing her experimental design is a double risk. If the data are "good," someone else could step in and find patterns before her, and take away the only piece of credit that could advance her career. If the data are "bad," someone could find out about it and this would jeopardize her reputation as a researcher even before she could gain one in the first place. I discussed with several scientists the problem of invisible labor related to the release of large-scale data collections of genomics data released prior to publication. Hank, a computational biologist, had a very interesting point of view on this issue, where he thinks that this problem will only get worse over time.

Hank: I think it is a very healthy push that data should be freely-available. Data generated with public funds should be made publicly available. But I think we have to be very careful that in our zeal to make the data available, we do not jeopardize the next generation of researchers. […] If you talk to the data consumers, there's nothing better than open data. If you talk to people that had to spend two, three years setting up a sample collection and then another two years getting the money to generate data with that sample, you're talking about somebody investing five years of their life to generate certain datasets, that then with a click of a button is downloaded. And that disconnect between the amount of effort that sometimes goes into making this data available and then the effort in like, well, I downloaded the data, I did all the analysis without having the actual understanding, the concept of what it took to actually be able to provide that data to somebody to analyze. That disconnect is problematic.

Hank explained to me that computational biologists, especially those publishing in bio-statistics journals, increasingly mine multiple large genomics datasets at once with meta-analysis techniques that look for novel genes/phenotypes patterns, the so-called "long hanging fruits." When multiple datasets collected by different teams are downloaded and analyzed at an integrated level, it becomes a real challenge to give credit to each single data collector through authorship or citation. From my interview with Hank:

Hank: My worry at the end of the day, and you see that in the field of genetics as well, as a young investigator... I'm now talking about the field of genetics, like in genetic analysis. As a young investigator in the field of genetics, how do you make a name for yourself these days? How do you get your first R01? How do you... What do you do?

Hank: Are you going to propose to do an exome sequencing study, a whole genome sequencing study, when there are massive, massive consortia doing tens of thousands of samples. How are you safeguarding your next generation of geneticists coming out of that? Is everything going to be commoditized, so that you just have computational scientists working?

Hank: You can be a genetics lab that did investigation of syndrome X, syndrome Y. You generate data for that, right? You will do a genome screen.

Hank: Now, you... Everything is going to very large datasets, right? Especially for common disorders.

Hank: You want to have sample sets of thousands of thousands, tens of thousands or maybe even bigger datasets, which you can only get if everybody starts pooling their samples into one repository and have a central organization that types all of that. But then, how do you safeguard all these different genetic groups of that sample collection and then want to do analysis? How do you say exactly what you did? What was

your part?

Hank: Yeah. What was your contribution? And how do you get recognition for your contribution in that? That was... That then becomes difficult when you're talking about a completely different model of doing science, right? Because in those large consortia, there are usually analysis groups that are associated, they are either going to be one big mass of metadata analysis in all of that, and that's going to be the thing that catches the spotlight, right? That's going to be the publication or a set of publications that generate the hits and the citations. And each individual group that contributed the sample set, what is their... What is then their...

Hank: How do you profile yourself among such a large consortium? And there are going to be winners and losers there. And that's, I think...

Hank: I think it's potentially problematic for young investigators. But I think it's the early days, and it's going to be interesting to see how that develops in the coming years.

Hank: At the end of the day, people need to be motivated to do research. And if you're working as a cog and a wheel for some amorphous big consortium, where you do not know what's actually going to happen with the data that you're analyzing would generate, but that it's also very field specific. And I think genetics is where this is going to play earlier in our types of fields where sort of the gene-based focus is still very much considered [1:00:25] ____.

Hank: And the problem with that is it's going to only appear down the line 'cause people have already invested, these data are going to keep flowing for a while. But at some point, people are going to be dincentivized to start collecting large datasets and there's going to be a lag between when that's going to start appearing down the line. But again, it's as simple as that, basic [1:03:00] ____ but it is a potential risk of more open data. That's not to say that data shouldn't be open but there's a risk that we need to be very aware of and make sure that we anticipate and that there is all credit and attribution to the people

that...

When I asked the scientists what piece of information is the most valuable for them in order to decide to run a secondary data analysis on someone else's data, they would say "the contact information" of the person who collected the data. Overall, data reuse seems to be a highly social activity regulated by many un-official community norms. Emanuela explains how she values collaboration over competition when it comes to re-using others' data. Kristina talks about the "etiquette of co-authorship," where co-authorship is given to those who provide access to closed data, and contribute to their re-analysis.

Irene: Okay. Do you ever download data collected by someone else and run a new analysis asking a different research question?

Emanuela: Not usually, no. […] If somebody has an interesting data set and they could do that analysis or ask that question, and I try to approach it from like a, "Let's collaborate on this question." kind of a thing, instead of, "I'll take your data and do something with it myself."

Kristina: While I was working on my story (research study), I contacted another person who was working on the area of chromosome I was working on. He has access to five more databases I didn't get to look at, he looked up my gene of interest in those databases to see if anything was of interest for me.

Irene: But the person gave you access to these databases as a personal favor?

Kristina: Yes, basically the thing was authorship. If we were to find out something which I was going to publish in my paper, I would have given him authorship definitely and acknowledged his thing, because it just makes the story more interesting. I will obviously acknowledge that this part of data is from these

collaborations and this database, and this guy has helped me with it. We will offer them authorship. It is an etiquette, and then if they want it it's fine, if they do not then we acknowledge them and that's it.

To sum up, the practice of accessing others' data at a "raw" level to conduct secondary analyses is complicated by many factors. First, there seems to be an epistemic challenge in asking new questions from old data, especially when these are conceptualized as "hypothesis free." Some researchers reported to be unsure about how to independently pose novel research questions to mine others' data that have been supposedly collected in the context of hypotheses free studies. At the same time, other scientists do not believe that their data are truly hypothesis free, and think that – given the fairly specialized nature of craniofacial researcher – potential reusers will need to contact them to make sense of their data.

Second, there seems to be a problem of trust in "unpublished" big datasets. The researchers I interviewed tend not to trust the quality and accuracy of high-throughput data when these are released right after data collection and prior to publication. Also, it is by no mean guaranteed that "raw" data are shared along with complete and up-to-date metadata and annotations – which help to conduct secondary analyses. In relation to human data, obtaining a new IRB accreditation for data reuse could take a long time and be extremely time consuming.

It emerged that the participants in this study tend to re-analyze others' raw data with the help of the data creators. Data creators help data reusers to make sense of "raw" data by sharing with them the specialized knowledge and the technical skills that are

necessary to extract novel findings from "raw" data. Often these collaborations end up in co-authorship. At least in the cases I have discussed with the participants in this study, those who end up co-authoring the final papers are in most cases involved in their analysis, but not always. Co-authorship is often a means to reward those who collected the data – especially young academics – but it could be used as a currency to access closed data.

A data reuse story: the case of DNA-based facial reconstruction

"You know me by my face, you know me as a face and you never knew me in any other way. There-fore it could not occur to you that my face is not myself."

Milan Kundera, Immortality

The DataFace Genome Wide Association Studies (GWAS) datasets have been reused in a variety of research projects. The Consortium received hundreds of download requests for the GWAS datasets, which include DNA sequences, facial images, statistics of facial measurements, and metrics (landmarks and linear distances) for mapping and quantifying the human face. As previously discussed, the DataFace GWAS datasets – differently from most of the other DataFace datasets – were collected from the non-syndromic "general population." The low specificity of these datasets – especially compared to the other DataFace datasets – makes them valuable for reuse in many different contexts. The DataFace GWAS phenotypical data are available on the DataFace website, while the GWAS genotypical data are stored in the dbGaP database.

Some DataFace GWAS datasets have been reused in a line of research that in turn informs the design of DNA-based facial reconstruction technologies. Several scholars from the biomedical domain, as well as in legal studies and bio-ethics, criticized the

current research on DNA-based facial reconstruction that claims to digitally reconstruct individuals' facial portrays solely from their DNA samples. As I will discuss, criticism focuses on the scientific validity of this line of research, and on the ethics of employing DNA-based facial reconstruction in criminal investigations. Evaluating to what extent reconstructing faces from found-DNA at a crime scene is an ethical practice is beyond the scope of this dissertation. However, it is a fact that DNA-based facial reconstruction is a controversial research subject towards which the scientific community (and society at large) is highly divided. This "data reuse story" brings to surface the scientific debate surrounding this line of research. The overarching goal of this chapter is to provide empirical ground to reflect on the unexpected consequences of making research data openly available for reuse. My analysis raises the following concerns, which I will examine further in the discussion section:

- Given the instability and controversies surrounding the science of DNA-facial based reconstruction, especially in relation to its uses in criminal investigations, should the data donors be involved in choices related to the reuse of their data (in this case, DNA samples and facial images)?
- Given that we cannot predict with certainty how data and the associated analyses will be reused once made open, are IRB protocols and the Informed Consent the right tools to evaluate uses and reuses of open research data?

Reconstructing faces from DNA samples: appeal and controversies

The idea that we can reconstruct a human face from a DNA sample has great appeal: the faces of prehistoric peoples could be reconstructed from their remains, the face of a

child could be predicted in utero from amniocentesis, and DNA from a crime scene could be used to create a facial image of a suspect (Hallgrimsson, Mio, Marcucio, & Spritz, 2014; Zhang, 2017). In technical terminology, the latter application is called Forensic DNA Phenotyping (FDP). "DNA-Snapshots" obtained via FDP technologies are promoted as means to narrow down a search for a suspect under special circumstances (Kayser, 2015). For example, when there are no witnesses to a crime, and there is no match between the DNA found on the scene and the DNA of the suspects. DNA-Snapshots can be shared with the public, and they can also be run against thousand of law-enforcement mug shot databases, as well as against government databases containing citizens' ID photos (Gannon, 2017).

Forensic analyses of found DNA samples have been used for a long time to categorize suspects by "group-level" factors, namely gender and ancestry (Hindmarsh & Prainsack, 2010; M'charek, 2008). FDP is gaining quite some traction among law enforcement agencies because it aims at providing information about isolated facial traits, such as eyes and hair color, and – even more specific – nose, chin, and jaw shapes. These can be used in combination to group-level factors (i.e., sex and ancestry) to reconstruct supposedly highly-accurate facial profiles (Kayser, 2015). Also referred to as the "biological witness," FDP promises to be more objective than "human witnesses" (Kayser, 2015). The use of FDP in police search is spreading fast. Since 2015, when FDP technologies became commercially available in the US, American law enforcement agencies released over ten DNA-generated mug-shots to the general public, via television and online media, asking the public for help to identify the suspects (Biswas, 2015;

Purcell, 2016). At the time of writing this dissertation, two companies currently sell these services to law enforcement agencies in the US.

There is little regulatory oversight over how law enforcement can use DNA tests in criminal investigation and in court. DNA left at a crime scene is considered abandoned material, and police can use the information encoded in it in any way they consider useful for investigative purposes (M'charek, 2008). The use of FDP technologies have been criticized by several scholars in a multitude of disciplines especially in relation to potential ethical issues that can occur with the rapid and unregulated diffusion of these technologies (Dewey-Hagborg, 2017; Hallgrimsson et al., 2014; Toom et al., 2016). For example, researchers in the field of forensic medicine, and also in the social studies of science, argued that FDP has a problem of racism. Contrary to the ways in which it is promoted, it cannot provide data capable of probabilistically identifying unique individuals. Instead, the researchers argue, DNA phenotyping groups people in "suspect populations." Researchers observed that FDP "provides typological information about common but variable personal properties of relatedness to others, features of visual appearance, or aspects of biogeographic ancestry" (Toom et al., 2016). Researchers are concerned that DNA-generated mug-shots can be used to justify "genetic dragnets" in which hundreds of individuals belonging to a sub-population are asked by law enforcement to provide their genetic material for testing and profiling. Scholars further argued that while human witnesses provide contextual information along with their testimonies, "biological witnesses" like FDP simply return a statistically inferred stereotypical facial image. If stereotypical facial reconstructions are run in police

databases to find matches, these could return highly biased results, also given that police databases and algorithms are already known to be biased toward minorities (Dressel & Farid, 2018). In some occasions, pro-FDP commentators sustained that because human external appearances are visible to everybody, their records should not be protected by privacy regulations (Kayser, 2015). However, critics of FDP rebutted that "there is a difference between spotting someone on the street once, and keeping and reusing data on someone physical traits" (Toom et al., 2016). For all these reasons, concerned researchers urge caution in using FDP for criminal investigations, and ask to keep facial images reconstructed from DNA-analysis strictly confidential to the investigators, do not share them with the public, to avoid stigmatization of entire sub-populations.

Research on FDP technologies

Another problem with commercial FDP is that "the algorithm" behind it is not officially published in any academic journal. The companies' technology is closed-sourced, and its machine-learning algorithm black-boxed. However, the company's technique is based partly on the work of an international team of physical anthropologists and computer engineers who published their model for predicting facial shapes from DNA starting from 2012. From now on, I will refer to this team as the "the team of physical anthropologists," to differentiate this lab – which is external to the DataFace Consortium – from the labs that collected the DataFace GWAS datasets and are part of the DataFace Consortium. This team is currently testing multiple versions of a machine-learning algorithm on a series of different training datasets. The latest model, published on *Nature Genetics* in February 2018, employs one of the DataFace GWAS datasets as a training set for the algorithm. In an commentary to *PLOS Genetics* called "Let's Face it:

Complex Traits are not That Simple," some of the scientists involved in the DataFace Consortium – but not in the collection of this specific dataset – highlighted a series of concerns related to the accuracy and scientific validity of this model for facial reconstruction. They argue that the science behind the genetic causes of facial morphology is still in its infancy, and that other studies showed fewer or different correlations between genes and facial traits. In what follows, I first describe the datasets involved in this controversy and the state of art of the research on DNA-based facial reconstruction. I then present the debate over its validity, and I conclude by pointing out the role that the DataFace datasets played in this context. Direct citations to academic articles have been removed to protect (to some extent) the confidentiality of this research study's participants.

The DataFace GWAS Datasets

In order to predict faces from DNA, researchers and developers need, first of all, to know to what extent facial development and morphology are dictated by genetics, beyond obvious parental-siblings resemblance. They need, in other words, precise knowledge of the genetic causes behind the formation of facial features like eyes color, or nose shape, which are generally referred to as "Externally Visible Characteristics" (EVCs). While eyes and hair color have been (statistically) associated for a long time to certain locations on the genome, the research on specific genes that have an impact on individual facial traits such as nose, chin, and jaw dimensions, is still in its infancy (Hallgrimsson et al., 2014). As pointed out by several commentators, one of the main challenges with studying the genetic causes of EVCs is that researchers need to have access to a large number of genetic and phenotypical data related to the human face, which collection requires

considerable investments in time, human resources, and infrastructures (Kayser, 2015; Roosenboom, Hens, Mattern, Shriver, & Claes, 2016). In this context, the human craniofacial datasets collected and made available by the DataFace Consortium constitute a quite precious, and scarce, resource. These datasets are described on the DataFace website as "an excellent resource for exploring questions relating to patterns of human facial variation and growth; e.g., how does the face change over different life stages, how are sex differences manifested in facial structure, how are different facial structures integrated during growth, and what are the major facial differences among different ethnic/ancestral groups […]." The DataFace database for GWAS data is described as "of particular interest to those working within the fields of physical anthropology, orthodontics and forensics." The database technical notes also suggest that: "the thousands of 3D facial surfaces available through the 3D Facial Norms Database provide a unique resource for computer science and computer vision experts to develop novel surface-based methods for representing and analyzing human faces."

Overall, the research design for craniofacial GWAS studies can be summarized in four main steps, namely (1) the collection of the DNA samples and the 3D facial images from human subjects, (2) the extraction of quantitative measures and landmarks from the 3D facial images, (3) the sequencing of the DNA samples, and (4) the statistical investigation of possible relations between genes and facial variations. Two different DataFace teams conducted two separated GWAS with data collected from two distinct populations. I called the two resulting datasets "the Caucasian dataset," and "the Tanzania dataset." "Caucasian" and "Tanzania" are terms used by the teams themselves

to refer to the "populations" sampled in each GWAS.

The Caucasian dataset

Data for the Caucasian population are stored and made accessible through the 3D Facial Norms (3DFN) Database, which is hosted on the main DataFace website. The Consortium described the database as "a web-based resource designed to provide the research and clinical community with access to high-quality craniofacial anthropometric normative data." The database includes anthropometric data of different kinds. One data type consists in a variety of metrics and statistics about facial measurements. Metrics include 3D coordinates (x,y,z) for facial surface landmarks (i.e., reference points on the face), linear distances calculated between the landmarks, and actual face and head measurements using traditional anthropometric methods (i.e.,, calipers). Another type of data consists in the facial images themselves. The 3D facial surfaces were released in the Object Wavefront (.obj) format. Some demographic descriptors were released in association with the data, and these include age, sex, and ancestry. Finally, the consortium released the genotypic markers associated with the data (the raw data are available on dbGaP).

The Caucasian dataset was collected during DataFace first grant phase. Subjects who volunteered their DNA and their facial images for the study include 3500 unrelated males and females of European-Caucasian ancestry between the ages of 3-40 years. The participants were recruited at three main sites within the US. Recruitment strategies vary from site to site but include the use of public print advertisements, word-of-mouth, direct mailing, university and hospital-based research registries, kiosks in public venues (e.g., commercial malls), and collaborations with general dental and medical clinics.

The Caucasian dataset can be consulted in multiple ways. Indeed, the 3DFN database allows users to interact with the data via a graphical interface, a type of reuse that I previously characterized as "background" reuse of others' data. Summary-level data include things like sex- and age-specific means and standard deviations for selected anthropometric measurements (e.g., the average distance for five year old males). As such, all summary-level phenotypic data are non-restricted and available to all users directly on the database website.

By giving proper credentials, users can also gain access to individual-level data, which allows investigators to perform their own analyses and reuse the data to ask novel questions (what I previously characterized as "foreground" reuse of others' data). Individual-level data refer to the unique data elements that comprise the summary-level data, and include things like the measurements, landmark coordinates and 3D facial surface files for each individual in the database. As pointed out on the DataFace website, "one major advantage with individual-level data is that the users can carry out their own statistical analyses on original raw data as if they collected the data themselves." Access to all individual-level phenotypic data (facial 3D images) is restricted to users with the proper permission. In order to access restricted datasets and reuse them in their own research studies, potential reusers need to obtain an IRB approval from their institutions. The users then need to submit a "Data Access Request form" and associated documentation to the "DataFace Data Access Committee" for review. Upon approval, the DataFace Hub grants the user permission to access and download the requested data

through the DataFace website.

DataFace researchers obtained the data for this GWAS study using "consistent and semi-automated methods for facial quantification and measurements" (Weinberg et al., 2016). These methods were developed by the DataFace team "to address the shortcomings of traditional craniofacial datasets," which – according to the DataFace scientists – "are limited to measures obtained with handheld calipers and tape measurers." As reported in the craniofacial literature, measurements obtained through direct anthropometry often result in some degree of deformation, caused by the soft tissues of the face with the caliper tips, which leads to lack of standardization between datasets obtained by different teams and at different sites. Lack of standardization makes comparative studies challenging. The most well-known and most comprehensive dataset of this kind was compiled by Dr. Leslie Farkas and colleagues in the 1980s and 1990s (Farkas, 1996). Kolar (1993) has pointed out numerous problems with this particular dataset, the most serious of which – as indicated by the scientists- seems to be the inconsistency of data collection protocols. In this context, the anthropometric data collected by the DataFace researchers aims at being "the right tool for the right job." Obtained via 3D digital stereophotogrammetry – a method of 3D imaging increasingly used for capturing human facial surface morphology – DataFace images promise consistency and precision. Measurements computed from the 3D surface models are supposed to be more "objective" because they involve much less deformation.

From each image, the researchers located, with the help of a computer program, a set

of 24 “facial surface landmarks” (i.e., reference points on the facial area). Based on these landmarks, the researchers used basic Euclidean geometry to calculate the precise measurements of 29 linear distances between different sets of multiple landmarks. Starting from the 24 initial landmarks, the 3DFN database users can calculate any number of alternative inter-landmark distances directly from the raw coordinate data available through the 3DFN Database.

The measurements obtained via 3D images were still complemented with additional linear distance measurements obtained through traditional direct anthropometry, which are also available through the 3D Facial Norms Database. These were large measurements of the head and face, which were difficult to capture through indirect 3D surface anthropometry.

The team who collected the Caucasian dataset conducted the primary analysis of this dataset in collaboration with a second DataFace team, the one responsible for the collection of the Tanzania dataset (see next section). In the resulting overview paper, the two teams reported to have found evidence of genetic associations involving measures of eye, nose, and facial breadth. This study represents the second round of GWAS on facial morphology for Caucasian populations, the first round appeared in 2012 (Liu et al., 2012; Paternoster et al., 2012). The findings from the DataFace Caucasian dataset partially replicated the findings from the previous two studies. For example, both round of GWAS studies found the gene PAX3 having a role in shaping the nose’s “bridge elevation.” However, the DataFace researchers were not able to replicate most of the findings from

the previous round of GWAS studies. In the overview paper, the DataFace researchers reported that their ability to find significant genetic associations "was limited by a lack of directly comparable phenotypes, which is related to differences in data collection methods and the type and number of measurements available." In addition, they point out that "the prior two European GWA studies each used imaging modalities different from the kind used here." In the same publication, the DataFace researchers also noted that: "(…) fortunately, several promising approaches are on the horizon, such as the BRIM method." The "BRIM method" – as I will shortly discuss – is one of the early iterations of the research design that team of physical anthropologists will eventually use to reconstruct faces from DNA. In a survey paper called "New Entries in the Lottery of Facial GWAS Discovery," the physical anthropologists conducted a secondary meta-analysis of the data obtained from the first two rounds of Caucasian GWAS studies. In this second paper, the physical anthropologists further discuss (and visualize) the significance of the association between the PAX3 gene with the nose formation in the Caucasian population.

The Tanzania dataset

The DataFace consortium funded the collection of a second GWAS study with participating subjects of African ancestry. A different team of researchers located at different US institution handled this second study. This is "the first GWAS of facial morphology for an African population." I will refer to this dataset as the Tanzania dataset. The African GWAS cohort included 3,505 non-syndromic African Bantu children and adolescents ages 3–21 from the Mwanza region of Tanzania, a region that the team described as being "both genetically and environmentally relatively

homogeneous."

For this second study, methods for data collection and annotation slightly differed from the ones used for the Caucasian dataset. For example, they used the same methodology for 3D facial imaging collection – 3D digital stereophotogrammetry – but a different camera, namely the Creaform MegaCapturor (MC) camera. Also, while the Caucasian dataset is annotated for 24 landmarks on the human face, the Tanzania dataset is annotated for 29 landmarks.

Also for the Tanzania dataset, the landmarks on facial images were identified using a semi-automated computational method. The landmark data were then used to calculate linear distances and multivariate measures to be used as phenotypes, as in the Caucasian study. Still, some images from 163 subjects were landmarked manually, "as they could not be landmarked automatically." The team explained: "This was mostly due to imaging artifacts on non-critical regions of the face that do not interfere with manual landmark placement." The details of the method and of the automated landmarking algorithm used to identify the landmarks on the human faces are available on the DataFace website.

The creators of the Tanzania dataset published an overview paper in late 2016 on *Plos Genetics*. The data analysis suggests that only two locations on the genome are significantly associated with "measures of facial size." For one of these loci, the team conducted an experiment in which they knocked out the loci from mice. This showed developmental anomalies in the palate and in the snout, indicating that the locus plays

indeed a role in facial development. This is one of the first studies to show a genome-wide genetic association of human facial morphometric phenotypes in an African population.

To some extent, results from the analysis of the Tanzania dataset differed from those on Caucasian populations, including from the analysis of the DataFace Caucasian dataset. For example, the Tanzania team was not able to replicate the influence of gene PAX3 on the shape of the nose bridge. In their overview paper, the team suggests caution in interpreting the possible causes behind such differences: "It is possible that facial morphology differences in different human populations have different genetic underpinnings," but, they add, "alternatively, as noted above, our study cohort was young and almost universally lean, and therefore may be less influenced by environmental factors than study cohorts of adults from Europeans' populations." The researchers seem to argue that differences in genetic markers associated with the face between populations could be explained by the fact that the Caucasian populations are on average older, of a different body constitution, and were also exposed to several environmental factors to which the Tanzania population was not. Overall, in the discussion section, the researchers seem to make the point that craniofacial GWAS studies are not design to specifically to find variation "in between" populations, but within the same population.

Modeling 3D Facial Shape from DNA

"Our notion of symmetry is derived from the human face"

Blaise Pascal (1623–1662)

In a study titled "Modeling 3D Facial Shape from DNA," a team of physical

anthropologists (not a member of the DataFace Consortium) analyzed a sample of 592 individual genotypes for 540,000 SNPs, expressed in 46 different genes. The research design for this study is quite different from the GWAS studies conducted by the DataFace teams. The team of physical anthropologists chose not to use a GWAS approach on purpose, for multiple reasons. According to the authors, GWAS approaches have a problem of statistical power. They argue: "the fundamental problem with taking a naïve whole genome scan approach is one of statistical power […] the larger the number of markers that are tested for significant effects on facial variation, the larger the number of false positive results and the harder it will be to know which among those that show significant effects are actually having important effects that should be used for DNA-based facial composites." From this perspective, using GWAS to look for genotypes/phenotypes associations in facial morphology is like looking for a needle in a haystack. However, by using what is referred to as a "candidate genes" approach, researchers can narrow down the search for gene/phenotype correlations to few usual suspects, namely genes that have been previously identified as potentially expressed in facial development (46 genes, in this case).

The physical anthropologists' study differed from a GWAS also because they used an "admixture" approach to data sampling and analysis. An "admixture" approach enables researchers to purposefully look for variation across individuals with supposedly difference ancestry backgrounds. The admixture approach uses ancestry informative markers (AIMs) to estimate individual genomic ancestry from DNA (African, European, Native American, East Asian etc.), also called Biogeographical Ancestry (BGA) (Halder

& Shriver, 2003). In simple terms, BGA identifies "racial percentages" that are expressed in the heritable component of the face. The idea behind admixture sampling is that "non-random mating and continuous gene flow in admixed populations results in admixture stratification or variation in individual ancestry" (Halder & Shriver, 2003). The process of admixture results in "admixture linkage" or "non-random association" among AIMs and traits to vary between individuals with different ancestry backgrounds (e.g., skin pigmentation). This is similar to the technology used by commercial companies that offer direct-to-consumer genetic ancestry testing.

According to the authors of this paper, GWAS studies for facial traits are also limited in their description of facial morphology. As we have seen, DataFace GWAS studies relied on a limited set of pre-determined landmarks and linear distances to calculate facial variation. The physical anthropologists are convinced that facial variation can go well beyond this pre-determined measurement techniques. In order to bring to surface hidden variations of the human face, the team of physical anthropologists had previously developed a fully automated computational method for the mapping and quantification of the "full" facial morphology. This method relies on the use of a "digital anthropometric mask" made of thousands semi-landmarks. The mask is graphically imposed over the 3D facial images of participating subjects to map them "onto a common coordinate system." The mask is applied automatically, "eliminating the difficult and error-prone procedure of manually indicating facial landmarks."

Finally, in order to find genotypes-phenotypes correlations, the authors used a novel

statistical methodology for shape prediction that was developed with the help of machine-learning algorithms. The authors named it "bootstrapped response-based imputation modeling" (BRIM). According to the authors, the advantage of the BRIM method is that it allows researchers to estimate facial shape from a single multidimensional factor, such ancestry, sex, or a single gene. By using the BRIM method, "the effects of sex and ancestry can be isolated and optionally removed from the model, thereby providing the ability to extract the effects of individual genes." In the authors' words, "(with this method) the hypothesis *Does this gene have significant effects on facial shape* can be addressed with a single statistical test."

In the abstract of the paper, the researchers announced that they "uncover the relationships between facial variation and the effect of sex, genomic ancestry, and a subset of craniofacial candidate genes." Results from the BRIM data analysis method suggested that many parts of the face are affected by both ancestry and sex. More specifically, findings from this study suggested that sex explains 13% of the total shape variation in the face, while ancestry 10%. In the authors' own words: "these results provide the means for identifying the genes that affect facial shape, and for modeling the effects of these genes to generate a predicted face."

In a second paper titled "Toward DNA-based facial composite: Preliminary Results and validation," the team conducted a secondary analysis on the same dataset in what they describe as "the first effort of generating facial composites from DNA." This time, the researchers first used genomic ancestry and sex as main factors to create a "base-face." Subsequently, they overlapped the effects of 24 individual SNPs in 20 genes on the

base-face, in a process akin to a photomontage. The team concluded that "physical accuracy of the facial predictions either locally in particular parts of the face or in terms of overall similarity is mainly determined by sex and genomic ancestry," and that "the SNP-effects maintain the physical accuracy while significantly increasing the distinctiveness of the facial predictions, which would be expected to reduce false positives in perceptual identification tasks."

The debate over method: complex traits are not that simple

The very idea that complex facial traits could be predicted from DNA samples raised concerns among some members of the DataFace community. One thing is to say that a set of genes influences facial development processes, a very difference thing is to say that a set of genes predicts individuals' facial traits with statistical accuracy. In an article titled "Let's Face it: Complex Traits Are Not That Simple," a group of researchers from the craniofacial research community, including some DataFace principal investigators, argued that the physical anthropologists' model for predicting human faces based on genes, sex, and ancestry, needs to be replicated in the context of existing methodologies, in particular GWAS studies, in order to be considered methodologically valid.

In the "formal comment" to PLOS Genetics, the researchers stated that the claim that facial shape can be predicted from DNA is troubling because, "it is not actually supported by the work done in this study." First of all, they argue, the physical anthropologists' work is "genes biased" because they used as a starting point a list of genes previously known in the literature to be expressed in the head of animal models with craniofacial

abnormalities. This fact, the commentators argue, "it does not mean that those genes contribute to normal variation in the face" and, they add, "it is quite possible that many genes not known to play important roles in craniofacial development contribute to normal variation in the face." The commentators further pointed out that "only one of the 46 genes identified in this study would have survived the Bonferroni adjustment for multiple testing." In the absence of multiple tests, "the study contributes nothing new to our understanding of how genes influence the shape of the face since the genes tested may or may not actually contribute anything to normal variation in the shape of the face."

The commentators also observed that the physical anthropologists' finding that ancestry has a great impact on facial shape variation (i.e., 10%) is "unexpectedly high" and further characterized this finding as a "surprising result." As previously discussed, also the DataFace GWAS studies conducted on the Caucasian and the Tanzania dataset showed some differences in genes expression in between the two populations, such as for the PAX3 gene and its effect on nose shape. However, the DataFace PIs were generally careful in providing genetically deterministic interpretations of such results, suggesting instead that such differences could be related to the environment or other factors. A significant difference between the physical anthropologists' model for facial prediction and the DataFace GWAS studies on facial variation lies in the way they use "ancestry background" as a factor for facial prediction. In the GWAS studies, ancestry background is used as a statistical tool to normalize the sample and obtain a seemly homogeneous population. This homogeneous population is then employed to find genetic and phenotypical variation within the population itself. While this type of approach surely

relies on racial classifications, at the same time, it does not explicitly aims at finding differences in between populations, quite the opposite, it aims at finding differences among individuals belonging to the same "racial classification." On the contrary, the admixture research approach used by the physical anthropologists, and employed in combination with the BRIM statistical method for facial prediction, it explicitly uses ancestry as a discriminatory factor (along with sex) to determine the "basic" features of a predicted human face. As a result, the "base-face" fundamentally corresponds to a stereotypical "base-face" of the population to which the sampled-subjects allegedly belong (M'charek, 2017).

The commentators wondered whether "genomic prediction of complex morphologies is even feasible." They insist that the genotype-phenotype map for morphological traits like the shape of the face is "incredibly complicated," and that changes to developmental processes can have much greater effect on facial shape than single genes or sex or race. These researchers, who conducted craniofacial GWAS studies themselves, conclude by reminding the physical anthropologists that current GWAS studies are finding very few genome "loci" to be causatively related to complex facial traits.

Refining the prediction model for human faces

The commentators challenged the physical anthropologists to design a second study that consistently shows that certain genes can be significantly associated to certain facial phenotypes, either at a group-level (ancestry, sex), or at an individual-level (nose, jaw, chin etc. dimensions). In other terms, the researchers called for replication of the physical anthropologists' findings within a different research design. Without replication, their

prediction model for facial reconstruction from DNA loses credibility.

Open data resource for craniofacial research include many different datasets collected in the context of GWAS studies, animal studies, dysmorphology studies, populations studies, and family studies (Roosenboom et al., 2016). But the commentators specifically called for the replication of the physical anthropologists' results in a GWAS study, given that the "genes candidate" research design previously used by the team is considered genes-biased. Since organizing and conducting a GWAS study is expensive and time consuming, most of these studies are conducted in the context of large consortia, such as the DataFace Consortium. To replicate their study, the team of physical anthropologists conducted a secondary analysis of multiple craniofacial GWAS datasets, which included the DataFace Caucasian dataset. The results were published on *Nature Genetics* in early 2018. The secondary analysis of the DataFace Caucasian dataset was conducted in collaboration with the DataFace team who collected the Caucasian dataset, but not with the team who collected the Tanzania dataset.

In the Nature paper, the physical anthropologists used a refined methodology that aims at improving traditional GWAS studies, which they describe as being "phenotypic-first" types of approaches. From this perspective, in "phenotypic-first" studies, the search for correlations between genes and traits is limited by the fact that the phenotypes are pre-selected and used to classify individuals based on pre-fabricated linear distances. In the paper, the researchers applied a fully automated data-driven approach for facial mapping to the analysis of the DataFace Caucasian GWAS dataset (plus other datasets).

To be sure, this time the team used an improved version of their previous methods. The improved technique allows for the identification of the genetic effects on facial shape at multiple levels of organization, "from global to local." By employing an unsupervised machine-learning algorithm known as "hierarchical spectral clustering," the face is subdivided in different fragments on the base of 10.000 semi-landmarks, from general (i.e., global) to more specific (i.e., local) fragments. As explained by the physical anthropologists, the shift from facial "linear distances" to "independent modules" enables the researchers to study the human face "as a whole," instead of focusing on few pre-determined traits. In the authors' word: "This method provided an efficient and objective way means for subdividing facial shape into parts."

This approach resulted in the identification of 63 facial segments, that have then be tested over 9 million SNPs. The team used Generalized Procrustes Analysis (GPA) and Principal Component Analysis (PCA) to extract the major factors of shape variation characterizing each facial segment. The results suggest that, this time, 15 loci are involved in a variety of facial segments, mainly the nose and the chin. 9 of these loci have been previous found to be associated with facial variation, including locus 2q36.1 on the PAX3 gene for the nose bridge. 4 were completely new, and 2 were found to be associated with more than one facial segment.

Overall, this last study suggests that variation in the human face seems to be influenced by many genes that exhibit a range of effects, "with some influencing only localized parts of the face and others influencing more global aspects of morphology." In

this paper, the researchers built a method with the intention of studying variation "freely" from pre-determined linear distances, and, at the same time, across a large number of SPNs. In this sense, the method developed in here by the team of physical anthropologists aims at maximizing the chances of finding correlations between genetics and facial traits. The team of physical anthropologists' goal is not to simply respond to previous criticism, but to propose a method that can be adopted by virtually all researchers to analyze human faces in GWAS craniofacial studies. In their words, "[with this paper] we substantially advanced the literature on facial genetics on several fronts."

6. Discussion

This discussion starts by situating DataFace policies and infrastructures for data sharing and reuse in a broader framework, especially in relation to what I referred to as the "radical openness" regime for data sharing. Then, I examine how the scientists participating in the DataFace Consortium reuse open data for knowledge production. I conclude this discussion by considering ethical implications of making research data openly available for reuse, especially in relation to the impossibility of truly predicting what open data might be reused for.

Regimes for Data Governance and the DataFace Consortium

I have discussed how – in a data governance regime of semi-openness – the publication status of a dataset's primary analysis (i.e., published or unpublished) regulates when the dataset would be made publicly available for reuse. In a semi-open regime, researchers deposit their data in public repositories exclusively after publication, and they share their data prior to publication in closed collaborations inter-labs. Depositing the research data after publication guarantees a certain degree of transparency, while at the same time allows others to make use of the data. Retaining data until publication also ensures that the data creators receive credit for the design and execution of the experiments in which the data are collected. By sharing "unpublished" data solely within closed collaborations between the data creators and the data reusers, scientists control for what purposes their data are reused. In this data governance regime, researchers share their tacit and specialized knowledge (which is essential to properly reuse their data) with few trusted colleagues, and they negotiate with them credit attribution – mainly by co-authoring the resulting publications.

With the advent of the "genomics revolution" and the completion of the HGP, sequence data started to be shared prior to publication in open repositories (Hilgartner, 2017), *de facto* destabilizing pre-existing semi-open practices of data sharing. Hilgartner (2017) referred to research data made available prior to publication in open repositories as "Unpublished in Journal, Available in Databases" (UJAD) research data. The UJAD data-sharing regime partially emerged as a response to the deluge of sequence data that were generated during the HGP. As Hilgartner observed, it was also motivated by the anxiety of finding significant patterns in the freshly sequenced human genome. In this context, making data available prior to publication seemed to be a good way to guarantee fast reuse of such data, and, as a consequence, a return on the investment for the HGP. During the HGP, depositing data prior to publication became an accountability mechanism that the HGP leadership would use to measure the success of a participating sequencing center.

As pointed out by Hilgartner (2017), the idea of making sequence data available prior to publication to the whole research community was rooted in the belief that sequence data are fundamentally different from "results data." The results data are those coming out of experimental design and statistical analyses, and are normally deposited along with peer-reviewed publications. Contrary to results data, in this perspective, sequence data are seen as the products of factory-style sequencing facilities that extract "raw" resources that do not need any specialized labor or significant amount of time to be produced (Mike Fortun, 2008; Stevens, 2013). As a result, the HGP's leadership did not expect the HGP

sequencing centers to receive any academic recognition for the collection of sequence data. Sequence data of whole human or animal genomes were believed to be "generic enough" type of data that can be reused in all sorts of contexts. These two conditions of the HGP's sequence data – supposedly being labor and hypothesis free – motivated the funders to conceptualize sequence data as what Leonelli would call today "fungible resources" (Leonelli, 2016; Mirowski & Nik-Khah, 2017). Just like freshly extracted "crude oil," freshly sequenced data can be reused by anyone who has the means to transform them into usable knowledge. Given sequence data labor free commitment, high potential for reuse, and abundance, the funding body saw no reasons why the sequencing centers should have kept sequence data "hostage" in their labs until publication.

However, as Hilgartner (2017) also observed, HGP's sequence data were not hypothesis free commodities. Commentators showed how these data were collected by specialized research communities, such as the model organism communities, and in the context of specific research agendas (García-Sancho, 2012; Gaudillière & Rheinberger, 2004; Leonelli, 2016). At the sequencing centers, those researchers directly responsible for the sequencing of the data were very much interested in analyzing them and using them in their publications. HGP data were also not labor free. As discussed, because of its repetitive and quasi-automated character, sequencing was surely seen as a "factory style" type of research activity. But, in addition of being "boring," the practice of sequencing would also take away from the scientists precious time (i.e. labor) that they could have used for conducting primary data analyses and publish papers. Sequencing was still time-consuming and did not provide credit in terms of career achievements.

As I have discussed in the findings section, the participants in the DataFace Consortium refer to the data collected in the context of this project as "hypothesis free" research data. Like the sequence data during the HGP, also the DataFace data are marketed as not being collected by one lab for the benefit of that lab, to answer a pre-defined set of research question. I have shown that both the DataFace leadership and the DataFace participants conceptualized DataFace data as resources to be reused by many labs to answer many research questions in relation to craniofacial syndromes and beyond. Like the sequence data during the HGP, the DataFace datasets are understood and promoted as fungible commodities. Another point of similarity between the HGP and the DataFace Consortium is the presence of a strong leadership. This leadership, in both cases, saw "openness" of the data as an accountability mechanism for measuring the success of the participating labs. The Consortium was funded as U01 cooperative agreement grant, which – as I explained in the finding section – motives the collection of hypothesis free genomics data. At the same time, it enables the leadership to retain control on setting up the goals of the project and on monitoring its advancements. Finally, like in the case of the sequence data coming out of the latest phase of the HGP, the DataFace datasets are Unavailable in Journals, Available in Databases (UJAD) datasets. All datasets are shared prior to publication in an open repository. In the DataFace context, sharing data before publication is regarded as an optimal solution to advance research in the craniofacial field, by making the research process faster and optimized. In this frame, by opening data right after collection, raw dataset can be re-purposed immediately,

multiple research questions can be investigate at the same time, and replication of efforts are avoided.

But the DataFace Consortium is also obviously very different from the HGP, and not only for its much more limited scope. The data types collected by the participants are not only sequence data, but they also include facial images, measurements and statistics, and gene/RNA expression data coming out of experimental practices, such as candidate gene experiments and function validation studies. The DataFace datasets – like in the case of the Green and the Blue spoke (see findings section) – are after all still collected in the context of specific experimental designs, with specific research questions in mind. The only datasets that are to some extent "generic enough" (i.e., supposedly "hypothesis free") to be widely reused in all sorts of research contexts are the human subjects GWAS datasets collected from children and adult general populations (Caucasian and African). It is not a chance that these datasets are also the most reused datasets among all datasets collected from the DataFace collaboration. Set apart the GWAS datasets, most of the DataFace datasets were collected in the context of specialized research on rare craniofacial syndromes and craniofacial developmental processes. Most of the participants in this study, maybe with the exception of the Pink spoke, were relatively new to the practices of genomics. As I have discussed, the majority of the participants conducts "gene-centric" research projects (see the workflow of the Blue spoke), or aims at integrating genomics screenings to developmental biology research designs (see the workflow of the Green spoke). This condition of being "in transition" from gene-centric to genomics methodologies translated in both enthusiasm and skepticism toward high-

throughput data collections and analyses. While the participants are eager to incorporate genomics approaches in their experimental practices, at the same time they are quite suspicious of the idea of sharing and reusing others' "big data," especially when these are shared prior to publication.

Another dimension that differentiates the DataFace data collections from the HGP sequences is that most of the DataFace datasets were collected in the context of rare disease research. Rare syndrome research – such as research on craniosynostosis diagnosis and treatment –is a "data scarce" environment. Craniofacial syndromes are many, but very diverse. For example, very few people are born with craniosynostosis, and, as a consequence, little information and data are available about the syndrome. Because of its limited nature – the data on craniosynostosis are very specialized. Given the scarcity of these data, and the over-specialization of the field, the competition over their use by multiple labs is a zero-sum game. Not that many people are interested in these data, and those who are interested they want to use them to ask very similar research questions. When a lab working on craniosynostosis makes available their specialized and rare data – prior to publication – the chances that these data are re-purposed to ask a "novel" research question are much lower. Obviously, this situation creates issues of credit attribution. As discussed next, following a well-established tradition in molecular biology, the researchers would deal with this situation by sharing co-authorship between data creators and data reusers.

To sum up, the DataFace Consortium shared a vision for open data somehow similar to the "UJAD" data governance model that emerged during the completion of the Human Genome Project. Like HGP's sequences, also DataFace datasets were made available right after data collection in an open repository and were conceptualized as "hypothesis free" data. UJAD datasets are the product of a regime of "radical openness." What is relevant about this radical openness is that it differs from precedent semi-open traditions of data sharing that characterized the biology field for a long time, such as the semi-open data sharing practices of model organism communities. The regime of radical openness challenges the DataFace community on several fronts. Most DataFace datasets are quite specialized resources obtained via experimental practices. The fact that DataBase datasets are specialized makes them difficult to reuse to ask infinite research questions, and, at the same time, increases the bar for competition over their reuse. Finally, because the craniofacial researchers operate in a relatively traditional "gene-centric" knowledge production domain, they are concerned over reusing others' high-throughput data in their research settings, unless these are related to a publication, and accessible through easy-to-use data visualizations.

Re-purposing others' data: background and foreground reuse

Because the value of scientific data lies in the possibility of using them as "evidence for phenomena" (Borgman, 2015), in order for the data to be reused such value – the data's evidentiary "power" – needs to travel with the data from context to context when these are made open and reused. Scholars who investigate data reuse practices in the sciences researched how scientists "trust" the evidentiary power of others' data (Birnholtz & Bietz, 2003; Jirotka et al., 2005; Wallis et al., 2013; Zimmerman, 2008).

Trust in the "data" and trust in the "system"

As discussed by several scholars, and certainly confirmed in this study, metadata and ontologies play an essential role in enabling trust in others' datasets (P. N. Edwards, Mayernik, Batcheller, Bowker, & Borgman, 2011; Wallis et al., 2013). Metadata allow scientists to verify the quality and accuracy of the data (Leonelli, 2016). Ontologies enable scientists to understand how relevant new data are to their research (Faniel & Jacobsen, 2010). Ontologies link conceptually different sets of data and incorporate them in specific knowledge representation schemas. Especially in biology and biomedicine, ontologies are crucial to enable reuse. As discussed by several commentators, higher levels of data integration can lead to higher rates of data reuse (Buneman, 2005; Jones, Schildhauer, Reichman, & Bowers, 2006). Some studies of digital repositories show that the researchers also value the functionality of a specific database, the reputation of the repository that hosts the database, and the type of organization responsible for the data curation process (Faniel & Yakel, 2017; Peer, Green, & Stephenson, 2014; Ross & McHugh, 2006).

Trust in the data creators

While trust in the data and trust in the system are important for reusing others' data, commentators also observed that the judgment of trustworthiness is ultimately determined by the perceptions of the individual(s) performing the judgment, rather than solely by the essential properties of the dataset (Prieto, 2009). Another dimension involved in data reuse processes is indeed interpersonal trust, such as trust in the

individual who produced the dataset. For example, Jirotka et al.'s study of distributed readings of mammograms revealed strategies for assessing trustworthiness based on reputational familiarity with the data producer, i.e.,, whether the producer was known to produce reliable data (Jirotka, 2005). Zimmerman (2008) discussed how ecologists assess data by disciplinary standards involved in their production and by reputation of the data producer (Yakel, Faniel, Kriesberg, & Yoon, 2013).

While these studies surely inform us about the factors that scientists take into consideration when selecting a dataset for reuse, they do not say much about how research datasets are actually employed once they have been selected for reuse. We can think of filtering and selecting through sets of open data resources as the tip of the iceberg of data reuse practices. In this dissertation research project, I took a closer look at what researchers do with open data in their daily research routines. I developed a typology of reuse practices. Knowing how researchers reuse others' data for knowledge production is a good starting point for understanding the subtleties of how trust in others' data is established. The examination of such data reuse practices – and of how they may vary – allows us to identify those factors that enable a type of reuse versus another one.

The strategy behind this investigation originates from the observation that scientists reuse others' data for more than one goal. Wallis et al. initially observed this phenomenon in a study titled "If We Share Data, Will Anyone Reuse Them?" (Wallis et al., 2013). Wallis et al. explain: "Foreground data are the focus of the research, whether a field deployment or laboratory study. These forms of data are described as core or

primary data, distinct from background data that serve other purposes." Building on Wallis et al.'s initial formulation, and my empirical findings, I developed a typology of data reuse practices of the DataFace participants. This typology matters because if the scientists indeed reuse others' data in more than one way, as discussed next, different sets of sociotechnical challenges relate to different reuse practices. This typology, however, differs from Wallis et al.'s one because it does not classify data reuse practices by the type of data that is reused (background data vs. foreground data), but rather by the type of research purpose (background reuse vs. foreground reuse).

Data reuse for background research

I found that the researchers participating to this study reuse others' data in at least two distinct research practices, which relate to quite different sets of socio-technical implications. Researchers reuse others' data *daily* by accessing these at an aggregate or summary level – through data visualizations – on open databases and bioinformatics tools. Datasets reused for what I call "*background*" research purposes are highly curated (i.e., with metadata and ontologies) collections of research data. "Background" research consists – in this study – in re-using others' data to set up experiments, to annotate novel sequences, and to interpret preliminary results from statistical analyses. The bioinformatics platforms where these data are hosted have user-friendly GUIs that require not more that typing couple of simple queries in order to filter and visualize the datasets.

Trust in others' data and the "publication status"

When the DataFace researchers consult others' data for background research, for example when they visualize genome tracks on the Genome Browser, they trust the data behind the scene. There is no doubt that the fact that these datasets are highly curated and semantically and logically linked to each other is what makes them usable as "small facts" (Leonelli, 2016). However, in my observations of the DataFace participants, it emerged that the scientists trust these data better when these are known to be associated to academic publications. Most of the highly curated and integrated data hosted on platforms such as the Genome Browser, OMIM, and GEO are harvested by the database bio-curators from academic publications, a practice that – we have seen – is common among model organism communities. Increasingly, the datasets are submitted to the databases as a requirement for submission to the journals, a practice that – as discussed – started to emerge in the US in the late 80s and early 90s. Over the years, these datasets have been used and reused a number of times, by many researchers, in the context of multiple research studies. When others' data are harvested from publications, integrated in a system of organized knowledge, and reused over and over again, they constitute a corpus of known and validated knowledge about certain biological entities of interest. This validation does not solely come from the ways in which datasets are linked to each other – as discussed by Leonelli and many others – but also "through experience," by the fact that a record exists of how these datasets have been reused *before*. When a dataset is released along with a peer-reviewed publication, this means that such dataset demonstrated – at least once – to be useful and "good enough" for research purposes. This "proof of effective reuse" – along with the metadata and ontologies – constitutes a

main factor that makes the datasets highly trustworthy for the scientists participating in this study.

Data Reuse for foreground research

I use the expression "foreground research" to refer to what constitutes a full realization of the big data promise: the possibility of extracting novel knowledge from old data (boyd & Crawford, 2012). Researchers conduct foreground reuse of others' data when they statistically re-analyze the data to find new patterns that can inform novel scientific claims. In the findings section I provided an account of how a GWAS dataset collected in the context of DataFace has been reused in the context of a new experimental design, which led to a re-calculation of the genes-phenotypes correlations as they appear in the primary analysis. In the biomedical sciences, genomics datasets represent the "big data" par excellence. As the reader might recall, genomics datasets are considered "hypothesis free" data and, as such, they are increasingly made available in "raw" formats. In the context of the DataFace consortium, all datasets were made available in raw formats and prior to publications. All DataFace datasets were released with granular metadata information, but at a low level of integration.

The main difference between background and foreground reuse lies at the epistemic level: in background research others' data perform as validated "small facts," while in foreground research others' data are resources that can be used to test new statistical hypotheses. The same dataset can be used for background and for foreground, depending on the analysis that we perform on it. In Aristotle's philosophy, "actuality" and

"potentiality" are states of the same being, which is subject to motion and transformations. The concept of potentiality, in this context, generally refers to any "possibility" that a thing can be said to have. Actuality, in contrast to potentiality, is the motion, change or activity that represents the instance in which a possibility becomes real in the fullest sense. When used for background purposes, others' data perform as established facts, as fulfilled possibilities. While when used for foreground research, the same data perform as facts "in potency." For example, the findings from the primary analysis of the Caucasian GWAS dataset could be reused by a researcher conducting a literature review on those genes that have been observed as having a role in shaping the development of the human face in the embryo (background reuse). However, in order to use the Caucasian GWAS dataset itself to look for "new" genes that might be involved in other biological processes – or in the same processes – a researcher would need to conduct a secondary analysis of the genes/phenotypes associations (foreground reuse). See table 2 below.

Table 2: Features of background and foreground reuse.

	Background Reuse	**Foreground Reuse**
Goal of reuse	Knowledge contextualization: comparison and interpretation	Novel knowledge production: correlation and causation
Example of reuse	Sequence annotation	Statistical analysis
Frequency of reuse	Frequent – routine practice	Rare – emergent practice
Data processing level at moment of access	Aggregate or results' data, shared at a summary-level	"Raw" data, released at a low level of processing
Epistemic value of the data	Data represent validated knowledge – small facts	Data perform as uncharted territories – facts *in potency*

Data creator and data reuser collaboration

The DataFace participants run novel statistical analyses on others' data exclusively in collaboration with the original creators of the data. I found that there are many reasons why the data creator/data reuser collaboration is deeply valued by the scientists, but two main reasons stood up. First, the data creator/data reuser collaboration allows the data reusers to interpret the findings derived by novel analyses thoroughly. As I have discussed in few occasions already, craniofacial datasets are collected in the context of highly specialized research designs. In the craniofacial domain, often those who generate a certain dataset are those who possess the most up to date knowledge of the literature related to that field of inquiry. The data creators mastered this specialized knowledge over time, but the data reusers would need to invest a great amount of time to master it as well. Collaboration over reuse allows for the transfer of "specialized knowledge" – some of which is tacit knowledge – from the data creators to the data reusers. Second, collaborating with the data creators help the reusers to establish trust in large-scale datasets released prior to publication. When datasets have no records of reuse, the reputation of the data creator works as a substitute for the lack of publication records. When data are collected and made available for the first time, nobody in the research community knows whether they will turn out to be useful for anything. Simply put, the data creators simply have a better sense of what are the chances that the dataset will be useful for novel knowledge production.

Data Reuse and Co-authoring Papers

In a context in which raw data are made available prior to publication, the collaboration between data creators and data reusers also facilitates the process of

attributing credit to the data creators. Indeed, co-run secondary data analyses often result in the data collector and the data reuser co-authoring the research papers that comes out of the novel data analysis. In various conversations with the researchers, it emerged that the "etiquette" of co-authorship between data creators and data reusers is highly valued by these researchers, and often elevated to the status of community norm. As I have reported in my findings, in the context of the DataFace Consortium research data are exclusively collected by early researchers whose careers depend on the possibility of extracting results from the data. It is a fact that sometimes co-authorship is given simply in exchange to accessing data that are not openly available. However, nobody is forcing researchers working in open environments to give co-authorship to those researchers who collected the data, but they still do so. Among this study's participants, the heads of the laboratories approve and encourage co-authorship with the data creators. Far from being a way to receive "free credit for doing nothing," co-authorship empowers researchers to use research data as collective properties. In an open data environment, all researchers have access to most data, at any time. Contrary to early twentieth-century model organism newsletters, researchers are not required to share their own fly stocks in order to reuse others' stocks. Nor they have to show to be active members of the craniofacial community. The research community has nicknames for those who reuse others' data without giving recognition to the data creators, and without contributing their own data: these are called the "free riders" or the "data parasites." By exchanging authorship at the time of reuse, researchers promote a work environment of cooperation and reciprocal trust.

In summary, I found that DataFace researchers tend to reuse others' open data either for background or for foreground research. The scientists who participated in this study trust others' data when these are highly curated with metadata and ontologies, but also when the related primary analyses are already published in trusted academic journals. Researchers reuse others' data for background research daily, for example to annotate new statistical findings. Foreground reuse is more technically challenging because it requires scientists to conduct a new analysis that test new hypotheses on old data. Novel analyses are particularly challenging when conducted "manually," on data shared at a lower level of processing. Making "raw" data available prior to publication further complicates their reuse for foreground purposes because researchers tend to be suspicious of "unpublished data." For all these reasons, the participants in this study tend to re-analyze others' raw data in the context of a collaboration with the creators of the data. Finally, collaboration for data reuse normally results in co-authorship between the data creators and the data reusers. Via co-authorship, scientists make sure that the data creators – who are early career academics – receive credit for their labor, and also reward them for their contribution to the secondary analyses.

Socio-technical challenges to reusing DataFace open datasets

So far, my findings examined how the craniofacial researchers who participated in this study reuse others' open data in their daily practices. Given my observations, I now discuss the socio-technical challenges that the community of craniofacial researchers at large could face when reusing the datasets collected in the context of the DataFace Consortium.

Reusing DataFace open datasets for background research

As we have seen, the participants in this study reuse others' data on a daily basis to contextualize findings or set up experiments. They do so by accessing others' data at an aggregate level through easy-to-generate data visualizations. Online platforms and bioinformatics tools provide the primary means for data retrieval and consultation. Given these observations, to what extend can DataFace datasets be reused for background research?

As I discussed, during the second grant phase of DataFace, the engineering hub built a solid infrastructure that allows scientists to perform granular and complex searchers of individual datasets. The database engineers used agreed-upon metadata and ontological terms to curate the data. By collaborating with ontology experts, the hub standardized the naming process for new pieces of data. Via annotation strategies, each datasets was further linked to relevant known information such as genotypes, anatomical images, developmental stages, and phenotypes. Database engineers uploaded the DataFace tracks of sequence data on the Genome Browser, and linked these to the main DataFace platform.

At the same time, it was never the goal of the DataFace infrastructure to include visualization tools that would allow scientists to access data through a level of data integration similar to the one offered by the Genome Browser, the ExAc browser, or the OMIM and GEO databases. The development and design of the Human Genomics Analysis Interface consisted in one attempt to create such a tool. By providing summary

level visualizations for GWAS studies datasets, the Interface would have enabled scientists to visualize summary level information about individual datasets. At the time of writing this dissertation, the Interface is released in beta.

With the exception of DataFace genome tracks, which have been integrated in the UCSC Genome Browser, the DataFace platform never reached a level of data integration that would enable scientists to visualize others' data at an aggregate level. A main factor that prevented this from happening is the fact that the participants collected highly heterogeneous types of data, using different methods for data collections and analysis. As Andrea (informatician), Hank (computational biologist), and Rose (NIH officer), pointed out during our conversations (see findings section), the DataFace database was truly envisioned as a cross-searchable repository of craniofacial large-scale datasets that could be downloaded and re-analyzed by the community. In other words, the DataFace datasets are best suited for foreground research than for background research.

Reusing DataFace open datasets for foreground research

As I have already discussed, in this research community, foreground reuse of open data (i.e., running secondary analyses of others' raw sequence data) is an emergent practice. The community is fairly skeptical of high throughput data collections, especially when these are released prior to publication. The scientists have hard time trusting Unpublished in Journal, Available in Databases (UJAD) high-throughput data that haven't been explored by anyone yet, not even by the data creators. At the time the DataFace data are released, these datasets have neither the reputation, nor the history, to

speak for their value as evidence of something (to speak for their *potentiality*). Significantly, the Caucasian GWAS dataset was re-analyzed after the data creators published their primary data analysis.

The problem with reusing raw data released prior to publication is not simply a problem of trust in the data quality. Metadata and ontological terms inform scientists about the ways in which the data were collected, and about the biological entities described in the data. But metadata and ontologies cannot inform a data reuser on the potential for a certain dataset to contain novel information worth a scientific publication. For example, the RNAs expression data collected by Jane and Travis (the Green spoke) could, or could not, turned out to be relevant for understanding facial development processes in the mouse embryo. If the RNA fragments studied by Jane and Travis will be differentially expressed in the tissues embryos, the underlying data could then be re-purposed to study the emergence of craniofacial syndromes in the human embryos. But Jane and Travis will not know whether this is the case until they analyze their data. At the time these data are released, these are "virgin" of any findings. Again, when researchers conduct novel statistical analyses, research data perform as facts "in potency." In an attempt to address this issue, the hub added some links close to the datasets that redirect the users to the data creator's publication history. The question that the DataFace Consortium is now facing is whether curation "before the fact" (see chapter on tools for daa sharing and reuse) can compensate for the lack of reputation and usage history of these datasets. This is a challenge of temporal nature. The sooner the data creators will

publish their primary analyses, the sooner other researchers will be encouraged to reuse their data.

Another challenge for foreground reuse of DataFace data is related to the fact that the consortium partially addressed the issues of credit attribution and invisible labor that originated with the decision of releasing data prior to publication. As a mechanism for credit attribution, the leadership encourages data reusers to cite the consortium as a whole as the source of the datasets. However, as I have discussed, researchers belonging to this community, in their daily practices, tend to attribute credit for reuse following a quite different strategy, which is by co-authoring the final paper with the team responsible for the production of the data.

The consortium leadership actively promotes and encourages the design of "shared research questions" in between the participating labs. However, this attempt is challenged by the fact that participating scientists are already busy working on the analyses of their own datasets, and they do not have the time, the resources, and the interest in analyzing other participants' data. As I also mentioned, the leadership eventually released small grants to encourage novel analyses of DataFace data. The applicants for these small grants must be external to the Consortium and not have been involved the creation of the datasets. While this funding strategy makes sure that resources reach a higher number of laboratories, at the same time it alienates the data creators from being involved in the reuse of the data they collected and – consequently – it disrupts established collaborative practices and credit attribution strategies.

Unpredictable reuses: from craniofacial syndromes to facial reconstruction

In a regime of radical openness, research data are made available right after collection and mined as "hypothesis free" resources. The practice of data reuse is the locus of knowledge production: novel knowledge will be extracted from old data. By marketing research data as fungible communities, and actively encouraging scientists to access and reuse others' data, open repositories like DataFace increase the chances for epistemic re-negotiation, which is neither a good nor a bad thing *per se*, but it is a fact.

By "following the data," I reconstructed step by step the process that led the DataFace GWAS Caucasian dataset to be reused for training machine-learning algorithms for DNA-based facial shape prediction models. In turn, these models are informing the development of Forensic DNA Phenotyping (FDP) services. As discussed, FDP is a controversial technology. Commentators from multiple fields are skeptical of FDP's claim to be able to re-construct with high accuracy a suspect's face from a small amount of DNA left at a crime science. Since the commercialization of FDP services in 2015, law enforcement agencies started to use FDP services to generate DNA-based mug-shots of suspects, which are then routinely shared with the public.

This story of data reuse exposes the convergence of few methodological traditions that inform today's design of Forensic DNA Phenotyping (FDP) services. Metrics and techniques for classifying human beings into distinct "racial" groups have come in full circle, from being phenotypes-based, to DNA-based, to both phenotypes and DNA based. In the nineteenth-century, the discipline of craniometrics used to categorize us in distinct

races based on the measurement of their skulls. More recently, population genetics and admixture sampling techniques group individuals based on genetic markers (SNPs, AIMs, etc.). Craniofacial GWAS studies classify human beings on the basis of both facial phenotypical appearances (head measurements) and genetic markers (AIMs). The metrics and techniques developed in the context of GWAS studies, following a long tradition tracing back to the Bertillon's system of identification, ended up informing a series of studies that aim at producing novel systems of identification (FDP services, indeed).

The scientific validity of DNA-based facial reconstruction models is currently debated among the scientific community. The validity of this prediction model seems to be contested between the biomedical craniofacial community (DataFace teams), whose members study the diagnosis and treatment of rare craniofacial syndromes, and, on the other hand, the physical anthropology and computer engineering community, which is in this case specialized in human variation and prediction algorithms. I have shown how the same DataFace GWAS dataset has been analyzed and re-analyzed in two different research designs. In the primary analysis, conducted by the data creators, few genome loci were found to be associated with facial variation in humans. Among these few loci, only a couple replicated previous GWAS studies conducted on the same population (Caucasian). In the secondary analysis, the researchers have found 15 genes associated with human facial morphology.

A main difference between these two studies is the way in which the researchers designed the facial phenotypes (variation across nose, chin, jaw shapes etc.) that need to

be correlated with the genetic markers. As I have discussed, the design of the phenotypes is key in identifying variation among single facial traits. It is this variation among isolated traits that can be potentially used to reconstruct faces of individuals. In the primary analysis, the DataFace researchers constructed their phenotypes by using approximately 30 landmarks and the same amount of linear distances. However, as we have seen, for the physical anthropologists this "phenotypic first" approach potentially misses variation in the face that is expressed in traits that are not represented in a linear distance system. In the secondary analysis, the physical anthropologists' team used a machine-learning algorithm called hierarchical spectral clustering to map every inch of the human face. By using as a starting point 10.000 points on the face (as opposed to 24 landmarks), this algorithm produced a supposedly "comprehensive" representation of geometrical facial variation in humans. Based on proximity and variation, the algorithm clustered the spatially distributed 10.000 points into "facial fragments." Fragments vary from global (larger) to small (local), and are organized – in a Russian doll fashion – into 60 modules. In this way, a "global" module contains many "local" fragments. By applying this technique, which basically returns a higher number of phenotypes, the physical anthropologist's lab maximized their chances to find association between genes and facial traits. The claim is that the algorithm calculates the phenotype in a granular and objective way, as opposed to linear distances that are pre–determined by the researchers.

Given the ethical controversies surrounding FDPs technologies, and the epistemological instability of the associated prediction models for facial reconstruction, I

conclude this discussion by considering the right of the data donors (people who donated their data to biomedical research) to be informed about the reuses of their data. In relation to this issue, I also discuss the feasibility of predicting specific reuses of research data and results, once these are made openly available across different communities of practice.

Beyond privacy: the emerging politics of data reuses

In the literature review section of this dissertation I provided some examples of how the accumulation and availability of sequences and individual health data has generated a growing stock of information that can be mined to look for patterns, patterns that can be in turn used to make predictions about human health (Stevens, 2016). Large volumes of bio data are used to feed machine–learning algorithms, which are responsible for returning the prediction models. Joanna Radin observed that "machine learning and the related field of statistical pattern recognition have been the subject of increasing interest to the biomedical community because they offer the promise of improving the sensitivity and specificity of detection and diagnosis of disease, while at the same time purportedly increasing the objectivity of the decision–making process" (Radin, 2017a, p. 52). I also discussed that starting from the mid–2000s some bio–tech entrepreneurs saw a commercial opportunity for monetizing these predictions – such as in the case of the "personal genomics" services (Bliss, 2018; Mike Fortun, 2008; Nelson, 2016).

On the official website of the DataFace Consortium, the DataFace repository presents itself as "your curated, one–stop shop for facial research." The DataFace open repository constitutes a set of data resources made freely available to be mined for predictions, and

the development of FDP services by private companies is an example of how these predictions can been monetized. In her paper *Digital Natives*, Radin (2017) calls for an inclusion of the "politics of data reuse" in the discourse about human subject data reuse practices and policies. For Radin, when scientists reuse human subject data that they did not collect themselves, they should take into consideration the "histories" behind these collections. Scientists should ask themselves: Where do these data come from? How were these data collected? What were these data originally collected for? Radin's analysis highlights the fact that biomedical data comes from bodies, bodies that – in the Pima's case study analyzed by Radin – have been appropriated unethically and as a result of post–colonial dynamics.

The DataFace human subject datasets were collected and made available following procedures reviewed and approved by the Institutional Review Boards (IRBs) of all the participating labs. The parents and children who participated to the data collections were not compelled in doing so. Research subjects voluntary signed an Informed Consent form, which explains – along with the risks and benefits that come with the research participation – what researchers will use their data for. The Informed Consent is stipulated based on an overarching research protocol approved by an Institutional Review Board (IRB). To reuse DataFace human subject data, potential reusers need to submit a Data Access Request (DAR) to the Consortium. The DAR includes a detailed description of the research objectives and design. Each DAR needs also to include a research protocol approved by the requester's IRBs. If investigators plan to collaborate outside their own institution, the collaborators must submit an independent DAR using the same

project title as the one originally submitted. A novel DAR and a novel IRB need to be submitted for each reuse of the DataFace datasets. The DAR and the IRB do no guarantee access to the data. The consortium leadership is eventually responsible for approving the DAR and reviewing the IRB protocol.

Based on the DataFace policy for data access and reuse, the DataFace datasets should be employed in research studies that aim at understanding craniofacial development and disorders. But, as discussed, in practice the analysis of the DataFace datasets is already informing research well beyond these domains. The DataFace GWAS Caucasian dataset was reused in the context of a study which analytical results will advance research on the genetics causes of craniofacial birth defects. At the same time, the same data and the same analyses will also inform the design of prediction models for DNA–based facial reconstruction. This raises a question of how can IRBs and informed consents truly "inform" the data donors about the potential uses of their data when these are made available for reuse in interdisciplinary research. The more science becomes open and collaborative, the more scientists will be able to do novel and unexpected things with research data and results. When data are reused collaboratively and in an open environment, multiple teams can re–contextualize analytical results in multiple lines of research, which can have very different outcomes. Novel models for data governance are needed to make sure that those who donate their data can be properly informed about potential misuses.

Another limit of the IRBs and the inform consent is that this kind of regulations apply only to research data reused for research purposes. If research data are reused commercially, IRBs have no jurisdiction. This is also the reason why human subjects' "research data" collected on Facebook cannot be used in academic context in ways that are not officially approved by an IRB, but the same datasets can be reused in all sort of ways for marketing purposes (at least for now). Human subjects data like facial images – both from Facebook and DataFace – must be always de-identified following HIPAA rules (45 CFR 164.514). Individual-level data must not be made publicly available, also when used in commercial purposes. However, no limitations apply to the kind of service or product the data can be reused for. As I pointed out the literature review, some limitations exist for sequence data in terms of privacy protection. For example, if sequenced data are used to profile individuals, the Genetic Information Nondiscriminary Act (GINA) makes it illegal for employers and insurers to discriminate on the basis of genetic of genomic information (Hudson et al., 2008). But, not only this the law does not apply to life, disability, or long-term care insurance, it also – again – does not say anything about the types of commodities and services that can be obtained by mining the data.

7. Conclusions

The findings of this research project contribute to the socio-technical understanding of research data practices at multiple levels. First, this dissertation project informs the debate around the existence of different data governance regimes in the biomedical domain in the US. In a "semi-open regime" for data sharing, scientists deposit their research data voluntarily, once they have published their primary analyses. In a "radical

openness" regime, which emerged toward the completion of the Human Genome Project (HGP), scientists share their data in exchange for the funding for new collections, right after data production. Openness performs as an accountability mechanism, and data reuse ensures the return on the investments on data collections. In order to maximize data sharing and speed up data reuse, research data are conceptualized as "hypothesis free" resources, which are different from the "results data" shared along with peer-reviewed publications. Hypothesis-free data – in this perspective – do not need specialized labor neither to be collected nor to be analyzed. In this frame, hypothesis-free data can be exchanged as fungible resources, rhetorically separated from their creators: they can stand by themselves.

The very idea of producing hypothesis free data is troublesome for the scientists. As reported in my interviews, the participants in this study have hard time trusting research data that are marketed as being collected with no specific research design in mind. Even atlases of mouse RNA-expression data are collected to the goal of answering an initial set of questions. The participants in this study joined the DataFace Consortium to access the funding to collect large-scale datasets that they could mine to ask a set of pre-determined research questions informed by pre-established research agendas. Their overarching goal is to integrate these novel data collections with other data collections in an attempt to study the development of the human face as a system. Given these observations, the meaning of the expression "hypothesis free" is not clear. None of the DataFace datasets seemed to be collected without a clear research agenda, and a set of hypotheses, in mind.

Second, this study contributes to the debate about "what is needed" to reuse open data. My starting point was that, for research data to be reused, scientists need to trust that these data can be used as evidence of a natural phenomenon. I began this study by defying research data as "entities used as evidence of phenomena for the purposes of research or scholarship (i.e., science)" (Borgman, 2015, p. 29). We have also seen that – as observed by Leonelli (2016) – the evidentiary power of research data lies in the ways in which these are curated, or "packaged," for reuse by others than the data creators. My findings confirm the value of the data curation process as the locus of evidence in research data. DataFace scientists access others' data daily on databases and via bioinformatics tools – such as the Genome Browser, the ExAC Browser, or the OMIM database. Participants reuse these sets of curated and aggregate data to set up their experiments, annotate "raw" sequences, and interpret the results of the statistical analyses they perform on their own data. Bioinformaticians and data curators integrate open datasets in ways that can be easily visualized by the researchers at a summary-level.

However, my findings also suggest that – at least in this case study – data curation is necessary but not sufficient to enable data reuse. Other factors impact the evidentiary power of research data. One of these is the "publication status" (published vs. not published) of a dataset's primary analysis. Participants rarely reuse data that are not published along with peer-reviewed primary analyses. Before the advent of the radical openness regime, the data creators used to execute and publish the primary analysis of the data they produce. As I have discussed, if a dataset is published along with a publication, this means that – at very least once – it demonstrated its usefulness for the research

community. If a study is published in an academic journal, it means that it passed the peer-review process, but it also means that the quality of the underlying data is good enough to be used as evidence for "small biology facts" – at least theoretically. By accessing open datasets linked to peer-reviewed publications, reusers quickly assess the utility and quality of an open dataset. While metadata provide specific indications about the conditions of production, the "publication status" of the primary analysis indirectly suggests whether a dataset has the potential to be useful for research. In this sense, the publication status is the first indication of utility, the starting point for reuse. The publication of a primary analysis is not enough to enable scientists to re-contextualize others' data, that is the "job" of metadata and ontologies, but it makes the very first reuse of a dataset easier. In a semi-open regime of data sharing, the data creators were responsible for demonstrating the usefulness of the data they collected. Now, with the emergence and diffusion of the radical openness regime, this responsibility is distributed among the research community at large. In a radical openness regime, the line between primary uses and secondary uses is re-negotiated. On one hand, scientists are supposed to trust the evidentiary power of datasets that have no publication records. On the other hand, anyone could be the first to conduct a primary analysis on data collected by someone else.

Still, even when data are shared along with publications and are properly curated, there are cases in which the reusers need to collaborate with the data creators to re-analyze the data. They cannot do so independently. This is where the typology of data reuse practices comes in handy. I developed the background/foreground typology based

on Wallis et al.'s initial observation that – for the scientists – data reuse can mean more than one thing (Wallis et al., 2013). I further used it to provide an explanation of why researchers collaborate with the data creators to reuse others' data solely in certain instances. Overall, collaborating with the data creators does not seem to be necessary when data are accessed at an aggregate or summary level, and used for background. On the other hand, the practice of testing new statistical hypotheses on others' data – what I defined as "foreground reuse" – requires collaboration between the data creators and the data reusers. Let's suppose that a researcher is about to conduct a secondary data analysis of a research dataset that she did not collect herself – for example, she is about to re-analyze Jane's RNAs expression dataset of the mouse embryonic facial tissue. The dataset's metadata provide the data reuser with information about the conditions of data production, and annotations and ontologies link Jane's dataset to pre-existing datasets and other bits of information. The publication of Jane's primary analysis confirms that this dataset has already passed peer-review quality control. Given this set of meta-information about Jane's dataset, can a potential reuser run a new analysis on it? In my observations, in most cases she still cannot. Based on my observational and interview data, reusing others' data for foreground research requires the linguistic and interpersonal exchange of tacit and specialized knowledge that is not easily formalizable in metadata and ontologies. Jane is an expert on facial tissue formation processes. She has been collecting and analyzing RNAs data from mouse embryos for a long time. Jane conceptualized and designed the experimental setting in which the data were collected. When she analyzes her own data, Jane heavily relies on her specialized knowledge of her field and of her experimental procedures to interpret them. For the data reuser to gain the same amount of

tacit and specialized knowledge, it would take months of training and many hours of reviewing past and current literature on this subject. For this reason, data reusers tend to collaborate with the data creators when they indent to reuse their data for foreground research. Collaboration enables fast sharing of tacit and specialized knowledge.

Among the participants in this study, the data reusers reward the data creators with co-authorship because of the time they invested in collecting the data, and also for the specialized knowledge that they share along with their data. The conceptualization of research data as "hypothesis and labor free" resources falls short in the light of these observations. At least among the participants in this consortium – at this moment in time – the data creators collected their datasets in the context of specialized research designs, of which they are "the experts." Tools developers could question this observation by showing how certain visualization tools are able to suddenly visualize "with one click" patterns in the data that were not there before. In this perspective, data discovery tools are promoted as substitute for specialized expertise and tacit knowledge. I would reply by asking: is the knowledge retrieved by this visualization enough for a scientist to submit it in the form of a publication to a panel of peers for review? In most cases, it is not. The Genome Browser, for example, by aligning and mapping many genome tracks on the same knowledge representation schema, enables the scientists to visualize and compare multiple genome tracks at once. But in order to publish a paper worth of the top journal in the field, the scientists still need to explain "why" variance can be observed across multiple genomes in the first place, and what are the implication of such variance. The deep knowledge that is necessary to explain the meaning and impact of a certain genetic

variance is produced by consulting many other databases and by accessing specialized and tacit knowledge that the researchers accumulate over time during their careers.

Data reusers need the data creators' collaboration to re-analyze data efficiently and properly interpret them. This observation raises a set of questions about how epistemic power is distributed in data reuse practices. The data creators, on one hand, retain a certain degree of control on who can access their specialized knowledge. At the same time, the data creators are concerned that their data can be misinterpreted and misused by those who re-analyze them independently.

The case study of the reuse of the DataFace GWAS datasets is a perfect opportunity to reflect on "the politics of data reuse." We have seen that DataFace GWAS datasets have been reused in the context of a research design that aims at associating variation in human facial traits to variation in human genotypes. My examination revealed that the science behind this kind of studies is highly contested between different epistemic communities (biomedical researchers vs. computer scientists and physical anthropologists). The fact that FDPs are being employed by law enforcement agencies to search for suspects while the underlying science is not settled yet, makes their uses ethically questionable.

Given these observations, this case study further exposes the limitations that the IRB and the Informed Consent face when biomedical research data are reused across contexts of production. It highlights the fact that it is impossible to predict – and consequently

formalize – how open research data will be reused, by whom, and to what purposes. This is true in research, and even more problematic in the private sector, where there is no oversight over the products and services that research data can be reused for. We should be asking whether it is even feasible to truly inform the data donors about the ways in which research data will be reused.

It might be useful to think about the reuse of research data as a "situated action" – to use Lucy Suchman's famous expression (Suchman, 1987). The reuses of research data vary based on the research design – the situation – that informs such practice. Not all reuse situations are the same. Metadata and ontologies might be sufficient to enable reuse when research data are accessed at an aggregate or summary level, and used for comparison and control. Meta-information might not be sufficient to enable reuse when data are accessed at a low level of processing, to run novel statistical analyses. No research dataset was born "hypothesis free." Even an atlas of mouse's facial images is collected within a certain research design, in the context of a certain research agenda. When reusers who are not familiar with the research design in which data were originally collected try to independently re-analyze them, they lose bits of tacit and specialized knowledge that informed the initial collection of the data, as well as the interpretation of related statistical findings. By collaborating with the data creators, data reusers have easy and fast accessed to such knowledge.

In my observations of the data reuse practices of the DataFace Consortium, the scientists are aware of what is gained and what is lost in reusing open data independently

vs. in collaboration with the data creators. The scientists in this community seem to be mostly in favor of reusing data in collaborative settings. Collaborations are formed spontaneously in relation to pre-existing research agendas and hypotheses that the scientists are interested in testing. The biomedical researchers who participated in this study did not download and re-analyze "hypothesis free" research data from open repositories out of the blue. Beyond metadata and ontology schemas, the process of testing novel statistical hypotheses on "old" data is facilitated by the availability of primary analyses, and the existence of a shared research agenda, a shared research question, and a shared hypothesis between the data creators and the data reusers.

References

Adameyko, I., & Fried, K. (2016). The Nervous System Orchestrates and Integrates Craniofacial Development: A Review. *Frontiers in Physiology*, *7*. https://doi.org/10.3389/fphys.2016.00049

Atkins, D. E., Bindoff, N., Borgman, C. L., Ellisman, M., Feldman, S., Foster, I., … Ynnerman, A. (2010). *RCUK Review of e-Science 2009* (pp. 1–77). London: Research Councils, United Kingdom. Retrieved from http://www.epsrc.ac.uk/research/intrevs/escience/Pages/default.aspx

Atkins, D. E., Droegemeier, K. K., Feldman, S. I., Garcia-Molina, H., Klein, M. L., Messina, P., … Wright, M. H. (2003). *Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon panel on Cyberinfrastructure* (p. 84). Washington, D.C.: National Science Foundation. Retrieved from http://www.nsf.gov/cise/sci/reports/atkins.pdf

Babbie, E. R. (2012). *The Practice of Social Research* (13 edition). Belmont, CA: Cengage Learning.

Baker, K. S., Duerr, R. E., & Parsons, M. A. (2015). Scientific Knowledge Mobilization: Co-evolution of Data Products and Designated Communities. *International Journal of Digital Curation*, *10*(2). https://doi.org/10.2218/ijdc.v10i2.346

Bietz, M. J., Baumer, E. P. S., & Lee, C. P. (2010). Synergizing in Cyberinfrastructure Development. *Computer Supported Cooperative Work (CSCW)*, *19*(3–4), 245–281.

https://doi.org/10.1007/s10606-010-9114-y

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* (pp. 339–348). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/958160.958215

Biswas, S. (2015, June 18). Building the face of a criminal from DNA. *BBC News*. Retrieved from http://www.bbc.com/news/science-environment-33054762

Bliss, C. (2018). *Social by Nature: The Promise and Peril of Sociogenomics*. Stanford University Press.

Boas, F. (1922). Report on an Anthropometric Investigation of the Population of the United States. *Journal of the American Statistical Association*, *18*(138), 181–209. https://doi.org/10.2307/2277526

Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*. Cambridge, MA: MIT Press.

Borgman, C. L., Bowker, G. C., Finholt, T. A., & Wallis, J. C. (2009). Towards a Virtual Organization for Data Cyberinfrastructure. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 353–356). New York, NY, USA: ACM. https://doi.org/10.1145/1555400.1555459

Borgman, C. L., Darch, P. T., Sands, A. E., & Golshan, M. S. (2016). The durability and fragility of knowledge infrastructures: Lessons learned from astronomy. In *Proceedings of the Association for Information Science and Technology* (Vol. 53, pp. 1–10). Copenhagen, Denmark. Retrieved from http://dx.doi.org/10.1002/pra2.2016.14505301057

Borgman, C. L., Wallis, J. C., Mayernik, M. S., & Pepe, A. (2007). Drowning in Data: Digital library architecture to support scientific use of embedded sensor networks. In *Joint Conference on Digital Libraries* (pp. 269–277). Vancouver, British Columbia, Canada: Association for Computing Machinery. https://doi.org/10.1145/1255175.1255228

Bowker, G. C. (2005). *Memory Practices in the Sciences*. Cambridge, Mass.: MIT Press.

Bowker, G. C., Baker, K., Millerand, F., & Ribes, D. (2009). Toward Information Infrastructure Studies: Ways of Knowing in a Networked Environment. In J. Hunsinger, L. Klastrup, & M. Allen (Eds.), *International Handbook of Internet Research* (pp. 97–117). Dordrecht: Springer Netherlands. Retrieved from http://link.springer.com/10.1007/978-1-4020-9789-8_5

Bowker, G. C., & Star, S. L. (1999). *Sorting Things Out: Classification and Its Consequences*. Cambridge, Mass.: The MIT Press.

boyd, danah, & Crawford, K. (2012). Critical Questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Buhr, S. (2018, February 8). Human sequencing pioneer George Church wants to give you the power to sell your DNA on the blockchain. *TechCrunch*. Retrieved from http://social.techcrunch.com/2018/02/08/human-sequencing-pioneer-george-church-wants-to-give-you-the-power-to-sell-your-dna-on-the-blockchain/

Buneman, P. (2005). Curated databases. In *Data Integration in the Life Sciences, Proceedings* (Vol. 3615, pp. 2–2). Retrieved from ://000232263300002

Carlson, S., & Anderson, B. (2007). What Are Data? The Many Kinds of Data and Their Implications for Data Re-Use. *Journal of Computer-Mediated Communication*, *12*(2), 635–651. https://doi.org/10.1111/j.1083-6101.2007.00342.x

Carusi, A., Darch, P. T., Lloyd, S., Jirotka, M., de la Flor, G., Schroeder, R., & Meyer, E. (2010). *Shared Understandings in e-Science Projects: A report from the 'Embedding e-Science Applications: Designing and Managing for Usability' project*.

Caspari, R. (2003). From Types to Populations: A Century of Race, Physical Anthropology, and the American Anthropological Association. *American Anthropologist*, *105*(1), 65–76. https://doi.org/10.1525/aa.2003.105.1.65

Castells, M. (1996). *The Rise of the Network Society: The Information Age: Economy, Society, and Culture Volume I* (1 edition). Chichester, West Sussex ; Malden, MA: Wiley-Blackwell.

Chen, A. (2016). Is It Time To Stop Using Race In Medical Research? Retrieved from https://www.npr.org/sections/health-shots/2016/02/05/465616472/is-it-time-to-stop-using-race-in-medical-research

Clarke, A. E., & Fujimura, J. H. (2014). *The Right Tools for the Job: At Work in Twentieth-Century Life Sciences*. Princeton University Press.

Collins, F. S. (2011). *The Language of Life: DNA and the Revolution in Personalized Medicine* (1 edition). New York: Harper Perennial.

Collins, F. S., Morgan, M., & Patrinos, A. (2003). The Human Genome Project: lessons from large-scale biology. *Science*, *300*(5617), 286–290.

Collins, H., & Evans, R. (2008). *Rethinking Expertise*. University of Chicago Press.

Collins, H. M., Evans, R., & Gorman, M. (2007). Trading zones and interactional expertise. *Studies in History and Philosophy of Science Part A*, *38*(4), 657–666. https://doi.org/10.1016/j.shpsa.2007.09.003

Contreras, J. L. (2011). Bermuda's Legacy: Policy, Patents, and the Design of the Genome Commons. *Minnesota Journal of Law, Science & Technology*, *12*, 61.

Darch, P. T. (2014). Managing the Public to Manage Data: Citizen Science and Astronomy. *International Journal of Digital Curation*, *9*(1), 25–40. https://doi.org/10.2218/ijdc.v9i1.298

Daston, L. J., & Galison, P. (2007). *Objectivity* (1 edition). New York : Cambridge, Mass: Zone.

David, P. A., & Spence, M. (2003). *Towards Institutional Infrastructures for E-Science: The Scope of the Challenge* (Oxford Internet Institute Research Reports No. 2). Oxford: University of Oxford. Retrieved from http://129.3.20.41/eps/le/papers/0502/0502002.pdf

Day, R. E. (2014). *Indexing it all: the subject in the age of documentation, information, and data*. Cambridge, Massachusetts: The MIT Press. Retrieved from https://mitpress.mit.edu/books/indexing-it-all

De Chadarevian, S. (2004). 5 Mapping the worm's genome. *From Molecular Genetics to Genomics: The Mapping Cultures of Twentieth-Century Genetics*, 95.

Derijcke, A., Eerens, A., & Carels, C. (1996). The incidence of oral clefts: a review. *British Journal of Oral and Maxillofacial Surgery*, *34*(6), 488–494. https://doi.org/10.1016/S0266-4356(96)90242-9

Dewey-Hagborg, H. (2017). Postgenomic Identity: Art and Biopolitics. *Leonardo*, *50*(5), 531–531.

Donovan, J., Pasquetto, I., & Pierre, J. (2018). Cracking Open the Black Box of Genetic Ancestry Testing. Retrieved from http://scholarspace.manoa.hawaii.edu/handle/10125/50105

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, *4*(1), eaao5580. https://doi.org/10.1126/sciadv.aao5580

Duster, T. (2005). *Race and reification in science*. American Association for the Advancement of Science.

Duster, T. (2006). The molecular reinscription of race: unanticipated issues in biotechnology and forensic science. *Patterns of Prejudice*, *40*(4–5), 427–441. https://doi.org/10.1080/00313220601020148

Edwards, P., Mayernik, M. S., Batcheller, A., Bowker, G., & Borgman, C. (2011). Science friction: Data, metadata, and collaboration. *Social Studies of Science*, 0306312711413314.

Edwards, P. N. (2010). *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. Cambridge, MA: The MIT Press.

Edwards, P. N., Jackson, S. J., Bowker, G. C., & Knobel, C. P. (2007). *Understanding Infrastructure: Dynamics, Tensions, and Design: Report of a Workshop on History & Theory of Infrastructure, Lessons for New Scientific Cyberinfrastructures*. Washington, DC: National Science Foundation. Retrieved from http://hdl.handle.net/2027.42/49353

Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C., & Borgman, C. L. (2011). Science Friction: Data, Metadata, and Collaboration. *Social Studies of Science*, *41*(5), 667–690. https://doi.org/10.1177/0306312711413314

Emerson, R. M., Fretz, R. I., & Shaw, L. L. (2011). *Writing Ethnographic Fieldnotes, Second Edition* (Second Edition edition). Chicago: University Of Chicago Press.

Evans, P. D., Gilbert, S. L., Mekel-Bobrov, N., Vallender, E. J., Anderson, J. R., Vaez-Azizi, L. M., … Lahn, B. T. (2005). Microcephalin, a Gene Regulating Brain Size, Continues to Evolve Adaptively in Humans. *Science*, *309*(5741), 1717–1720. https://doi.org/10.1126/science.1113722

FaceBase. (2017). Craniofacial Resources Hub | FaceBase. Retrieved December 19, 2017, from https://www.facebase.org/resources/

Faniel, I. M., & Jacobsen, T. E. (2010). Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues' Data. *Journal of Computer Supported Cooperative Work*, *19*(3–4), 355–375. https://doi.org/10.1007/s10606-010-9117-8

Faniel, I. M., & Yakel, E. (2017). Practices Do Not Make Perfect: Disciplinary Data Sharing and Reuse Practices and Their Implications for Repository Data Curation. In L. R. Johnston (Ed.), *Curating Research Data, Volume One: Practical Strategies for Your Digital Repository* (pp. 103–126). Chicago, Illinois: Association of College and Research Libraries. Retrieved from http://www.oclc.org/research/publications/2017/practices-do-not-make-perfect.html

Farkas, L. G. (1994). Anthropometry of the attractive North American Caucasian face. *Head and Face*.

Farkas, L. G. (1996). Accuracy of anthropometric measurements: past, present, and future. *The Cleft Palate-Craniofacial Journal*, *33*(1), 10–22.

Ferryman, K., & Pitcan, M. (2018). Fairness in Precision Medicine. Retrieved March 21, 2018, from https://datasociety.net/research/fairness-precision-medicine/

Fortun, Michael. (2001). Mediated speculations in the genomics futures markets. *New Genetics and Society*, *20*(2), 139–156.

Fortun, Mike. (2008). Promising genomics. *Iceland and DeCODE Genetics in a World of Speculation. Berkeley*.

Friedberg, I. (2016). The Research Parasite Awards. Retrieved March 12, 2018, from http://researchparasite.com/

Frizzo-Barker, J., Chow-White, P. A., Charters, A., & Ha, D. (2016). Genomic Big Data and Privacy: Challenges and Opportunities for Precision Medicine. *Computer Supported Cooperative Work (CSCW)*, *25*(2–3), 115–136. https://doi.org/10.1007/s10606-016-9248-7

Galison, P. (1987). *How experiments end*. Chicago: University of Chicago Press.

Gannon, M. (2017, July 12). Amazing DNA tool gives cops a new way to crack cold cases. *NBC News*. Retrieved from https://www.nbcnews.com/mach/science/amazing-dna-tool-gives-cops-new-way-crack-cold-cases-ncna781946

García-Sancho, M. (2012). From the genetic to the computer program: the historicity of 'data' and 'computation' in the investigations on the nematode worm C. elegans (1963–1998). *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(1), 16–28. https://doi.org/10.1016/j.shpsc.2011.10.002

García-Sancho, M. (2015). *Biology, Computing, and the History of Molecular Sequencing: From Proteins to DNA, 1945-2000* (2012 edition). Palgrave Macmillan.

Gaudillière, J.-P., & Rheinberger, H.-J. (2004). *From Molecular Genetics to Genomics: The Mapping Cultures of Twentieth-Century Genetics*. Routledge.

Gilliland, A., & Mckemmish, S. (2006). Building an Infrastructure for Archival Research. *Archival Science*, *4*(3–4), 149–197. https://doi.org/10.1007/s10502-006-6742-6

Gitelman, L. (2013). *Raw Data Is an Oxymoron*. MIT Press.

Glaser, B., & Strauss, A. (1967). The discovery grounded theory: strategies for qualitative inquiry. *Aldin, Chicago*.

Goodman, A. A., Pepe, A., Blocker, A. W., Borgman, C. L., Cranmer, K., Crosas, M., … Slavkovic, A. (2014). Ten Simple Rules for the Care and Feeding of Scientific Data. *PLoS Computational Biology*, *10*(4), e1003542.

https://doi.org/10.1371/journal.pcbi.1003542

Griffiths, A. J., Gelbart, W. M., Miller, J. H., & Lewontin, R. C. (1999). Linkage Maps. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK21358/

Hacking, I. (2005). Why Race Still Matters. *Daedalus*, *134*(1), 102–116.

Halder, I., & Shriver, M. D. (2003). Measuring and using admixture to study the genetics of complex diseases. *Human Genomics*, *1*(1), 52–62. https://doi.org/10.1186/1479-7364-1-1-52

Hallgrimsson, B., Mio, W., Marcucio, R. S., & Spritz, R. (2014). Let's Face It—Complex Traits Are Just Not That Simple. *PLOS Genetics*, *10*(11), e1004724. https://doi.org/10.1371/journal.pgen.1004724

Hilgartner, S. (2017). *Reordering Life: Knowledge and Control in the Genomics Revolution*. MIT Press.

Hindmarsh, R., & Prainsack, D. B. (Eds.). (2010). *Genetic Suspects: Global Governance of Forensic DNA Profiling and Databasing* (1 edition). Cambridge ; New York: Cambridge University Press.

Hudson, K. L., Holohan, M. K., & Collins, F. S. (2008). Keeping Pace with the Times — The Genetic Information Nondiscrimination Act of 2008. *New England Journal of Medicine*, *358*(25), 2661–2663. https://doi.org/10.1056/NEJMp0803964

Hughes, T. P. (1983). *Networks of Power: electrification in Western society, 1880-1930*. Baltimore: John Hopkins University Press.

Jackson, S. J., Edwards, P. N., Bowker, G. C., & Knobel, C. P. (2007). Understanding infrastructure: History, heuristics and cyberinfrastructure policy. *First Monday*, *12*(6). https://doi.org/10.5210/fm.v12i6.1904

Jackson, S. J., Ribes, D., & Buyuktur, A. (2010). Exploring Collaborative Rhythm: Temporal Flow and Alignment in Collaborative Scientific Work, 1–6.

Jirotka, M., Procter, R., Hartswood, M., Slack, R., Simpson, A., Coopmans, C., … Voss, A. (2005). Collaboration and trust in healthcare innovation: The eDiaMoND case study. *Computer Supported Cooperative Work (CSCW)*, *14*(4), 369–398. https://doi.org/10.1007/s10606-005-9001-0

Jones, M. B., Schildhauer, M. P., Reichman, O. J., & Bowers, S. (2006). The New Bioinformatics: Integrating Ecological Data from the Gene to the Biosphere. *Annual Review of Ecology, Evolution, and Systematics*, *37*(1), 519–544. https://doi.org/10.1146/annurev.ecolsys.37.091305.110031

Kayser, M. (2015). Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes. *Forensic Science International. Genetics*, *18*, 33–48. https://doi.org/10.1016/j.fsigen.2015.02.003

Keller, E. F. (1984). *A Feeling for the Organism, 10th Aniversary Edition: The Life and Work of Barbara McClintock* (Anniversary edition). New York: Times Books.

Keller, E. F. (2002). *The Century of the Gene*. Cambridge, Massachusetts London: Harvard University Press.

Kelty, C. M. (2012). This is not an article: Model organism newsletters and the question of 'open science.' *BioSocieties*, *7*(2), 140–168. https://doi.org/10.1057/biosoc.2012.8

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

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

Kolar, J. C. (1993). Methods in anthropometric studies. *The Cleft Palate-Craniofacial Journal: Official Publication of the American Cleft Palate-Craniofacial Association*, *30*(4), 429–431.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). University of Chicago Press.

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. Cambridge, MA: Harvard University Press.

Leonelli, S. (2016). *Data-Centric Biology: A Philosophical Study*. Chicago, IL: University of Chicago Press. Retrieved from http://www.press.uchicago.edu/ucp/books/book/chicago/D/bo24957334.html

Lessig, L. (2001). *The Future of Ideas: The Fate of the Commons in a Connected World*. New York: Random House.

Levin, N., Leonelli, S., Weckowska, D., Castle, D., & Dupré, J. (2016). How Do Scientists Define Openness? Exploring the Relationship Between Open Science Policies and Research Practice. *Bulletin of Science, Technology & Society*, *36*(2), 128–141. https://doi.org/10.1177/0270467616668760

Lewontin, R. C. (1972). The apportionment of human diversity. In *Evolutionary biology* (pp. 381–398). Springer.

Liu, F., Van Der Lijn, F., Schurmann, C., Zhu, G., Chakravarty, M. M., Hysi, P. G., … Ikram, M. A. (2012). A genome-wide association study identifies five loci influencing

facial morphology in Europeans. *PLoS Genetics*, *8*(9), e1002932.

Lofland, J., Snow, D. A., Anderson, L., & Lofland, L. H. (2005). *Analyzing Social Settings: A Guide to Qualitative Observation and Analysis* (4 edition). Belmont, CA: Cengage Learning.

Longo, D. L., & Drazen, J. M. (2016). Data Sharing. *New England Journal of Medicine*, *374*(3), 276–277. https://doi.org/10.1056/NEJMe1516564

Mamoshina, P., Vieira, A., Putin, E., & Zhavoronkov, A. (2016). Applications of Deep Learning in Biomedicine. *Molecular Pharmaceutics*, *13*(5), 1445–1454. https://doi.org/10.1021/acs.molpharmaceut.5b00982

Margolis, R., Derr, L., Dunn, M., Huerta, M., Larkin, J., Sheehan, J., … Green, E. D. (2014). The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. *Journal of the American Medical Informatics Association*, *21*(6), 957–958. https://doi.org/10.1136/amiajnl-2014-002974

Marx, V. (2013). Biology: The big challenges of big data. *Nature*, *498*(7453), 255–260. https://doi.org/10.1038/498255a

M'Charek, A. (2005). *The Human Genome Diversity Project: An Ethnography of Scientific Practice*. Cambridge University Press.

M'charek, A. (2008). Silent witness, articulate collective: DNA evidence and the inference of visible traits. *Bioethics*, *22*(9), 519–528.

M'charek, A. (2017). Data-Face and Ontologies of Race — Cultural Anthropology. Retrieved November 18, 2017, from https://culanth.org/fieldsights/835-data-face-and-ontologies-of-race

McLellan, D. (2001, October 13). John Moore, 56; Sued to Share Profits From His Cells. *Los Angeles Times*. Retrieved from http://articles.latimes.com/2001/oct/13/local/me-56770

Mirowski, P. (2011). *Science-Mart*. Harvard University Press.

Mirowski, P., & Nik-Khah, E. (2017). *The Knowledge We Have Lost in Information: The History of Information in Modern Economics*. Oxford University Press.

Mossey, P. A., & Catilla, E. E. (2003). Global registry and database on craniofacial anomalies : report of a WHO Registry Meeting on Craniofacial Anomalies. Retrieved from http://www.who.int/iris/handle/10665/42840

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1978). *Belmont Report: Ethical principles and guidelines for the protection of human subjects of research*. Washington, DC: United States Government Printing Office.

National Institutes of Health. (2017a). *Grants for Secondary Analyses of Existing Data Sets and Stored Biospecimens (R03)*. Retrieved from https://grants.nih.gov/grants/guide/pa-files/PA-17-299.html

National Institutes of Health. (2017b). NIH Research Project Grant Program (R01). Retrieved December 20, 2017, from https://grants.nih.gov/grants/funding/r01.htm

National Institutes of Health. (2017c). Types of Grant Programs. Retrieved December 20, 2017, from https://grants.nih.gov/grants/funding/funding_program.htm#u01

National Institutes of Health (NIH). (2018). NIH Strategic Plan for Data Science. Retrieved from https://grants.nih.gov/grants/rfi/NIH-Strategic-Plan-for-Data-Science.pdf

Nature. (2014). single nucleotide polymorphism / SNP | Learn Science at Scitable. Retrieved December 30, 2017, from https://www.nature.com/scitable/definition/single-nucleotide-polymorphism-snp-295

Nature. (2017). Definition of enhancer. Nature. Retrieved from https://www.nature.com/scitable/definition/enhancer-163

Nature Chemical Biology. (2005). *Nature Chemical Biology*, *1*(2), 63. https://doi.org/10.1038/nchembio0705-63

Nelson, A. (2016). *The Social Life of DNA: Race, Reparations, and Reconciliation After the Genome* (1 Reprint edition). Beacon Press.

Next Generation Identification (NGI) | Biometrics. (2016). Retrieved March 23, 2018, from https://www.leidos.com/transportation-security/biometrics/next-generation-identification-ngi

Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., … Shendure, J. (2009). Targeted Capture and Massively Parallel Sequencing of Twelve Human Exomes. *Nature*, *461*(7261), 272–276. https://doi.org/10.1038/nature08250

NHGRI. (2000). Policy on Release of Human Genomic Sequence Data (2000). Retrieved from https://www.genome.gov/10000910/Policy-on-Release-of-Human-Genomic-Sequence-Data-2000

NIH. (2016). Commons Home Page | Data Science at NIH. National Science Foundation. Retrieved from https://datascience.nih.gov/commons

November, J. A. (2012). *Biomedical Computing: Digitizing Life in the United States*. JHU Press. Retrieved from https://books.google.com/books/about/Biomedical_Computing.html?id=-cIdzHlhGYgC

Olson, G. M., & Olson, J. S. (2000). Distance matters. *Human-Computer Interaction*, *15*, 139–178.

Open Humans. (2016). Home - Open Humans. Retrieved March 24, 2018, from https://www.openhumans.org/

Open Knowledge Foundation. (2015). Open Definition: Defining Open in Open Data, Open Content and Open Knowledge. Retrieved September 22, 2015, from http://opendefinition.org/od/

Organisation for Economic Co-operation and Development. (2007). *OECD Principles and Guidelines for Access to Research Data from Public Funding* (p. 24). Paris: Organisation for Economic Co-Operation and Development. Retrieved from http://www.oecd.org/dataoecd/9/61/38500813.pdf

Palmer, C. L., Weber, N. M., & Cragin, M. H. (2011). The Analytic Potential of Scientific Data: Understanding Re-use Value. *Proceedings of the American Society for Information Science and Technology*, *48*(1), 1–10.

Panofsky, A., & Bliss, C. (2017). Ambiguity and Scientific Authority: Population Classification in Genomic Science. *American Sociological Review*, *82*(1), 59–87. https://doi.org/10.1177/0003122416685812

Parker, S. E., Mai, C. T., Canfield, M. A., Rickard, R., Wang, Y., Meyer, R. E., … for the National Birth Defects Prevention Network. (2010). Updated national birth prevalence estimates for selected birth defects in the United States, 2004–2006. *Birth Defects Research Part A: Clinical and Molecular Teratology*, *88*(12), 1008–1016. https://doi.org/10.1002/bdra.20735

Pasquetto, I. V., Sands, A. E., Darch, P. T., & Borgman, C. L. (2016). Open Data in Scientific Settings: From Policy to Practice. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 1585–1596). San Jose, CA. https://doi.org/10.1145/2858036.2858543

Paternoster, L., Zhurov, A. I., Toma, A. M., Kemp, J. P., Pourcain, B. S., Timpson, N. J., … Smith, G. D. (2012). Genome-wide association study of three-dimensional facial morphology identifies a variant in PAX3 associated with nasion position. *The American Journal of Human Genetics*, *90*(3), 478–485.

Peer, L., Green, A., & Stephenson, E. (2014). Committing to Data Quality Review. *International Journal of Digital Curation*, *9*(1), 263–291.

Pomerantz, J., & Peek, R. (2016). Fifty shades of open. *First Monday*, *21*(5). https://doi.org/10.5210/fm.v21i5.6360

Prieto, A. G. (2009). From conceptual to perceptual reality: trust in digital repositories. *Library Review*, *58*(8), 593–606. https://doi.org/10.1108/00242530910987082

Purcell, A. (2016, December 17). Could DNA phenotyping construct a likeness of the Gold Coast rapist? *The Sydney Morning Herald*. Retrieved from http://www.smh.com.au/national/drawing-an-offenders-face-from-a-drop-of-blood-20161117-gss377.html

Qu, H.-Q., Tien, M., & Polychronakos, C. (2010). Statistical significance in genetic association studies. *Clinical and Investigative Medicine. Medecine Clinique et Experimentale*, *33*(5), E266–E270.

Rabeharisoa, V., Callon, M., Filipe, A. M., Nunes, J. A., Paterson, F., & Vergnaud, F. (2014). From 'politics of numbers' to 'politics of singularisation': Patients' activism and engagement in research on rare diseases in France and Portugal. *BioSocieties*, *9*(2), 194–217. https://doi.org/10.1057/biosoc.2014.4

Radin, J. (2017a). "Digital Natives": How Medical and Indigenous Histories Matter for Big Data. *Osiris*, *32*(1), 43–64. https://doi.org/10.1086/693853

Radin, J. (2017b). *Life on Ice: A History of New Uses for Cold Blood*. University of Chicago Press.

Ragas, J. (2018, March 5). How Will Facial Recognition Tech Change the Future? Look at Its Problematic Past. *Slate Magazine*. Retrieved from https://slate.com/technology/2018/03/with-apples-face-id-its-time-to-look-at-facial-recognition-techs-problematic-past.html

Ramello, G. B. (2005). Intellectual property and the markets of ideas. *Review of Network Economics*, *4*(2).

Reardon, J. (2017). *The Postgenomic Condition: Ethics, Justice, and Knowledge after the Genome* (1 edition). University of Chicago Press.

Rheinberger, H.-J. (1997). *Toward a history of epistemic things: synthesizing proteins in the test tube*. Stanford, Calif: Stanford University Press.

Ribes, D., & Bowker, G. C. (2009). Between meaning and machine: Learning to represent the knowledge of communities. *Information and Organization*, *19*(4), 199–217. https://doi.org/10.1016/j.infoandorg.2009.04.001

Ribes, D., & Lee, C. P. (2010). Sociotechnical Studies of Cyberinfrastructure and e-Research: Current Themes and Future Trajectories. *Journal of Computer Supported Cooperative Work*, *19*(3–4), 231–244. https://doi.org/10.1007/s10606-010-9120-0

Richardson, S. S., & Stevens, H. (Eds.). (2015). *Postgenomics: Perspectives on Biology after the Genome*. Durham: Duke University Press Books.

Roetzer, A., Diel, R., Kohl, T. A., Rückert, C., Nübel, U., Blom, J., … Niemann, S. (2013). Whole Genome Sequencing versus Traditional Genotyping for Investigation of a Mycobacterium tuberculosis Outbreak: A Longitudinal Molecular Epidemiological Study. *PLOS Medicine*, *10*(2), e1001387. https://doi.org/10.1371/journal.pmed.1001387

Roosenboom, J., Hens, G., Mattern, B. C., Shriver, M. D., & Claes, P. (2016). Exploring the Underlying Genetics of Craniofacial Morphology through Various Sources of Knowledge. *BioMed Research International*, *2016*, e3054578. https://doi.org/10.1155/2016/3054578

Rosenberg, D. (2013). Data before the Fact. In L. Gitelman (Ed.), *"Raw Data" is an Oxymoron* (pp. 15–40). Cambridge MA: MIT Press.

Ross, P. S., & McHugh, M. A. (2006). The role of evidence in establishing trust in repositories. Retrieved from http://eprints.erpanet.org/112/

Rowen, L., Wong, G. K. S., Lane, R. P., & Hood, L. (2000). Publication Rights in the Era of Open Data Release Policies. *Science*, *289*(5486), 1881–1881. https://doi.org/10.1126/science.289.5486.1881

Rung, J., & Brazma, A. (2012). Reuse of public genome-wide gene expression data. *Nature Reviews Genetics*, *14*(2), 89–99. https://doi.org/10.1038/nrg3394

Shaffer, J. R., Orlova, E., Lee, M. K., Leslie, E. J., Raffensperger, Z. D., Heike, C. L., … Weinberg, S. M. (2016). Genome-Wide Association Study Reveals Multiple Loci Influencing Normal Human Facial Morphology. *PLOS Genetics*, *12*(8), e1006149. https://doi.org/10.1371/journal.pgen.1006149

Shankar, K. (2006). Recordkeeping in the Production of Scientific Knowledge: An Ethnographic Study. *Archival Science*, *4*(3–4), 367–382. https://doi.org/10.1007/s10502-005-2600-1

Skloot, R. (2006, April 16). Taking the Least of You. *The New York Times*. Retrieved from https://www.nytimes.com/2006/04/16/magazine/taking-the-least-of-you.html

Skloot, R., & Turpin, B. (2010). The immortal life of Henrietta Lacks.

Sonis, S. T. (Ed.). (2015). *Genomics, Personalized Medicine and Oral Disease*. Springer International Publishing. Retrieved from //www.springer.com/gp/book/9783319179414

Star, S. L. (1995). *Ecologies of Knowledge: Work and Politics in Science and Technology*. Albany, NY: State University of New York Press.

Star, S. L., & Griesemer, J. (1989). Institutional ecology, "translations," and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-1939. *Social Studies of Science*, *19*, 387–420.

Star, S. L., & Ruhleder, K. (1994). Steps towards an ecology of infrastructure: complex problems in design and access for large-scale collaborative systems. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work* (pp. 253–264). New York, NY, USA: ACM. https://doi.org/10.1145/192844.193021

Stevens, H. (2013). *Life Out of Sequence: A Data-Driven History of Bioinformatics* (1 edition). Chicago: University Of Chicago Press.

Stevens, H. (2016). *Biotechnology and Society: An Introduction* (1 edition). Chicago: University of Chicago Press.

Strasser, B. J. (2011). The Experimenter's Museum: GenBank, Natural History, and the Moral Economies of Biomedicine. *The University of Chicago Press on Behalf of The History of Science Society*, *102*(1), 60–96. https://doi.org/10.1086/658657

Strasser, B. J. (2012). Data-Driven Sciences: From Wonder Cabinets to Electronic Databases. *Studies in History and Philosophy of Science Part C*, *43*(1), 85–87.

Strauss, A., & Corbin, J. (1990). *Basics of Qualitative Research: Grounded Theory Procedures and Techniques* (Second Edition edition). SAGE Publications, Inc.

Suchman, L. A. (1987). *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge, UK: Cambridge University Press.

Tavory, I., & Timmermans, S. (2014). *Abductive Analysis: Theorizing Qualitative Research*. University of Chicago Press.

Temple, R., & Stockbridge, N. L. (2007). BiDil for heart failure in black patients: The US Food and Drug Administration perspective. *Annals of Internal Medicine*, *146*(1), 57–62.

Teslow, T. (2014). *Constructing race: The science of bodies and cultures in American Anthropology*. Cambridge University Press.

The Broad Institute. (2017). ExAC Browser. Retrieved December 14, 2017, from http://exac.broadinstitute.org/

Toom, V., Wienroth, M., M'charek, A., Prainsack, B., Williams, R., Duster, T., … Murphy, E. (2016). Approaching ethical, legal and social issues of emerging forensic DNA phenotyping (FDP) technologies comprehensively: Reply to 'Forensic DNA phenotyping: Predicting human appearance from crime scene material for investigative purposes' by Manfred Kayser. *Forensic Science International: Genetics*, *22*, e1–e4.

https://doi.org/10.1016/j.fsigen.2016.01.010

Trust, W. (2003). Sharing data from large-scale biological research projects: a system of tripartite responsibility. In *Report of a meeting organized by the Wellcome Trust and held on 14–15 January 2003 at Fort Lauderdale, USA*. Wellcome Trust London.

UCLA Office of Research Administration. (2015). Office of the UCLA Human Research Protection Program (OHRPP). Retrieved November 4, 2015, from http://ora.research.ucla.edu/ohrpp/Pages/OHRPPHome.aspx

U.S. National Library of Medicine. (2017a). What are whole exome sequencing and whole genome sequencing? *Whole exome sequencing and whole genome sequencing*. Retrieved from https://ghr.nlm.nih.gov/primer/genomicresearch/sequencing

U.S. National Library of Medicine. (2017b). What is genome annotation? [NCBI Support Center]. Retrieved December 14, 2017, from https://support.ncbi.nlm.nih.gov/link/portal/28045/28049/Article/755/What-is-genome-annotation

Van Otterloo, E., Williams, T., & Artinger, K. B. (2016). The old and new face of craniofacial research: How animal models inform human craniofacial genetic and clinical data. *Developmental Biology*, *415*(2), 171–187. https://doi.org/10.1016/j.ydbio.2016.01.017

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. *PLOS ONE*, *8*(7), e67332. https://doi.org/10.1371/journal.pone.0067332

Weinberg, S. M., Raffensperger, Z. D., Kesterke, M. J., Heike, C. L., Cunningham, M. L., Hecht, J. T., … Moreno, L. M. (2016). The 3D Facial Norms Database: Part 1. A web-based craniofacial anthropometric and image repository for the clinical and research community. *The Cleft Palate-Craniofacial Journal*, *53*(6), 185–197.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018.

Yakel, E., Faniel, I. M., Kriesberg, A., & Yoon, A. (2013). Trust in Digital Repositories. *International Journal of Digital Curation*, *8*(1), 143–156. https://doi.org/10.2218/ijdc.v8i1.251

Yang, W., Wu, G., Broeckel, U., Smith, C., Turner, V., Haidar, C., … Relling, M. (2016). Comparison of genome sequencing and clinical genotyping for pharmacogenes. *Clinical Pharmacology & Therapeutics*, *100*(4), 380–388. https://doi.org/10.1002/cpt.411

Yoon, A., & Kim, Y. (2017). Social scientists' data reuse behaviors: Exploring the roles of attitudinal beliefs, attitudes, norms, and data repositories. *Library & Information Science Research*, *39*(3), 224–233. https://doi.org/10.1016/j.lisr.2017.07.008

Young, N. M., Hu, D., Lainoff, A. J., Smith, F. J., Diaz, R., Tucker, A. S., … Marcucio, R. S. (2014). Embryonic bauplans and the developmental origins of facial diversity and constraint. *Development*, *141*(5), 1059–1063. https://doi.org/10.1242/dev.099994

Zhang, S. (2017, September 22). The Genomic Revolution Reaches the City Crime Lab. *The Atlantic*. Retrieved from https://www.theatlantic.com/science/archive/2017/09/next-generation-dna-sequencing-forensics/540603/

Zimmerman, A. S. (2007). Not by metadata alone: the use of diverse forms of knowledge to locate data for reuse. *International Journal on Digital Libraries*, *7*(1–2), 5–16. https://doi.org/10.1007/s00799-007-0015-8

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), 631–652. https://doi.org/10.1177/0162243907306704