UNIVERSITAT
RAMON LLULL

# At the crossroads of big science, open science, and technology transfer

Laia Pujol Priego

http://hdl.handle.net/10803/669220

C.I.F. G: 59069740 Universitat Ramon Llull Fundació Rgtre. Fund. Generalitat de Catalunya núm. 472 (28-02-90)

# DOCTORAL THESIS

| | |
|---|---|
| Title | At the crossroads of big science, open science, and technology transfer |
| Presented by | Laia Pujol Priego |
| Centre | Escola Superior d'Administració i Direcció d'Empreses ESADE |
| Department | Operations, Innovation and Data Sciences |
| Directed by | Dr. Jonathan Wareham |

*I dedicate this PhD thesis to my husband David and my two daughters: Ona and Maria*

# Acknowledgements

I could not imagine a better person to walk me through this journey: thank you, Jonathan. You have been a real mentor to me and you always will be. Thanks for trusting me and making me aim high. I have grown intellectually, analytically, and personally with you. You have taught me how to *think*, but also how to *be*. I have learned in every discussion, trying to answer your questions, but also observing you. You have taught me to interrogate our analysis with honesty, to dissect every argument, and to strive to say something meaningful, interesting, and good enough. These have been very intensive years and I can only feel very honoured to have spent them close to you. Thanks for all this and much more.

I wish to thank the members of the dissertation committee for generously offering their time and valuable comments throughout their review of this document.

Thanks, Markus Nordberg, Pablo Garcia Tello and all the people I had the privilege to meet at CERN for reminding me why I do research and how I want to do it. You are a real inspiration for me. I would also like to thank Javier, Erik, Maciej, Eduardo, Cesar, Felipe, and all White Rabbit community and extended family. You have been a fundamental piece in my journey and I feel very privileged to have been able to observe the amazing job you did. I would like to thank Jessica Vamathevan, David Hulcoop, and all the Open Targets people for being so generous with me and sharing the incredible and difficult work you are doing.

I would like to thank Antonio Dávila for sharing his passion for teaching and research, and for believing in me. This journey started close to you and I will always be grateful for your wise counsels.

I wish to express my gratitude to Paul Almeida for his support and for making possible such an intellectually growing experience at Georgetown. This was a year I will always remember.

I am also grateful to Sabine Brunswicker for trusting me, for teaching me that results come after hard work, and for motivating me to strive for excellence.

Thanks to Jean-Claude Burgelman for directing my attention towards a hot policy-relevant phenomenon that we are still striving to understand.

Special thanks to Vicenta, Pilar, Silvia, and François for making this journey possible, as well as to Laura Castellucci and Victoria Cochrane for supporting me and making things simpler and easier along this process. Thanks for your professionalism but most importantly for your kindness.

I cannot forget to say thank you to the amazing professors of the MRes who guided me with huge amounts of generosity. Thanks to Professor Eduard Bonet for inspiring me with your narratives and for your kind support. I will always be grateful to Joan Manuel Batista who seeded the statistical pillars where I built my analytical thinking. I promise you not to follow the rituals.

Finally, I wish to thank Ferran and Kobi for their treasured support, to my dearest friends for always being there, to my family who taught me that happiness is for the brave, to my daughters whose love makes us invincible, and to my husband David for sharing with me the passion for what we do, for his unbroken patience, for believing in this adventure, for his endless love.

*This research has been financially supported by SUR (Secretaria d'Universitats i Recerca) of the DEC (Departament d'Economia i Coneixement) of the Government of Catalonia and the European Social Fund (grants 2017FI_B 00345; 2018FI_B1 00176; 2019FI_B200169).*

*The research mobility has been partially supported by Spanish Ministry of Education, Culture and Sports (CEI15-31) and Campus of International Excellence Aristos Campus Mundus (ACM2016_Movilidad_10).*

# Abstract

Big science infrastructures are confronting increasing demands for public accountability, not only within scientific discovery but also their capacity to generate secondary economic value. To build and operate their sophisticated infrastructures, big science often generates frontier technologies by designing and building technical solutions to complex and unprecedented engineering problems. In parallel, the previous decade has seen the disruption of rapid technological changes impacting the way science is done and shared, which has led to the coining of the concept of Open Science (OS). Governments are quickly moving towards the OS paradigm and asking big science centres to "open up" the scientific process. Yet these two forces run in opposition as the commercialisation of scientific outputs usually requires significant financial investments and companies are willing to bear this cost only if they can protect the innovation from imitation or unfair competition. This PhD dissertation aims at understanding how new applications of ICT are affecting primary research outcomes and the resultant technology transfer in the context of big science and OS. It attempts to uncover the tensions in these two normative forces and identify the mechanisms that are employed to overcome them. The dissertation is comprised of four separate studies: 1) A mixed-method study combining two large-scale global online surveys of research scientists (2016, 2018), with two case studies in high energy physics and molecular biology scientific communities that assess explanatory factors behind scientific data sharing practices; 2) A case study of Open Targets, an information infrastructure based upon data commons, where the European Molecular Biology Laboratory-EBI and pharmaceutical companies collaborate and share scientific data and technological tools to accelerate drug discovery; 3) A study of a unique dataset of 170 projects funded under ATTRACT—a novel policy instrument of the European Commission led by European big science infrastructures—which aims to understand the nature of the serendipitous process behind transitioning big science technologies to previously unanticipated commercial applications; and 4) A case study of White Rabbit technology, a sophisticated open-source hardware developed at the European Organization for Nuclear Research (CERN) in collaboration with an extensive ecosystem of companies.

**Keywords**: big science, open science, open-source hardware, data commons, information infrastructures, transaction costs economics, collective action, serendipity, epistemic cultures.

# Table of Contents

# List of Figures

## 1. Introduction

## 2. Overarching Theoretical Framework

## 3. The Stickiness of Scientific Data: Epistemic cultures and a collective action dialogue

## 4. Opaque Spaces of the Commons: Governing Information Infrastructures in Life Science

## 5. Serendipity in Big Science Infrastructures

## 6. From Bits to Atoms: White Rabbit at CERN

## 7. Discussion and conclusion

# List of Tables

## 1. Introduction

## 2. Overarching Theoretical Framework

## 3. The Stickiness of Scientific Data: Epistemic cultures and a collective action dialogue

## 4. Opaque Spaces of the Commons: Governing Information Infrastructures in Life Science

## 5. Serendipity in Big Science Infrastructures

## 6. From Bits to Atoms: White Rabbit at CERN

## 7. Discussion and conclusion

# 1

# Introduction

This chapter introduces the topic of the PhD thesis, and presents its structure and content

## 1.1 Introduction to the topic of the PhD thesis

***Big science***, defined as large-scale and capital-intensive scientific collaborative efforts (Weinberg 1961), has provided societies with frontier technologies that have impacted businesses, markets, and people's lives. One major characteristic of these infrastructures is that they cannot use off-the-shelf technologies to conduct their experiments and measurements. Instead, they require unique solutions to unprecedented engineering problems that severely challenge technology suppliers and thereby serve as drivers of innovation. Famous examples of technologies impacting business are the World Wide Web (specifically HTTP, URL, HTML), the capacitive touch screen conceived at first for mastering the controls of the Super Proton Synchrotron at the European Organization for Nuclear Research (CERN), and also the detection, imaging, and computational technologies developed for advanced scientific measurement and analysis which have demonstrated tremendous potential for many industries such as advanced manufacturing, medical diagnostics and imaging, biotechnology, and microelectronics (Bressan 2014a, 2014b).

The tremendous potential of big science centres to innovate has equally not gone unnoticed by policymakers who, after the "carte blanche" attitudes of early big-science endeavours (Autio, Bianchi-Streit et al. 2003) and the softening of their geopolitical ethos (Hellström and Jacob 2012; Weinberg 1961, 1963, 1964), have increasingly demanded a broader higher return on investment via commercialisation of their technologies and research outputs (normative *vector* 1: technology transfer) (Autio 2014; Autio et al. 2004; Autio, Hameri et al. 2003; Castelnovo et al. 2018; Hallonsten 2014; Heidler and Hallonsten 2015). Although there is extensive literature describing big science's technological contribution to business and society, most of those studies have treated these organisations as a "black box", just as "sources of spin-offs, licenses and new knowledge" (Autio et al. 2004 p. 107). Little is known about how big science infrastructures can actively cultivate the transfer of their technologies to unanticipated and "outside" applications to fulfill the public policy plea (Autio 2014; Autio et al. 2004).

In parallel, the previous decade has seen the disruption of rapid technological changes, primarily in ***Information and Communication Technologies*** (ICT), impacting the way science is done and shared. Several terms have appeared to mark the impact of ICT in science: e-science (Crowston et al. 2008, 2009; David and Spence 2003; Stockinger 2005), the fourth paradigm in scientific discovery (Atkins et al. 2003; Hey 2009), and cyberscience or science 2.0 (Borgman 2010; Edwards 2019). The scientific process has long been one of the leading application areas of ICT and the rapid technological evolution in the last decade has been transforming science with the emergence of new research methods that capitalise on advanced computational resources, distributed infrastructures that support long-term sharing and reusing of data collections and scientific instruments. Data intensity, powered by computational hardware, software and research processes, is allowing scientists to carry out experiments at unprecedented levels in terms of scale and volume (Borgman 2010; Dougherty and Dunne 2011; Hey 2009). Social computing is also enabling new behaviours in communication and collaboration among scientists, where researchers are using multiple tools to share elements of their research: from literature reviews (e.g. Zotero[1]), to data (e.g. Figshare[2]) or their electronic lab notebooks (e.g. Scinote[3]) (OECD 2015)

Within such a framework, opening up and sharing data, code, scientific experimental devices, and any primary research output has become increasingly important. This phenomenon has led to the coining of the concept of ***Open Science*** (OS), which describes an approach to research based on greater access to any primary outputs of research with minimal restriction while fostering broader collaboration and transparency through all the stages of the scientific process (David 2003; OECD 2015). In this context, opening up and sharing data through data infrastructures, code (open-source software), engineering tools (open-source hardware), notes (electronic lab notebooks) and any possible primary research output have become increasingly crucial in particular in big science settings which have the policy mandate to widely disseminate all possible scientific outputs generated (Atkins 2003; Borgman 2015; European Commission 2014, 2019; OECD 2015). Governments are quickly moving towards the OS paradigm and asking public research institutions and big science centres to "open up" the scientific process—often making these practices requisite for continued funding (*vector 2*: open science) (European Commission 2014).

Yet these two forces run in opposition: Big science centres must negotiate a ***tension*** between their goal of generating revenue and economic stimulus via transferring their technologies to the market (*vector* 1: technology transfer), with the additional plea by policymakers of increasing openness in the way science is performed and diffused (*vector* 2: open science). The commercialisation of scientific outputs usually requires significant financial investments and companies are willing to bear this cost only if they can protect the innovation from imitation or unfair competition (Caulfield et al. 2012; David 2003; Dosi et al. 2006; Perkmann and Schildt 2015).

Hence, the **overarching goal** of this PhD dissertation is:

**(1)** *To understand the tension between the two normative forces that big science infrastructures face (i.e. technology transfer and open science (OS)) by uncovering the mechanisms that are employed to overcome the challenges that lie at the root of such tension.*

To understand this tension requires understanding the macro phenomenon of each of the vectors (i.e. open science and technology transfer) as external forces that at times can work contradictorily. Hence, we seek:

[1] About Zotero: https://www.zotero.org/

[2] About Figshare: https://figshare.com/

[3] About Scinote: https://scinote.net/

**(1.1.)** *To understand the dynamics behind the aim of opening up primary research outputs (i.e. open science vector).*

**(1.2.)** *To understand the dynamics of steering big science activities towards transferring their technological solutions to previously unanticipated commercial applications (i.e. technology transfer vector).*

Understanding the dynamics of the exogenous influence of the openness vector (1.1) and the technology transfer vector (1.2.) will inform our goal (1) to understand the tension between the two that at times can work contradictorily, and to elucidate the specific mechanisms that organisations use to reconcile the tensions caused by the two vectors.

## 1.2 Structure

This PhD thesis adopts the form of four studies, all written for publication. Each of these four studies responds to a step in the research strategy to accomplish the aforementioned overarching research goal of understanding the dynamics and tension within and across open science and technology transfer vectors and uncovering the mechanisms that are employed to overcome them.

- Chapter 2 presents the overarching framework in which the four studies are developed and introduces the research gaps and research questions that each study addresses. It provides an overview of the four complementary studies.
- Chapter 3 is the first of the four studies, which aims at understanding the first vector: open science, responding to the first sub-goal (1.1.) of our PhD investigation. In particular the study will shed light on the dynamics behind opening up scientific data and the explanatory factors behind the gradual and disparate adoption of data sharing practices across scientists. With this purpose, the study engages in a mixed-method design combining survey data collected in 2016 (n=1,162) and 2018 (n=1,029) and qualitative data from two case studies sequentially sampled of two information infrastructures within two scientific communities (i.e. high-energy physics and molecular biology). The study draws upon the notion of "epistemic cultures" originated from the sociology of science and a collective action theory perspective to understand the incentives and deterrents that scientists confront when considering contributions to the collective goods of data sharing (i.e. data commons).
- Chapter 4 is the second empirical study, which consists of a micro-study of a single case that gives us insight into the different mechanisms that help reconcile the main tensions between the first exogenous influence presented (i.e. open science) and technology transfer. The study analyses the governance processes in the development of an information infrastructure based upon *data commons* in the big science field of molecular biology. The study examines the exemplary case of

Open Targets (OT), a large-scale information infrastructure created by leading organisations in bioinformatics, genomics, and pharmaceuticals that include for-profit companies, non-profit foundations, and public research organisations. Under collective action theoretical lenses, the study theorises about the governance conditions of modularity and brokerage that enable a fluid process of transitioning between open and opaque spaces of work in the development of the information infrastructure. This fluid dynamic helped navigate many of the trade-offs between private and collective interests in the development of shared resource pools composed of heterogeneous members with different objectives.

- Chapter 5 is the third study, which aims at understanding the second vector: technology transfer, responding to the second sub-goal (1.2.) of our PhD investigation. The study seeks to understand the nature of the serendipitous process behind transferring big science technologies to alternative and previously unanticipated commercial applications by looking at the modes towards its realisation. Leveraging a unique dataset of 170 projects funded under ATTRACT,[4] a novel policy instrument of the European Commission aiming to harness the detection and imaging technologies of the leading European research infrastructures towards entrepreneurship, the study uncovers four serendipity modes showing the potential of directed interventions enabling organisations to find unexpected commercial applications of big science research.
- Chapter 6 is the fourth study which consists of a micro-study of a single case that gives us insight into the different operational levers that help reconcile the main tensions between the two exogenous influence presented (i.e. open science and technology transfer). In particular, the study will assess the development of White Rabbit (WR), an OSH initiated at CERN and deployed as a powerful precision and synchronisation technology in many industrial settings where time accuracy is critical. Through the investigation of WR, the study contributes to recent conceptualisations of digital objects by uncovering the differences from hybrids to purely non-material digital objects and elucidates what happens when we transpose the open-source model of development to a hybrid object. As a lens to understand how different attributes of objects require different development models, we adopt relevant constructs from Transaction Costs Economics (TCE) and examine its utility as a predictive theory of open source hardware development.

[4] The members of ATTRACT are as follows: the European Organization for Nuclear Research (CERN), European Molecular Biology Laboratory (EMBL), European Southern Observatory (ESO), European Synchrotron Radiation Facility (ESRF), European X-Ray Free-Electron Laser Facility (European XFEL), and the Institut Laue-Langevin (ILL), Aalto University, Esade Business School, and the European Industrial Research Management Association (EIRMA).

- Chapter 7 integrates the main conclusions, theoretical and practical contributions, limitations, and suggestions for future research from the four articles.

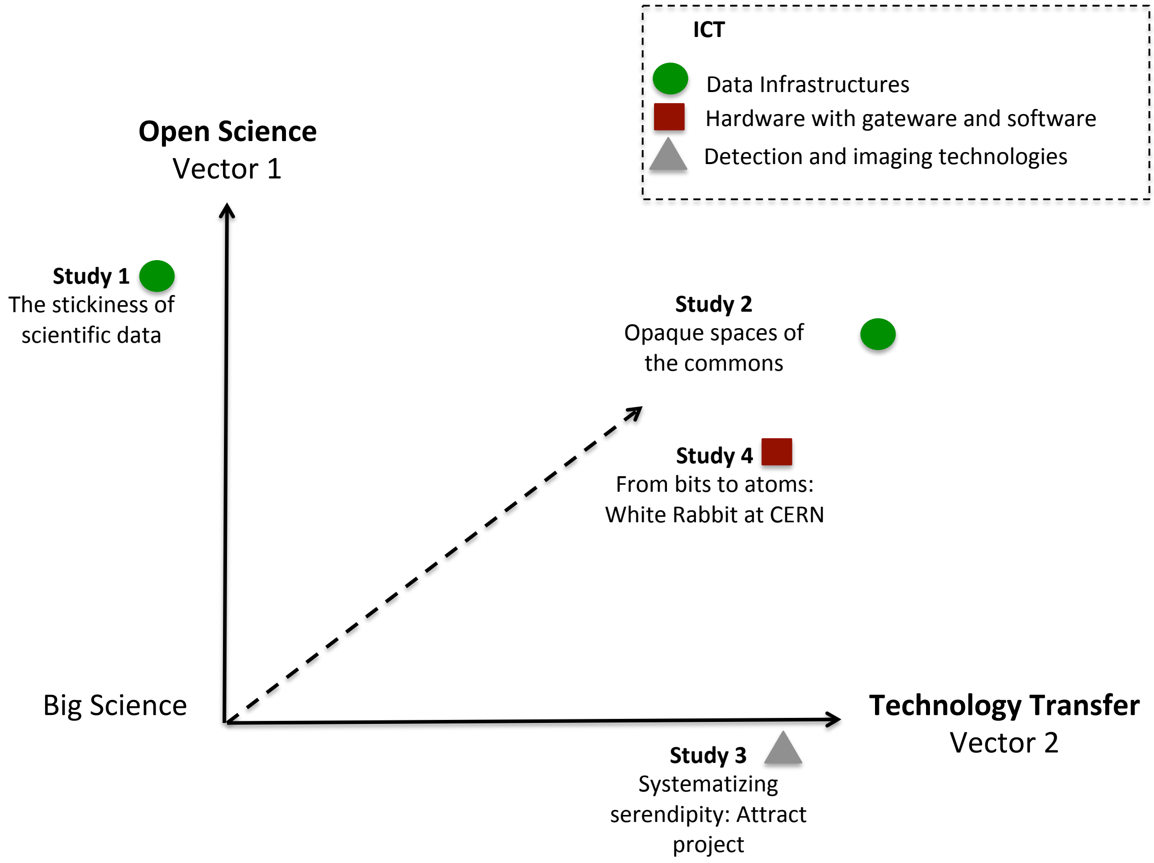

ICT
● Data Infrastructures
■ Hardware with gateware and software
▲ Detection and imaging technologies
Open Science
Vector 1
Study 1
The stickiness of scientific data
Study 2
Opaque spaces of the commons
Study 4
From bits to atoms: White Rabbit at CERN
Big Science
Technology Transfer
Vector 2
Study 3
Systematizing serendipity: Attract project

*Figure 1. Overview of the four studies integrating the PhD dissertation*

# 2

## Overarching Framework

This chapter discusses the phenomenon under investigation, related constructs and literatures, presents the research question, offers an overview of the four empirical studies that constitute chapters 3, 4, 5, and 6, and their theoretical foundations.

## 2.1 Big science

First coined by physicist Alvin Weinberg (1961), big science is a term used to refer to research organisation with high capital intensity, long-lasting facilities or networks, operating in monopoly or oligopoly conditions, and affected by externalities that produce social benefits via the generation of new knowledge, either pure or applied (Bozeman 2000). While big science was initially devoted to nuclear physics and astronomy, it has spread to other disciplines such as molecular biology, where scientists are tapping into data resources and computational infrastructures.

Under the label of big science infrastructures, there is (alphabetically): CERN, European Molecular Biology Laboratory (EMBL), European Synchrotron Radiation Facility (ESRF), European Southern Observatory (ESO), Institut Laue Langevin (ILL), Joint European Torus (JET), International Thermonuclear Experimental Reactor (ITER) and the now-completed Human Genome Project.

## 2.2 The role of ICT in big science

Science, and in particular big science (Au 2014; Borgman 2015; Hey 2009), has been one of the leading application domains of ICT. Multiple constructs have emerged to describe how ICT has transformed scientific research (e.g. eScience, eResearch, cyberscience or science 2.0) and the supporting systems that emerged to assist such transformation (e.g. cyberinfrastructures). Big science research could not be understood today without high-performance computing supporting the analyses of large volumes of data or without the diverse internet-enabled applications affording scientists access to a variety of resources including other scientists' workflows. Supercolliders, telescopes and a diverse set of large instruments are operated by large distributed research teams employing a wide range of ICT applications.

The fundamental nature of the data produced in such big science infrastructures has also changed: More than 200 petabytes of data are now permanently archived in CERN's tape libraries, which come from the particles collided in the Large Hadron Collider (LHC) detectors that generate about one petabyte of collision data per second (Gaillard 2017); more than 120 petabytes are archived in EMBL-EBI, which have experienced a deluge of biological data after the completion of the human genome in 2003 (Cook et al. 2018); and large volumes of data coming from remote sensing in satellites have revolutionised environmental sciences.

Within such a context, while seeking to accomplish their primary goal of conducting groundbreaking scientific research, big science centres face two exogenous demands that affect how they do it. First, they are requested to generate secondary socio-economic benefits and returns on investment by transferring the technologies developed for their scientific experimentation to other industrial settings. Second, they need to maximise the dissemination of their primary research outputs by disclosing with minimal restrictions their data, code, and the design of their engineering tools (OSH) so that others can re-use them.

## 2.3 Two normative vectors

### *2.3.1 First normative vector: Open Science*

Whereas the origins of OS are rooted in the norms of science articulated by sociologist Robert Merton (1973) who stressed the co-operative character of inquiry, the current developments of ICT transforming scientific practices have led to an emergent approach to research that reduces the barriers to sharing any form of research output, methods or tools at any stage of the research process (Friesike et al. 2015). The expression is used as an umbrella term that encapsulates open access to publications, open research data, open-source software, open-source scientific hardware, open distributed collaboration, open peer review, and citizen science.

OS was one of the clear political priorities of Commissioner Moedas. In 2014, the European Commission launched a public consultation about OS, which, in 2015, resulted in a policy agenda to foster it in Europe. This policy engagement led to the launch of the Open Science Cloud, a federated data infrastructure with cloud-based services to offer the scientific community an open environment for storing, sharing, and re-using scientific data, and the implementation of several actions contained in the Amsterdam Call for Action on OS. Also, in the United States, the Federal Crowdsourcing and Citizen Science Act were signed into law in January 2018. The requirements from funding agencies have incorporated the mandate of opening up research data and making it publicly available: US National Institutes of Health (NIH) in 2003 for grants over $500,000 (NIH 2003), the National Science Foundation (NSF) in 2010 (Borgman 2012), and the European Commission for Horizon 2020 program in 2014 (European Commission 2014).

Governments are quickly moving towards the OS paradigm and asking public research institutions and big science centres to "open up" their processes with particular attention to their data, meaning to make it freely available for other scientists to reuse. However, only recently has the literature started capturing the factors inhibiting scientific data sharing, suggesting that it imposes increased costs on scientists and their institutions without commensurate professional benefits (Borgman 2015; Edwards 2019; Edwards et al. 2011;

Tenopir et al. 2015; Wallis et al. 2013). Considering the tensions between policymakers and funding agencies' efforts to foster data sharing and the apparent barriers to its wide adoption, *we lack an understanding of the multifaceted and complex dynamics behind the normative force of sharing research data and the explanatory factors behind the drivers for and barriers to sharing research data*.

### 2.3.1 *Second normative vector: Technology Transfer*

Since the 1970s, public research institutions have faced demands for greater accountability of public spending, which have only grown in the current climate of budgetary austerity. In particular, big science accounts for a large proportion of publicly-funded research. With the conclusion of the Cold War, the direct link between big science, nuclear physics, and government expenditures on defence programs decreased. As a result, big science infrastructures faced more vigorous appeals to demonstrate their social value, not only in scientific discovery but also for the economy and society in general (Autio 2014; Autio et al. 2004; Schmied 1982).

Often associated with numerous technological innovations gestated during WWII, such as radar and wireless communication, big science infrastructures generate frontier technologies by severely challenging technology suppliers with sophisticated engineering problems that require never-seen technologies. By conducting experiments and measurements with unprecedented technological specifications, big science cannot use off-the-shelf technologies, thereby serving as an incredible driver of innovation, while significantly advancing the technical and organisational capacities of technology suppliers. Besides the immediate applications within experimentation and instrumentation, many of these technologies find alternative applications that were not part of their original scope within the scientific facility. Famous examples of research technologies gestated in big science centres impacting business are the World Wide Web (specifically HTTP, URL, HTML) at CERN, and also the detection, imaging, and computational technologies developed for advanced scientific measurement and analysis in the framework of such infrastructures which have demonstrated tremendous potential for many other industries such as advanced manufacturing, medical devices and imaging, biotechnology, and microelectronics.

The tremendous potential of big science centres to innovate has equally not gone unnoticed by policymakers who, after the "carte blanche" attitudes of early big-science endeavours (Autio, Bianchi-Streit et al. 2003), have increasingly demanded a broader higher return on investment via commercialisation of their technologies and research outputs (Guston 2000). However, the difficulties of transitioning these technologies and technical knowledge from the big science setting to "outside" their organisational setting, i.e. technology transfer (Bozeman 2000), are substantial. The primary goal of such big science infrastructures is to drive cutting-edge scientific research. Neither their cultures nor their governance is optimised for technology commercialisation. As such, different efforts have been put in place to provide the demand-side pull on these frontier technologies. For instance, most big science infrastructures set up specific structures, such as technology transfer offices (TTOs) (Siegel et al. 2003), as well as internal protocols and policies that seek to foster business collaboration. For example, the European Molecular Biology Laboratory (EMBL) created EMBL Enterprise Management Technology Transfer (EMBLEM), an affiliate and the commercial arm of the EMBL in 1999; CERN set up the Knowledge Transfer Group in 1997, which provides active service to CERN by managing and advising on all activities related to technology transfer. Moreover, since 2012, the organisation has set up business incubation centres (BICs) throughout its Member States (nine at present) to support entrepreneurs in taking CERN technologies and know-how to market.

While big science is famous for its capacity to bring new technologies to society in applications previously unanticipated, yet *there is a limited amount of rigorous empirical research on the nature of the serendipity behind such process*, which refers to a broad, multifaceted phenomenon related to the unanticipated discovery of something beneficial. Big science infrastructures are often treated as “black boxes” from which studies only grasp the outputs of the serendipity process by counting licenses or spin-offs (Autio 2014; Autio et al. 2004), but we lack knowledge on how such infrastructures can proactively realise such serendipity process (Autio 2014; Autio et al. 2004; Autio, Hameri et al. 2003; Castelnovo et al. 2018; Hallonsten 2014; Heidler and Hallonsten 2015).

Extant literature on serendipity has mostly been based on small-sample or anecdotal examples of scientific discoveries, and has mainly focused on the individual scientists' experiences as opposed to a more systemic level of analysis (Autio 2014). Hence, *questions remain around how to move towards the technology transfer vector and how to shift serendipity towards proactively finding market applications for big science.* There is a need for studies that put forward empirical examinations of serendipity to understand the dynamics that lie at the root of policy demands for leveraging big science technologies into market applications.

## 2.4 Managing the tension between technology transfer and open science

While OS promises to enhance the efficiency and quality of research by lowering data collection costs and fostering collaboration throughout the research process, it is unclear in the literature how a deliberate decision to share the scientific process openly with no Intellectual Property (IP) restrictions may affect the commercial exploitation of research outputs (Caulfield et al. 2012; David 2003, 2004; Dosi et al. 2006; Perkmann and Schildt 2015).

One of the most famous and unprecedented examples reflecting such friction between openness and economic returns via commercialisation of technologies was the World Wide Web when Tim Berners- Lee on April 30, 1993, convinced managers at CERN to place it in the public domain and make the IP freely available to everyone. By accepting this, CERN effectively agreed not to draw revenues or economic value from it. The tension emerges in other examples, for instance in the Super Proton Synchrotron (SPS), which came on-stream in 1976. It was the first accelerator to have a computerised control system. At that time, mastering the controls of the big new accelerator required technological ingenuity that led to the invention of the world's first capacitive touch screen. To develop the technology at that time, CERN worked with one of its suppliers. The development of the technology involved new techniques for metallisation on various substrates, which was the object of patent rights. However, when Bent Stumpe, CERN's scientist, was asked to sign a nondisclosure agreement, he refused, arguing that all inventions at CERN should be open. The supplier at this time was interested but unable to invest in the project unless the organisation could commit itself not to disclose the technology to third parties. As a consequence, CERN's involvement with the further development of touch screens ended and these technologies were put on hold and reinvented and brought to market in many applications by other players years later around the world (World Intellectual Property Organization, 2010).

Finally, the case of the Human Genome Project also provides an example from the life science field, when a global publicly-funded consortium, challenged by a parallel commercial effort, decided to open up all draft sequences of genes and made them available to everybody (Shreeve 2005). Had commercial pressure dominated, this could have led to a global “genome gold-rush” (Boulton et al. 2012).

These examples reflect the inherent *tension* regarding effects that greater openness in the scientific process powered by ICT may pose to the financial protection and exploitation of the technologies resulting from scientific activity (Dosi et al. 2006; Perkmann et al. 2013; Perkmann and Schildt 2015; West 2008). The commercialisation of scientific outputs usually requires significant investment, and companies are only willing to bear this cost if they can protect such outputs from imitation or unfair competition (Ågerfalk et al. 2015; Ågerfalk and Fitzgerald 2008).

## 2.5 Research question

Hence, the ***overarching research question*** guiding my dissertation is:

*How are new applications of ICT affecting research processes and the resultant technology transfer in the context of big science and open science?*

To answer our research question we will interrogate the two forces in isolation: open science (study 1) and technology transfer (study 3) and will explore through two single-case studies the different operational levers that help reconcile the main tensions between these two exogenous forces. In particular, we selected two of the constructs under the OS umbrella for their empirical prominence, significant impact on how businesses collaborate with big science and their theoretical relevance for the IS discipline:

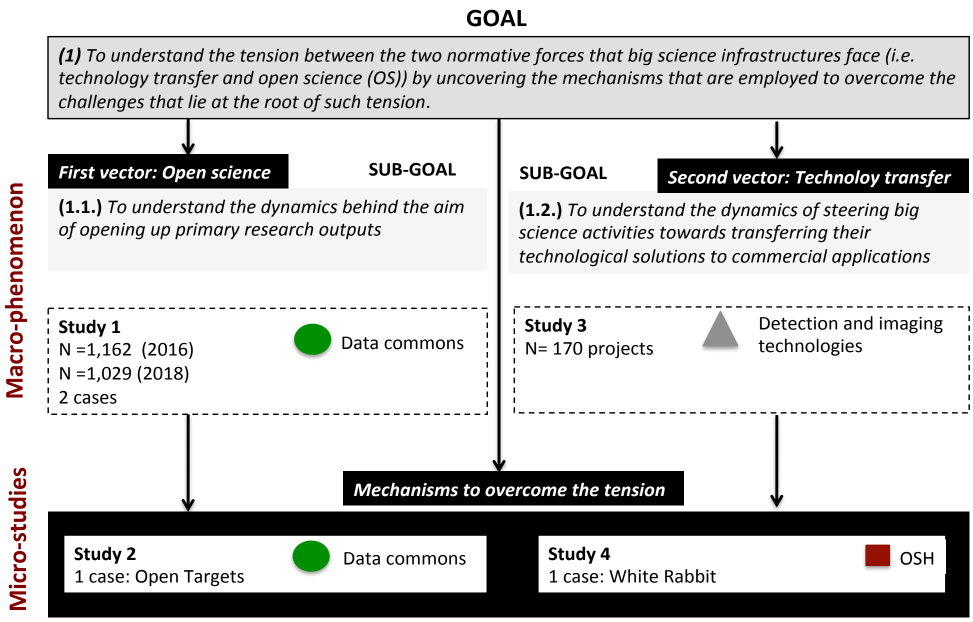
a) Data commons (also called "open data in research" or "data collaboratives"'). Data commons co-locate data, storage, and computing infrastructures with commonly used services and tools for analysing and sharing data to create an interoperable resource for the research community (Grossman et al., 2016).

b) Open source hardware (OSH) refers to tangible artifacts—machines, devices, or other physical things—whose design is made publicly available in such a way that anyone can study, modify, distribute, make, and sell the design or hardware based on that design (Open Source Hardware Association 2012).

## **2.6** Overview of four complementary studies

As briefly introduced in section 1.2, to respond to the aforementioned overarching research aim the dissertation is structured in four complementary empirical studies (Figure 2).

*Figure 1. Overview of the four complementary studies*


GOAL
(1) To understand the tension between the two normative forces that big science infrastructures face (i.e. technology transfer and open science (OS)) by uncovering the mechanisms that are employed to overcome the challenges that lie at the root of such tension.
First vector: Open science
SUB-GOAL
SUB-GOAL
Second vector: Technoloy transfer
(1.1.) To understand the dynamics behind the aim of opening up primary research outputs
(1.2.) To understand the dynamics of steering big science activities towards transferring their technological solutions to commercial applications
Macro-phenomenon
Study 1
N =1,162 (2016)
N =1,029 (2018)
2 cases
Data commons
Study 3
N= 170 projects
Detection and imaging technologies
Mechanisms to overcome the tension
Micro-studies
Study 2
1 case: Open Targets
Data commons
Study 4
1 case: White Rabbit
OSH

While the first and third studies offer us a contextual overview of the dynamics in the two vectoral forces of open science (study 1) and technology transfer (study 3), the studies on Open Targets (OT) (study 2) and White Rabbit (WR) (study 4) provide us with the required depth to understand the specific mechanisms that organisations use to reconcile the tensions caused by these two trajectories. Both studies, WR and OT, offer a complementary perspective on describing the tensions of these two exogenous forces by examining two different OS dimensions: OSH in WR, and data commons in OT. The two case studies come from two leading big science infrastructures in two different fields: high-energy physics (CERN) and molecular biology (OT), and offer us some variance when investigating the mechanisms that emerge to overcome the tensions. That is, in the case of WR it is hardware with embedded operating, middleware, or application-level software applied as a synchronisation device in multiple industrial settings (OSH), while for OT it is data and technological tools to accelerate R&D development processes in drug discovery (data commons). Both studies share the unit of analysis by looking at the ecosystem of organisations that contributed to the development of WR, namely firms' network and research organisations and their interactions when developing WR; and the organisations that develop OT information infrastructure, that is, big science infrastructure, pharmaceutical and biotech companies and other research organisations.

While these two investigations are paramount in our inquiry, the two larger studies addressing sub-goals 1.1. and 1.2 provide us with the contextual overview to understand the dynamics in each normative vector, that is, the difficulties behind opening up scientific data by exploring the factors behind the gradual and disparate adoption of data sharing practices across scientists (i.e. first study: open science vector); and the nature of the serendipitous process behind transferring big science technologies to alternative and previously unanticipated commercial applications (i.e. third study: technology transfer vector).

The units of analysis of the four studies are also complementary: the first study investigates the researcher perspective (individual level of analysis); the second and fourth study investigate the ecosystem of organisations participating in Open Targets and White Rabbit, and the third study is at a project level of analysis investigating the serendipitous mode in ATTRACT projects. This complementarity provides us with a holistic examination of the friction generated by the two exogenous forces.

In terms of research design (see Table 2), the four studies display heterogeneity in methods. The first study employs a mixed-method approach where the survey data of two global online surveys in 2016- 2018 are compared and combined with qualitative data of two case studies. The second and fourth studies consist of single-case qualitative studies based on primary and secondary data. The third study codifies 170 project proposals (3,000 words on average) to identify patterns across proposals and cluster them by identifying four main serendipity modes. The variety in methods responds to the needs of each specific study question, while it helps to provide a global view of the phenomenon under study, that is, the dynamics within and across open science and technology transfer in the context of big science.

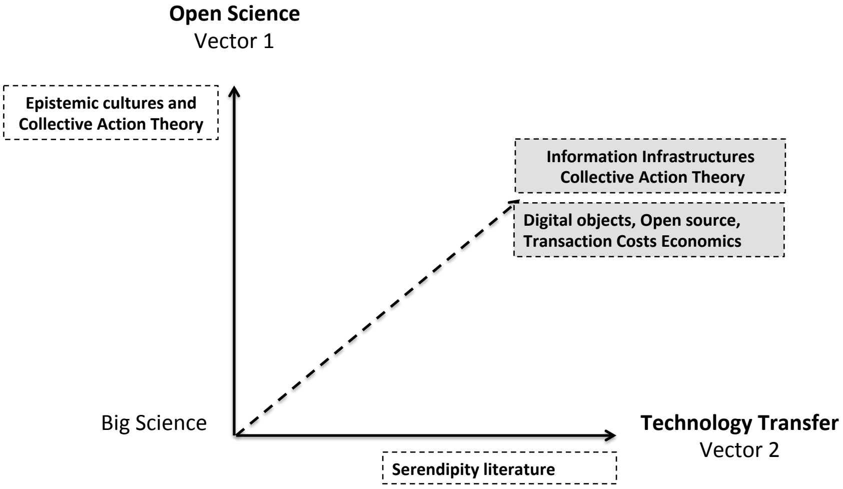
*Table 1: Overview of the four empirical studies*

| **Study #** | **1: The stickiness of scientific data sharing** | **2: Opaque spaces of the commons: governing information infrastructures in life sciences** | **3: Systematizing serendipity for big science infrastructures: the ATTRACT project** | **4: From bits to atoms: White Rabbit at CERN** |
|---|---|---|---|---|
| **Big science field** | High-energy physics<br>Molecular biology | Molecular biology | High-energy physics<br>Molecular biology<br>Astronomy | High-energy physics |
| **ICT** | Data infrastructures | Data infrastructures | Detection and imaging technologies | Hardware, gateware and software |
| **Open science dimension** | Data commons | Data commons | - | Open-source hardware |
| **Technology transfer** | | From big research on targets to acceleration of development phases in the drug discovery process | From big science research projects towards proof-of-concept and commercialisation phases | From the big science research setting to the commercialisation of the resulting open-source hardware in different industrial settings |
| **Data** | N =1,162 (2016)<br>N =1,029 (2018)<br>2 cases | 1 case | 170 projects | 1 case |
| **Unit of analysis** | Individual-level (the scientist) | Ecosystem of organisations | Project level | Ecosystem of organisations |

## 2.7 The theoretical basis of the four empirical studies

The disparity of the technological objects at the core of each work (open-source hardware, data commons), the different perspectives employed (from micro-foundations to the ecosystem level) and the heterogeneity in the specific interrogations of the four studies led to the adoption of different theoretical foundations to more appropriately and effectively guide each study towards progress in the common overarching research objective of the dissertation (see Figure 2 below).

*Figure 2. Overview of theoretical foundations employed across studies*


**Open Science**
Vector 1
Epistemic cultures and Collective Action Theory
Information Infrastructures Collective Action Theory
Digital objects, Open source, Transaction Costs Economics
Big Science
**Technology Transfer**
Vector 2
Serendipity literature

As a result, the first study of the dissertation combines a cultural perspective from the sociology of science that draws on the notion of "epistemic cultures" (Knorr Cetina 1999) and a collective action theory perspective (Hardin 1968, 1982; Olson 2009; Ostrom 1990) that seeks to understand the complex, intricate system of incentives and disincentives that scientists confront when considering whether to contribute to the collective goods of data sharing.

The second study relies on the literature on information infrastructures (Constantinides 2012; Constantinides and Barrett 2015; Hanseth and Monteiro 1997) and adopts a collective action theory approach (Hardin 1968, 1982; Olson 2009; Ostrom 1990) to identify the many trade-offs between private and collective interests in the development of shared resource pools composed of heterogeneous members with different objectives.

The third study is based on prior literature on serendipity (Fink et al. 2017; Garud et al. 2018; Yaqub 2018) to test serendipity models previously identified and identify two additional ones.

Finally, the fourth study builds upon the literature on digital objects and artifacts (Faulkner and Runde 2019; Kallinikos et al. 2013; Orlikowski and Iacono 2001; Yoo 2010) and open-source (Benkler 2002; Dahlander and Magnusson 2008; Feller and Fitzgerald 2002; Fitzgerald 2006; Fitzgerald and Feller 2002; Howison and Crowston 2014; O'Mahony and Ferraro 2007) while adopting transaction costs economics theoretical lenses (Williamson 1975, 1985, 1996). These theoretical underpinnings help us to effectively isolate how OSH differs from what we know about open-source software and better inform the challenges around developing an object with physical components.

*Table 2. Overview of the research design of the four empirical studies*

| **Study #** | **1: The stickiness of scientific data** | **2: Opaque spaces of the commons: Governing information infrastructures in Life Sciences** | **3: Systematising serendipity for big science infrastructures** | **4: From bits to atoms: White Rabbit at CERN** |
|---|---|---|---|---|
| **Research Question** | Do researchers share their data? How do they share their data? Which mechanisms emerge to enable researchers to share their data? | How do organisations develop commons-based information infrastructures that govern access to collective resources while simultaneously protecting members' private interests? | What are the formative conditions of serendipity transforming big science research towards commercial applications? | How do the attributes of a hybrid object and its components affect the open-source model of development? |
| **Theoretical Foundation** | Epistemic cultures<br>Collective action theory | Information Infrastructures<br>Collective Action theory | Serendipity | Digital objects and IT artifacts<br>Open source<br>Transaction Costs Economics |
| **Research Design** | Comparative | Longitudinal | Cross-sectional | Longitudinal |
| **Data** | Survey data 2016 (n=1,162)<br>Survey data 2018 (n=1,029)<br>2 cases: High energy physics (Reana) and molecular biology (Open Targets) | 1 case: Open Targets | 170 project proposals (3,000 words each) | 1 case: White Rabbit |
| **Analyses** | Mixed-method study: Comparing data practices across time (2016- 2018) and scientific communities (High energy physics and molecular biology) | Qualitative study: Analysis of primary and secondary data to identify governance mechanisms and conditions that align organisations' individual and collective interests | Qualitative study: Clustering and coding serendipity modes | Qualitative study: Analysis of primary and secondary data to relate object attributes with development models |

## 2.8 Research process and scholarly contributions

The four articles of this dissertation are in different stages of publication. The first article, co-authored with Jonathan Wareham, is based on the research done in the framework of the Open Science Monitor commissioned by the Directorate General of Research and Innovation of the European Commission and published by the European Commission:

> Pujol Priego L, Wareham J. "REANA: Reproducible Research Data Analysis Platform: Open Science Monitor Case Study". *European Commission Directorate-General for Research and Innovation*, B-1049 Brussels. 2019. http://publications.europa.eu/publication/manifestation_identifier/PUB_KI0219176ENN. Accessed March 20, 2019.
>
> Pujol Priego L, Wareham J. "Zenodo". *European Commission Directorate-General for Research and Innovation*, B-1049 Brussels. 2019. https://op.europa.eu/en/publication-detail/-/publication/b5187345-f3b1-11e9-8c1f-01aa75ed71a1/language-en/format-PDF/source-118580915. Accessed March 20, 2019
>
> Pujol Priego L, Wareham J. "Open Targets: Open Science Monitor Case Study". *European Commission Directorate-General for Research and Innovation*, B-1049 Brussels. 2018 http://publications.europa.eu/publication/manifestation_identifier/PUB_KI0518020ENN. Accessed March 20, 2019.
>
> Pujol Priego L, Wareham J. "Pistoia Alliance: Open Science Monitor Case Study". *European Commission Directorate-General for Research and Innovation*, B-1049 Brussels; 2018. http://publications.europa.eu/publication/manifestation_identifier/PUB_KI0618230ENN. Accessed March 20, 2019.
>
> Pujol Priego L and Wareham J. "Yoda: Open Science Monitor Case Study." *European Commission Directorate-General for Research and Innovation*, B-1049 Brussels. 2018. http://publications.europa.eu/publication/manifestation_identifier/PUB_KI0518019ENN. Accessed March 20, 2019

The manuscript is currently in preparation for submission to a leading IS journal.

The second article, also co-authored with Jonathan Wareham, was presented in:

> Pujol Priego, L. and Wareham, J. D. (2019) "Open Targets: Pre-competitive Collaborative Research in Life Sciences." *Academy of Management Proceedings*. Vol. 2019. No. 1. Briarcliff Manor, NY 10510: Academy of Management, 2019.

The manuscript is currently in preparation for submission to a leading IS journal.

The third article written with Jonathan Wareham, A. Romasanta, T. Wareham Mathiassen, M. Nordberg and P. Garcia Tello was revised and resubmitted (under the second round of review) to *Technovation* in March 2020.

The fourth article is co-authored with Jonathan Wareham and was presented in:

> Pujol Priego, L. and Wareham, J. D. (2018) "Time as a service: White Rabbit at CERN" *International Conference on Information Systems Proceedings*. Vol. 2018. San Francisco: Association for Information Systems.

It has been submitted in February 2020 to *MIS Quarterly* (under review)

Finally, this research has also informed the following studies:

> Brunswicker, S., Pujol Priego, L. and Almirall, E. (2019). Transparency in policymaking: A complexity view. *Government Information Quarterly*. Volume 36, Issue 3, July 2019, pp. 571-591

> Pujol Priego L. and J. Wareham "Emergent open strategies to accelerate innovation: Lessons from the Pharmaceutical industry" *Harvard Deusto Business Review*, No 289, 05/2019, p. 70-81

> Susanne Beck, Carsten Bergenholtzj, Marcel Bogers, Tiare-Maria Brasseura, Marie Louise Conradsen, Diletta Di Marcou, Daniel Dörler, Agnes Efferta, Benedikt Fecher, Despoina Filiou, Thomas Gillierh, Christoph Grimpeb, Marc Gruberk, Carolin Haeusslerl, Florian Heigl, Karin Hoislp, Katie Hyslopa, Olga Kokshaginat, Marcel LaFlammea, Cornelia Lawson, Wolfgang Lukas, Markus Nordberg, Maria Theresa Nornj, Marion Poetz, Gernot Pruschak, Laia Pujol Priego, Agnieszka Radziwon, Janet Rafners, Alexander, Rusero, Henry Sauermann, Julia Suess-Reyesa, Sonali K. Shahk, Jacob F. Shersons, Christopher L. Tucci, Philipp Tuertscher, Jane Bjørn Vedel, Roberto Verganti, Jonathan Wareham, Sunny Mosangzi Xu. 2019 "Open Innovation in Science" *Industry and Innovation* (lead article in the special issue on Open Innovation in Science).

*Book Chapter*: Osimo, D., Pujol Priego, L., and Vuorikari, R. (2017). Alternative Research Funding Mechanisms: Make Funding Fit for Science 2.0. Research 2.0 and the Impact of Digital Technologies on Scholarly Inquiry (pp. 53-67). IGI Global.

**References**

Ågerfalk, P. J., and Fitzgerald, B., 2008. “Outsourcing to an Unknown Workforce: Exploring Opensourcing as a Global Sourcing Strategy,” *MIS Quarterly* (32:2), pp. 385–409. (https://doi.org/10.2307/25148845).

Ågerfalk, P. J., Fitzgerald, B., and Stol, K.-J. 2015. *Software Sourcing in the Age of Open*, Springer Briefs in Computer Science, Cham: Springer International Publishing. (https://doi.org/10.1007/978-3-319-17266-8).

Atkins, D., 2003. *Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure*. (https://repository.arizona.edu/handle/10150/106224).

Autio, E. 2014. *Innovation from Big Science: Enhancing Big Science Impact Agenda*, Department of Business, Innovation & Skills. Imperial College Business School, p. 76.

Autio, E., Bianchi-Streit, M., and Hameri, A.-P. 2003. *Technology transfer and technological learning through CERN’s procurement activity*, p. 92.

Autio, E., Hameri, A.-P., and Bianchi-Streit, M. 2003. *Technology Transfer and Technological Learning through CERN’s Procurement Activity*. CERN Scientific Information Service, Geneva.

Autio, E., Hameri, A.-P., and Vuola, O. 2004. “A Framework of Industrial Knowledge Spillovers in Big-Science Centers,” *Research Policy* (33:1), pp. 107–126. (https://doi.org/10.1016/S0048-7333(03)00105-7).

Benkler, Y. 2002. “Coase’s Penguin, or, Linux and ‘The Nature of the Firm,’” *The Yale Law Journal* (112:3), p. 369. (https://doi.org/10.2307/1562247).

Borgman, C. L. 2010. *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*, MIT Press.

Borgman, C. L. 2012. “The Conundrum of Sharing Research Data,” *Journal of the American Society for Information Science and Technology* (63:6), pp. 1059–1078. (https://doi.org/10.1002/asi.22634).

Borgman, C. L. 2015. *Big Data, Little Data, No Data: Scholarship in the Networked World*, Cambridge, United States: MIT Press. (http://ebookcentral.proquest.com/lib/georgetown/detail.action?docID=3339930).

Bozeman, B. 2000. "Technology Transfer and Public Policy: A Review of Research and Theory," *Research Policy* (29:4–5), pp. 627–655. (https://doi.org/10.1016/S0048-7333(99)00093-1).

Bressan, B. 2014a. *From Physics to Daily Life Applications in Informatics, Energy, and Environment*, (Wiley Online Library.), Weinheim: Wiley-Blackwell.

Bressan, B. 2014b. *From physics to daily life: Applications in Biology, Medicine, and Healthcare*, (2nd ed.), Weinheim an der Bergstrasse, Germany: Wiley Blackwell.

Castelnovo, P., Florio, M., Forte, S., Rossi, L., and Sirtori, E. 2018. "The Economic Impact of Technological Procurement for Large-Scale Research Infrastructures: Evidence from the Large Hadron Collider at CERN," *Research Policy*. (https://doi.org/10.1016/j.respol.2018.06.018).

Caulfield, T., Harmon, S. H., and Joly, Y. 2012. *Open Science versus Commercialization: A Modern Research Conflict?* p. 11.

Constantinides, P. 2012. *Perspectives and Implications for the Development of Information Infrastructures*, IGI Global.

Constantinides, P., and Barrett, M. 2015. "Information Infrastructure Development and Governance as Collective Action," *Information Systems Research* (26:1), pp. 40–56. (https://doi.org/10.1287/isre.2014.0542).

Cook, C. E., Bergman, M. T., Cochrane, G., Apweiler, R., and Birney, E. 2018. "The European Bioinformatics Institute in 2017: Data Coordination and Integration," *Nucleic Acids Research* (46:D1), Oxford Academic, pp. D21–D29. (https://doi.org/10.1093/nar/gkx1154).

Crowston, K., Ribes, D., Sawyer, S., and Wiggins, A. 2009. *Little EScience, Big EScience*. (https://www.ideals.illinois.edu/handle/2142/15375).

Dahlander, L., and Magnusson, M. 2008. "How Do Firms Make Use of Open Source Communities?," *Long Range Planning* (41:6), pp. 629–649. (https://doi.org/10.1016/j.lrp.2008.09.003).

David, P. 2003. *The Economic Logic of "Open Science" and the Balance between Private Property Rights and the Public Domain in Scientific Data and Information: A Primer"*, (The National Academies Press., Vol. The Role of Scientific and Technical Data and Information in the Public Domain: Proceedings of a Symposium), Washington, DC: National Research

Council. (https://doi.org/10.17226/10785).

David, P. A. 2004. "Understanding the Emergence of 'Open Science' Institutions: Functionalist Economics in Historical Context," *Industrial and Corporate Change* (13:4), Oxford Academic, pp. 571–589. (https://doi.org/10.1093/icc/dth023).

David, P. A., and Spence, M. J. 2003. "Towards Institutional Infrastructures for E-Science: The Scope of the Challenge," SSRN Scholarly Paper No. ID 1325240, SSRN Scholarly Paper, Rochester, NY: Social Science Research Network, September 1. (https://doi.org/10.2139/ssrn.1325240).

Dosi, G., Llerena, P., and Labini, M. S. 2006. "The Relationships between Science, Technologies and Their Industrial Exploitation: An Illustration through the Myths and Realities of the so-Called 'European Paradox,'" *Research Policy* (35:10), pp. 1450–1464. (https://doi.org/10.1016/j.respol.2006.09.012).

Dougherty, D., and Dunne, D. D. 2011. "Digital Science and Knowledge Boundaries in Complex Innovation," *Organization Science* (23:5), pp. 1467–1484. (https://doi.org/10.1287/orsc.1110.0700).

Edwards, P. N. 2019. "Knowledge Infrastructures under Siege : Climate Data as Memory, Truce, and Target," *Data Politics*, Routledge, March 13, pp. 21–42. (https://doi.org/10.4324/9781315167305-2).

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

European Commission. 2014. "Data Management - H2020 Online Manual." (https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm, accessed March 10, 2020).

European Commission. 2019. "Facts and Figures of Open Research Data," *European Commission - European Commission*. (https://ec.europa.eu/info/research-and-innovation/strategy/goals-research-and-innovation-policy/open-science/open-science-monitor/facts-and-figures-open-research-data_en, accessed April 19, 2019).

Faulkner, P., and Runde, J. 2019. "Theorizing the Digital Object," *MIS Quarterly 43(4)*, pp. 1279-1302

Feller, J., and Fitzgerald, B. 2002. *Understanding Open Source Software Development*, Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Fink, T. M. A., Reeves, M., Palma, R., and Farr, R. S. 2017. "Serendipity and Strategy in

Rapid Innovation," *Nature Communications* (8:1), Nature Publishing Group, pp. 1–9. (https://doi.org/10.1038/s41467-017-02042-w).

Fitzgerald. 2006. "The Transformation of Open Source Software," *MIS Quarterly* (30:3), p. 587. (https://doi.org/10.2307/25148740).

Fitzgerald, B., and Feller, J. 2002. "A Further Investigation of Open Source Software: Community, Co-ordination, Code Quality and Security Issues," *Information Systems Journal* (12:1), pp. 3–5. (https://doi.org/10.1046/j.1365-2575.2002.00125.x).

Florio, M., and Sirtori, E. 2016. "Social Benefits and Costs of Large Scale Research Infrastructures," *Technological Forecasting and Social Change* (112:Nov), pp. 65–78. (https://doi.org/10.1016/j.techfore.2015.11.024).

Friesike, S., Widenmayer, B., Gassmann, O., and Schildhauer, T. 2015. "Opening Science: Towards an Agenda of Open Science in Academia and Industry," *The Journal of Technology Transfer* (40:4), pp. 581–601. (https://doi.org/10.1007/s10961-014-9375-6).

Gaillard, M. 2017. "CERN Data Centre Passes the 200-Petabyte Milestone," *CERN*. (https://home.cern/news/news/computing/cern-data-centre-passes-200-petabyte-milestone).

Garud, R., Gehman, J., and Giuliani, A. P. 2018. "Serendipity Arrangements for Exapting Science-Based Innovations," *Academy of Management Perspectives* (32:1), Academy of Management, pp. 125–140. (https://doi.org/10.5465/amp.2016.0138).

Grossman, R. L., Heath, A., Murphy, M., Patterson, M., and Wells, W. 2016. "A Case for Data Commons: Toward Data Science as a Service," *Computing in Science & Engineering* (18:5), pp. 10–20. (https://doi.org/10.1109/MCSE.2016.92).

Guston, D. H. 2000. "Retiring the Social Contract for Science," *Issues in Science and Technology* (16:4), the University of Texas at Dallas, pp. 32–36.

Hallonsten, O. 2014. "How Expensive Is Big Science? Consequences of Using Simple Publication Counts in Performance Assessment of Large Scientific Facilities," *Scientometrics*. (https://doi.org/10.1007/s11192-014-1249-z).

Hanseth, O., and Monteiro, E. 1997. "Inscribing Behaviour in Information Infrastructure Standards," *Accounting, Management and Information Technologies* (7:4), pp. 183–211. (https://doi.org/10.1016/S0959-8022(97)00008-8).

Hardin, G. 1968. "The Tragedy of the Commons," *Science* (162:3859), pp. 1243–1248. (https://doi.org/10.1126/science.162.3859.1243).

Hardin, R. 1982. *Collective Action*, Baltimore: Published for Resources for the Future by the Johns Hopkins University Press.

Heidler, R., and Hallonsten, O. 2015. “Qualifying the Performance Evaluation of Big Science beyond Productivity, Impact and Costs,” *Scientometrics*. (https://doi.org/10.1007/s11192-015-1577-7).

Hellström, T., and Jacob, M. 2012. “Revisiting ‘Weinberg’s Choice’: Classic Tensions in the Concept of Scientific Merit,” *Minerva*. (https://doi.org/10.1007/s11024-012-9203-9).

Hey, T. 2009. *The Fourth Paradigm: Data-intensive Scientific Discovery*, Redmond, Washington: Microsoft Pr.

Howison, J., and Crowston, K. 2014. “Collaboration Through Open Superposition: A Theory of the Open Source Way,” *MIS Quarterly* (38:1), pp. 29-A9.

Kallinikos, J., Aaltonen, A., and Marton, A. 2013. “The Ambivalent Ontology of Digital Artifacts,” *MIS Quarterly* (37:2), pp. 357–370. (https://doi.org/10.25300/MISQ/2013/37.2.02).

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

NIH. 2003. “NIH Data Sharing Policy and Implementation Guidance.” (https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm, accessed March 10, 2020).

OECD. 2015. “Making Open Science a Reality,” No. 25, OECD Science, Technology and Industry Policy Papers, Paris: OECD Publishing. (https://wiki.lib.sun.ac.za/images/0/02/Open-science-oecd.pdf).

Olson, M. 2009. *The Logic of Collective Action*, (Vol. 124), Harvard University Press.

O’Mahony, S., and Ferraro, F. 2007. “The Emergence of Governance in an Open Source Community,” *Academy of Management Journal* (50:5), pp. 1079–1106. (https://doi.org/10.5465/amj.2007.27169153).

Open Source Hardware Association. 2012. “Definition of Open Source Hardware,” *Open Source Hardware Association*, May 26. (https://www.oshwa.org/definition/, accessed May 16, 2019).

Orlikowski, W. J., and Iacono, C. S. 2001. “Research Commentary: Desperately Seeking ‘IT’ in IT Research - A Call to Theorizing the IT Artifact,” *Information Systems Research; Linthicum* (12:2), pp. 121–134.

Ostrom, E. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*, Cambridge University Press.

Perkmann, M., and Schildt, H. 2015. “Open Data Partnerships between Firms and

Universities: The Role of Boundary Organizations," *Research Policy* (44:5), pp. 1133–1143. (https://doi.org/10.1016/j.respol.2014.12.006).

Perkmann, M., Tartari, V., McKelvey, M., Autio, E., Broström, A., D'Este, P., Fini, R., Geuna, A., Grimaldi, R., Hughes, A., Krabel, S., Kitson, M., Llerena, P., Lissoni, F., Salter, A., and Sobrero, M. 2013. "Academic Engagement and Commercialisation: A Review of the Literature on University-Industry Relations," *Research Policy* (42:2), pp. 423–442. (https://doi.org/10.1016/j.respol.2012.09.007).

Schmied, H. 1982. "Results of Attempts to Quantify the Secondary Economic Effects Generated by Big Research Centers," *IEEE Transactions on Engineering Management* (EM-29:4), pp. 154–165. (https://doi.org/10.1109/TEM.1982.6448551).

Stockinger, H. 2005. "Message from the Program Chair," in *First International Conference on E-Science and Grid Computing (e-Science'05)*, Melbourne Vic., Australia, July, xiii–xiii. (https://doi.org/10.1109/E-SCIENCE.2005.57).

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., Pollock, D., and Dorsett, K. 2015. "Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide," *PLOS ONE* (10:8), (P. van den Besselaar, ed.), p. e0134826. (https://doi.org/10.1371/journal.pone.0134826).

Wallis, J. C., Rolando, E., and 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), (L. A. Nunes Amaral, ed.), p. e67332. (https://doi.org/10.1371/journal.pone.0067332).

Weinberg, A. M. 1961. "Impact of Large-Scale Science on the United States," *Science*. (https://doi.org/10.1126/science.134.3473.161).

Weinberg, A. M. 1963. "Criteria for Scientific Choice," *Minerva*. (https://doi.org/10.1007/BF01096248).

Weinberg, A. M. 1964. "Criteria for Scientific Choice II: The Two Cultures," *Minerva*. (https://doi.org/10.1007/BF01630147).

West, J. 2008. "Commercializing Open Science: Deep Space Communications as the Lead Market for Shannon Theory, 1960-73," *Journal of Management Studies* (45:8), pp. 1506–1532. (https://doi.org/10.1111/j.1467-6486.2008.00807.x).

Williamson, O. E. 1975. *Markets and Hierarchies, Analysis and Antitrust Implications: A Study in the Economics of Internal Organization*, New York: Free Press.

Williamson, O. E. 1985. *The Economic Institutions of Capitalism: Firms, Markets, Relational Contracting*, (1st Free Press pbk. ed.), London: Collier Macmillan Publishers.

Williamson, O. E. 1996. *The Mechanisms of Governance*, Oxford University Press.

World Intellectual Property Organization. 2010. "Managing IP at CERN," *WIPO Magazine* (6/2010). (https://www.wipo.int/wipo_magazine/en/2010/06/article_0003.html).

Yaqub, O. 2018. "Serendipity: Towards a Taxonomy and a Theory," *Research Policy* (47:1), pp. 169–179. (https://doi.org/10.1016/j.respol.2017.10.007).

Yoo, Y. 2010. "Computing in Everyday Life: A Call for Research on Experiential Computing," *MIS Quarterly* (34:2), pp. 213–231. (https://doi.org/10.2307/20721425).

# 3

## 3. The Stickiness of Scientific Data: Epistemic Cultures and a Collective Action Dialogue

The article that constitutes this chapter aims at understanding the first vector: open science, which responds to the first sub-goal (1.1.) of our PhD investigation. The study empirically investigates the dynamics behind sharing scientific data while interrogating the explanatory factors behind the gradual and disparate adoption of data sharing practices across scientists

## 3.1 Abstract

Researchers are generating unprecedented volumes of data. As the expectations of big scientific data grow, the expectations on the potential of sharing it and allowing others to mine, aggregate, and recombine it with other data for novel findings grow as well. As such, government funding entities, particularly in Western Europe and the US, have placed open data at the crux of scientific policy. While sharing scientific data has been positively promoted for some time now, only recently have several challenges become apparent, suggesting that data sharing imposes increased costs on scientists and their institutions without commensurate professional benefits. Considering the tensions between policymakers and funding agencies' efforts to foster data sharing and the apparent barriers to its wide adoption, we lack 1) a recent overview of data being shared across scientists (if and what); 2) how researchers share their data (how), and 3) what mechanisms enable research data sharing (why). Our study engages in a mixed-method design by combining survey data collected in 2016 (n=1,162) and 2018 (n=1,029) to explore data sharing behaviours of scientists across disciplines and countries; and qualitative data from two case studies sequentially sampled within two scientific communities of the disciplines surveyed (i.e. physics and life science): high-energy physics (HEP) and molecular biology (MB). As a lens to understand the factors behind data sharing practices, we draw upon the notion of epistemic cultures, originated from the sociology of science, and the collective action theory perspective to shed light on the incentives and deterrents that scientists confront when considering contributions to the collective goods of data sharing.

**Keywords:** open science, scientific data sharing, data commons, epistemic cultures, collective action theory.

## 3.2 Introduction

In September 2011 OPERA (Oscillation Project with Emulsion-tRacking Apparatus) researchers fired a 730 km beam of muon neutrinos from CERN (European Organization for Nuclear Research) to the Gran Sasso National Laboratory in central Italy at what appeared to be faster than the speed of light. Puzzled by these results, they decided to upload all the data with unprecedented granularity at arXiv.org. The scientific team included all the necessary procedural descriptions so that other scientists could search for an explanation for this surprising violation of physical law. More than 200 papers emerged and were shared at arXiv.org trying to explain the effect. With ruthless external scrutiny, the mystery was resolved within a year: the OPERA team announced the identification of two potential sources of timing error that corrupted measurements (Royal Society 2012). Many similar examples abound on the scientific and social value of data sharing; but if researchers were asked today whether they release their data, what would they answer?

Researchers are generating unprecedented volumes of data (Hey 2009). Although some disciplines have a long tradition of working with big data, particularly the big science research infrastructures (Weinberg 1961) for physics and astronomy (Atkins et al. 2003; Borgman 2015, 2015; Carillo and Papagni 2014), other disciplines have only recently begun to adopt the practice (EIROforum IT working group 2013). Examples of recent adopters of big data include computational social science (Lazer 2009), digital humanities (Kaplan 2015), sensor devices (Wallis et al. 2013), social media data (Plantin et al. 2018) citizen science research projects (Hochachka et al. 2012), and political science and public policy (Lee et al. 2016).

Perspectives have evolved, increasing the scale, role, and status of data in recent years. Scientific data is now its own scholarly object with dedicated journals such as *Nature-Scientific Data*. The increasing use of data-intensive methods has been labelled the "fourth paradigm" in science (Atkins et al. 2003; Hey 2009) that augments "the existing paradigms of experimental theoretical and computational science" (Edwards et al. 2011 p. 670). As the expectations of big scientific data grow, the expectations on the potential of *sharing* it and allowing others to mine, aggregate, and recombine it with other data for novel findings grow as well: *"*If the rewards of the data deluge are to be reaped, then researchers who produce those data must share them, and do so in such a way that the data are interpretable and reusable by others*"* (Borgman 2012 p. 1059). Data sharing describes the act of releasing data in a form that can be used by others (Pasquetto et al. 2017). If research data needs to be shared, it is expected to be FAIR (Findable, Accessible, Interoperable and Reusable)[5] so that it can be easily and effectively discovered and reused. Recent studies have estimated that the annual financial cost of not sharing FAIR data to be at least €10.2bn for the European economy; an additional estimate of the impact of FAIR on potential economic growth is €16bn annually (European Commission 2019b).

The growing importance of *sharing* (FAIR) data comes as part of a more general "open" movement embracing greater transparency in science (Edwards 2019). Starting with open access publishing, it expanded towards open scientific data, open standards, open repositories, open bibliography, open lab-notebooks, open-source software and hardware, with an endless list of 'open'- qualifiers to all activities in the scientific realm (Friesike et al. 2015). The urgency of sharing FAIR data is not only grounded in the reproducibility crisis

[5]. The term FAIR was launched in the Lorentz workshop celebrated in 2014. The resulting FAIR principles were published in 2016. See https://www.go-fair.org/fair-principles/

(Baker 2015) or concerns about fraudulent scientific practices (Kupferschmidt 2018) but also a recognition of the novel technological and scientific innovations resulting from data sharing (Borgman 2010).

As such, government funding entities, particularly in Western Europe and the US, have placed open data at the crux of scientific policy. Carlos Moedas, the former EU Commissioner for Research, Science and Innovation, made open research data one of the EU's priorities in 2015. Several expert working groups were put in place (e.g. High-level expert group on FAIR data; the Open Science Policy Platform; Expert group on altimetrics) to provide advice about how to foster and promote research data sharing in Europe. In 2016, the European Commission launched the Open Science Cloud initiative, a federated data infrastructure with cloud-based services to offer the scientific community an open environment for storing, sharing, and reusing scientific data. This policy evolution has been accompanied by requirements from funding agencies that scientific data be publicly available: US National Institutes of Health (NIH) in 2003 for grants over $500,000 (NIH 2003), the National Science Foundation (NSF) in 2010 (Borgman 2012), and the European Commission for Horizon 2020 program in 2014 (European Commission 2014).

Accompanying policy, new private and public entities have emerged to facilitate the aggregation and publication of research data. Examples include the Research Data Alliance, the National Data Service, as well as for-profit publishers who attempt to build on existing structures (e.g. Mendeley Data) (Borgman 2015). Platforms such as Dataverse (King 2007), FigShare (Thelwall and Kousha 2016), Dryad (White et al. 2008), Zenodo (Peters et al. 2017), DataHub (Bhardwaj et al. 2014), and EUDat (Lecarpentier et al. 2013) have also emerged, offering scholars new venues to archive and share their data (Cragin et al. 2010).

While sharing scientific data has been positively promoted for some time now, only recently have several challenges become apparent. In general terms, researchers have identified factors inhibiting data sharing, suggesting that it imposes increased costs on scientists and their institutions without commensurate professional benefits (Borgman 2015; Edwards 2019; Edwards et al. 2011; Tenopir et al. 2015; Wallis et al. 2013). Considering the tensions between policymakers and funding agencies' efforts to foster data sharing and the apparent barriers to its wide adoption, we lack 1) a recent overview of data being shared across scientists (*if* and *what*); 2) how researchers share their data (*how*), and 3) what mechanisms enable research data sharing (*why*).

Hence, the research questions that this study seeks to answer are:

*RQ1a: Do researchers share their data?*

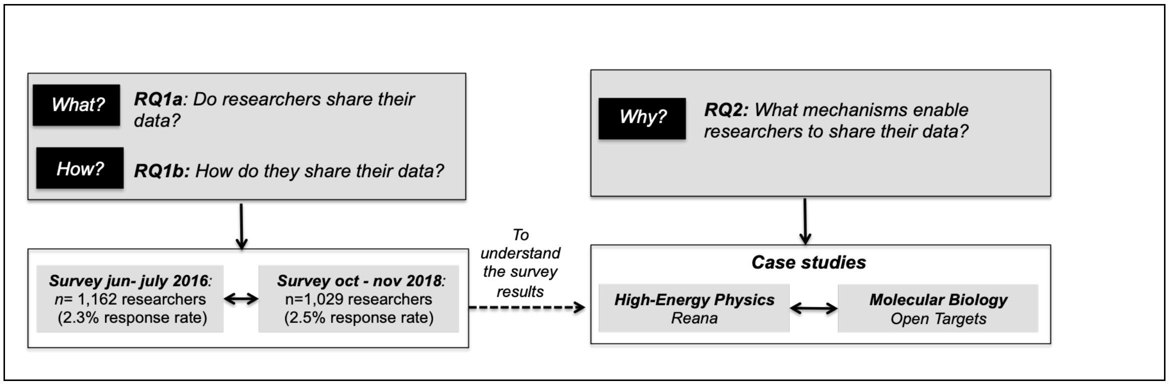
*RQ1b: How do they share their data?*

*RQ2: What mechanisms enable researchers to share their data?*

Our study engages in a mixed-method design to answer the research questions (Venkatesh et al. 2013) (Figure 1). First, to answer RQ1a and RQ1b, we employ survey data collected in 2016 (n=1,162) and 2018 (n=1,029) to explore data sharing behaviours of scientists across disciplines and countries. To explain the results from the survey and answer our RQ2 (*why*), we employ qualitative data from two case studies sequentially sampled within two of the disciplines surveyed (i.e. physics and life science). We chose these disciplines because they displayed the highest rates of data sharing and reuse in our survey findings, yet have significantly different scientific cultures, offering some variance needed to investigate the factors and boundary conditions behind data sharing practices. Specifically, we selected the communities of high-energy physics (HEP) and molecular biology (MB). Two information infrastructures (i.e. *Reana* (HEP) and *Open Targets* (MB)) were established to facilitate scientific data sharing within these communities. Our study complements the survey findings through an analysis of the architecture, practices, and governance of each infrastructure.

As a lens to understand the factors behind the data sharing practices, we draw upon both cultural and rational perspectives. The notion of 'epistemic cultures' originated from the sociology of science but has been subsequently applied in IS and organisational studies to understand information and knowledge sharing across communities (e.g. Kellogg et al. 2006; Mørk et al. 2008).  This perspective helps us explain the diversity and discontinuity across scientific communities and their heterogeneous data sharing practices. We augment a cultural perspective with a rational perspective to understand the incentives and disincentives that scientists confront when considering contributions to the collective goods of data sharing. Towards this, we employ collective action theory (Hess and Ostrom 2003; Olson 2009; Ostrom 1990), also used by IS scholars to explain how agents share heterogeneous information when developing common information infrastructures (Constantanides 2012; Constantinides and Barrett 2015; Vassilakopoulou et al. 2016). Collective action theory provides a useful framework to explain why researchers would contribute their data to collective resources by identifying incentives that are articulated for scientists to share. We believe that these complementary perspectives are useful in elucidating scientists' data sharing practices.

*Figure 1. A mixed-method approach to answer RQ1a,b, and RQ2*


What?
RQ1a: Do researchers share their data?
How?
RQ1b: How do they share their data?
Why?
RQ2: What mechanisms enable researchers to share their data?
Survey jun- july 2016: n= 1,162 researchers (2.3% response rate)
Survey oct - nov 2018: n=1,029 researchers (2.5% response rate)
To understand the survey results
Case studies
High-Energy Physics Reana
Molecular Biology Open Targets

The remainder of the paper is organised as follows. We first provide the research context by reviewing the background concepts from the IS and STS pieces of literature to delineate *what data is* and identify the reasons for sharing, or not sharing, scientific data. In a second step, we review the theoretical foundations of our research study and sequentially describe our methods and results. We follow Venkatesh et al.'s (2013) guidelines on how to present results of mixed-method studies: we first present the method and results from the survey data, and thereafter, the method and results from the case studies. We synthesise the findings and discuss the theoretical and practical implications of the study, its limitations, and future directions.

## 3.3 Research Context

An examination of research data sharing practices requires a brief review of the ontology of data as portrayed in the academic literature and the role of data in scientific knowledge production.

### *3.3.1 Conceptual considerations: what (or when) data is*

"Data are representations of observations, objects, or other entities used as evidence of phenomena for research or scholarship" (Borgman 2015 p. 18). A more operational definition from OAIS (Open Archival Information System) defines data as "a reinterpretable representation of information in a formalised manner suitable for communication, interpretation, or processing." Examples of data include: "sequence of bits, a table of numbers, the characters on a page, the recording of sounds made by a person speaking, or a moon rock specimen" (Consultative Committee for Space Data Systems 2012 p. 10).

The first important observation when examining these definitions is that data is *created* by people or machines. Entities may have a material existence or maybe a digital one (e.g. signals from sensors), which requires "acknowledge[ing] relationships between data, computers, models, and software" (Uhlir and Cohen 2011, as reported in Borgman, 2012a, p.

1061). These entities become data *when* scientists use them as evidence to understand a phenomenon better. In a seminal work, Susan Leigh Star describes how we may unveil the *processes* by which a scientific fact "emerges which is simultaneously stripped of its complexities and isolated from its relationship to a larger work/historical context" (Star 1983, pp. 224–225; Kallinikos and Tempini 2014). Thus, data typically involves a *process* (Kallinikos and Constantinou 2015) in which a scientist considers the observation or any other entity as evidence for a phenomenon and "collects, acquires, represents, analyses, and interprets those entities as data" (Borgman 2015 p. 62). Recognizing that data is a process implies that the research context is determinative to what becomes data and how it is processed (Kallinikos and Constantinou 2015). As such, it becomes paramount that all relevant contextual information is gathered in the description of the data, giving *metadata* a critical role in data sharing practices; metadata increases the utility of data across disciplines, time, geography or application domains (Edwards et al. 2011). For data to be reusable for those who did not create it, metadata needs to describe how it was generated, measured, and recorded. For instance, in biobanks, data creators need to publish highly-detailed descriptions of data collection parameters and procedures, including what was excluded or considered irrelevant (Demir and Murtagh 2013).

The conditions, instruments or mechanisms by which data is generated and recorded also informs the different *types* of research data (Kallinikos and Constantinou 2015). The US National Science Board distinguishes between: a) *observational data*, which results from identifying and recording facts or occurrences of a phenomenon; b) *computational data,* which results from implementing computer models or simulations; and c) e*xperimental data*, which is the product of implementing procedures in controlled conditions to test hypotheses or discover new laws (National Science Board 2005). Finally, *a record* is a fourth category that encompasses "everything else" not present in the former three. It is essential to acknowledge that this categorisation is permeable to some extent: observational data can be used in computational models or results from experiments may be used to refine the collection of observations.

The genesis of data may also affect an operational decision about whether to preserve the data and for how long (National Science Board 2005). For instance, it is considered essential to preserve observational data because it is the most difficult to replicate. Computational data requires extensive documentation on hardware, software, input data, and the workflow followed. Finally, the replicability of experimental data highly depends on the conditions of the experiment (Raphael et al. 2020). Nevertheless, who makes such decisions? Who has the authority to decide whether to destroy, share, or withhold data?

### *3.3.2 Incentives and deterrents for data sharing in science*

Where authors are initially the copyright holders of their academic publications, the jurisdiction of data is more ambiguous: uncertainty around ownership, control, and access over the data generates tensions. "Even when individuals and groups assign authority for data, the rights and responsibilities may remain unclear" (Borgman 2015 p. 43). What happens in practice is that raw data usually becomes the "intellectual and physical property of their creator" (Bowker 1999 p. 646). Policymakers, funders, and academic institutions are working towards an increased awareness that while publications and the knowledge produced from the research data pertains to the authors, the underlying data needs to be considered a public good (European Commission 2014; OECD 2015) so that its potential value can be unleashed (Järvenpää and Markus 2018; Vassilakopoulou et al. 2016).

Merton (1973) captured the norms of science in the imperatives of disinterestedness, communalism, universalism and organised scepticism. These principles highlight the cooperative spirit of scientific inquiry and emphasise that knowledge growth stems from collaboration where transparency and making scientific processes and outputs public are fundamental. Nevertheless, the transparency applied for research outputs such as publications does not play the same way for data: while scientists disseminate their publications, this differs when it comes to data (Tenopir et al. 2015).

While the practical concerns of making raw data transparent have been removed by lowering the cost of storing digital data and computing progress, other *reasons* behind non-sharing come into play (Table 1). First, there is a lack of incentives and rewards in the scholarly system which is heavily biased towards traditional journal and conference dissemination (Borgman 2015; Plantin et al. 2018); promotion and tenure decisions rarely take into account "subsidiary" products such as data or software contributions (Harley et al. 2010; Howison et al. 2015). Relatedly, some journals (e.g. *The Journal of Neuroscience* (Maunsell 2010)) recently announced that they would no longer publish supplementary data as reviewers are not able to spend the time required to scrutinise the material. Basically, for scholars driven by credit, sharing data offers little benefit, particularly in light of intentions to try to publish future articles out of the same data or aggregating it with complementary datasets (Harley et al. 2010; Meijer et al. 2017). Other factors include misuse, misinterpretation or liability concerns (Meijer et al. 2017; Tenopir et al. 2015; Wallis et al. 2013), in particular the fear that their work practices will come under scrutiny (Harley et al. 2010). A lack of skills, expertise, and tools to make their data available also hinders the practice (Borgman 2015). Finally, the real difficulty and costs associated with getting researchers to record detailed *metadata* are determinative (Edwards et al. 2011). Scientists' main interest is in using the data and they have little incentive to incur the additional overhead of a collective of unknown and future researchers "to whom they are not accountable and from whom they receive little if any benefit" (Edwards et al. 2011 p. 673; Gitter 2010). The production of metadata and the contextual descriptions of datasets require a critical amount of time to repair mistakes and misunderstandings, and researchers prefer to spend more time on new endeavours. Some have attempted to calculate the costs of metadata production, which could span an estimated two to three weeks from an average of a two-year research grant application (OpenAire 2019). In a dedicated study to examine high-energy physics practices, the vast majority of respondents (94.3%) thought that "the additional effort needed for preparing data for preservation in a re-usable form is substantial (more than 1% of the overall effort invested in producing and analysing the data) whereas 43.0% think that the supplementary effort is more than 10%" (Holzner et al. 2009 p.6). Table 1 summarises these arguments.

*Table 1. Reasons why data sharing is disincentivised in science*

| Reasons for not sharing | Description | Source |
|---|---|---|
| Lack of credit | There is a lack of consistency in the way data is cited. | (Borgman 2015; Meijer et al. 2017; Parsons et al. 2010; Piwowar and Vision 2013) |
| Lack of incentives and rewards | The scholarly system is heavily biased towards publications and secondary products such as data or code are rendered far less credit. | (Harley et al. 2010; Howison et al. 2015; Plantin et al. 2018) |
| Misuse, misinterpretation, liability concerns | Uncertainty over who is going to reuse the data and for what purposes and lack of understanding of the data and thus misuse. | (Meijer et al. 2017; Tenopir et al. 2015; Wallis et al. 2013) |
| Lack of skills | Lack of expertise and knowledge of tools to make their data available. | (Borgman 2015; European Commission 2019b; Meijer et al. 2017; OECD 2015) |
| Costs to input metadata | The effort and time-consuming activity of providing contextual information and detailed descriptions of the data. | (Edwards 2010; Holzner et al. 2009; OpenAire 2019) |

Nevertheless, despite such barriers, there is a consensus that sharing scientific data is beneficial, making it a clear objective for the research community at large. The reasons behind such consensus include (Table 2): to improve reproducibility; to accelerate scientific processes and research velocity; to increase scientific quality; to prevent scientific fraud; and to increase scientific productivity by reducing redundancy and innovation gains (e.g. Borgman 2015; Edwards et al. 2011; European Commission 2019a; OECD 2015; Tenopir et al. 2015).

*Table 2. Reasons for data sharing in science*

| Reasons sharing | Description | Source |
|---|---|---|
| Reproducibility | Sharing research data raises transparency and multiplies opportunities for the replicability of research findings. Making it easier to peer review data strengthens transparency and the potential of publishing negative results and enables accurate verifications of research findings. | (Baker 2015; Fecher et al. 2015; Lyon 2016; OECD 2015; Pujol Priego and Wareham 2019; Tenopir et al. 2015) |
| Accelerate scientific progress | The availability of the Gene Expression Omnibus (GEO) database at the US National Center for Biotechnology Information led to more than 1,150 published articles by third-party contributors by the end of 2010. | (Borgman 2015; Pasquetto et al. 2017; Piwowar et al. 2011) |
| Increase scientific quality | Sharing research data is related to the strength of the evidence supporting the results and the quality of the statistical results reporting. | (Wicherts et al. 2011) |
| Fraud prevention | Sharing research data contributes to the identification of scientific fraud and enables transparency and greater scrutiny of research. | (Kupferschmidt 2018) |
| Increase scientific efficiency | It increases the scientific efficiency of the research system by reducing duplication of costs and other costs stemming from data storage and transfer. More knowledge can be produced from the same data and thus increase returns on publicly-funded research. | (Lyon 2016; OECD 2015; Whyte and Pryor 2011) |
| Innovation gains | Data sharing fosters the reuse of data for R&D and innovation processes (e.g. in drug discovery processes). For instance, the use of data from PubMed Central at the US National Institutes of Health's repository has 17% unique daily users from companies versus 25% from universities. | (Khaladkar et al. 2017; Swan 2012) |

## 3.4 Theoretical underpinnings

Understanding the drivers for and barriers to sharing research data is both multifaceted and complex. We believe that they are shaped by the specific research community’s values and norms (cultural perspective) as well as the professional incentives that reconcile both individual and collective interests (rational perspective). Consequently, our study builds upon the notion of “epistemic culture” (Knorr Cetina 1999) and collective action theory (Hess and Ostrom 2003; Olson 2009; Ostrom 1990) to build a complementary perspective on the phenomenon.

### *3.4.1 Epistemic cultures*

Anthropologist Knorr Cetina (1999) coined the notion of *epistemic cultures* to describe “those amalgams of arrangements and mechanisms—bonded through affinity, necessity and historical coincidence—which, in a given field, make up how we know what we know” (Knorr Cetina 1999 p. 1). The notion of epistemic culture claims that the nature of scientific activities, types of reasoning, and practices of establishing evidence are variable across scientific fields. It is considered a *cultural* approach that disputes the ‘unity of science’ associated with the Vienna Circle (Knorr Cetina 1999 p. 3) and “reveals the fragmentation of contemporary science” (Mørk et al. 2008 p. 15). The main idea that Knorr Cetina argues is that different scientific fields exhibit different epistemic cultures.

The idea of different scholarly cultures can be drawn back to the (Fleck 1979 [1935]) idea of “styles of thought” shared by “thought collectives” (Knorr Cetina 1999) and also relates to a concept of “thought worlds” (Dougherty 1992) or the idea of “communities of knowing” (Boland and Tenkasi 1995). Haas (1992) also uses the notion of “epistemic communities” defining groups of people engaged in knowledge production. The general and universal idea across such notions is that knowledge is situated and local (Borgman 2012). "There is no 'view from nowhere'—knowledge is always situated in a place, time, conditions, practices, and understandings. There is no single knowledge but multiple bits of knowledge" (Cetina 2007; Gläser et al. 2015).

What makes Knorr Cetina's ideas attractive is that her definition of “culture” is rooted in *practice*, that is, when defining epistemic cultures, she designates the prevailing dynamics and aggregate patterns in scientists' practices. The “epistemic machinery” defines the shared tools, techniques, particular ontologies of instruments, conventional methods, and the architectures of shared empirical approaches that the epistemic subjects use to produce and distribute knowledge between them. She grounds the concept in the *making* of science, in practice and acts of making knowledge and the patterns in such practices. She constructs the concept of epistemic cultures, describing their interiorised process and arguing that scientists (or epistemic subjects) and the organisations and collectives that are part of the epistemic culture (e.g. labs and experiments) are shaped by conventional practices and these shared pieces of machinery of knowing, which also affects the nature of competition in the field (Knorr Cetina 2007).

Employing Knorr Cetina's lenses, we would expect that data sharing practices may be community-bound as a result of the epistemic culture of the community. Differences in data sharing practices across scientific communities would depend on whether the scientific community is more "communitarian" or "individualistic", using her terminology, resulting from how contributions are ascribed to individual scientists in the community and their norms and practices. A collective or communitarian epistemic culture compared to the individualised nature of another one may display predispositions to share and fewer concerns about individual incentives and rewards

The cultural explanation is useful when trying to account for the heterogeneity of data sharing practices across "field-specific research culture" (Gläser et al. 2015 p. 329). This is logical if we consider the long training cycles with which new members are trained, the specificities in the technological tools, the commonly accepted methods, particular financing sources, norms in collaboration dynamics, and how responsibility and authorship are assigned.

### *3.4.2 Collective action theory: managing the commons*

An alternative for explaining differences in data practices is to examine the mechanisms put in place by which self-interested researchers would contribute to a data *commons*. Commons designates a "resource shared by a group of people that is subject to social dilemmas" (Hess and Ostrom 2003). We bring an economic-rational perspective to the foreground of the data sharing conversation by revisiting classic collective action theory to uncover the intricate system of incentives and rewards behind the considerable amount of work needed to make data available to others and eventually FAIR.

Collective action theory has been widely used in sociology and economics to understand individuals' motivation to engage in collective action (Fulk et al. 2004; Monge et al. 1998). Research into collective action problems was originated with Olson's work in the classic Logic of Collective Action, which later Hardin (1968) developed with his thesis on the "tragedy of the commons" that argues how uncontrolled individual self-interested pursuits may corrupt the commons (Greco and Floridi 2004). In other words, the tragedy of the commons is an instantiation of the prisoner's dilemma (with *n*-people) where the rational pursuit of each self-interest results in suboptimal management of the commons (Greco and Floridi 2004; Fletcher and Zwick 2000; Ostrom 1986).

As Hardin (1982) describes, the community benefits if the individual perceives gains from their contribution to the commons. However, if no scientists perceive gains from contributing to the commons, the shared pool of resources is 'latent' and will not succeed by itself without external intervention. The social dilemma in contributing to the data commons arises when the incentive structure favours the free-riding of scientists on other contributions. Optimally, there should be a positive relationship between individual gains and individuals' contributions to the commons and the value of the commons and the collective resources that have been contributed. By increasing the number of contributors to the commons, the individual commitment to contribute is reinforced. In other words, there is *individual-collective interdependence*.

What makes collective action useful in understanding the scientific data sharing phenomenon is that the fundamental dynamic behind the commons is the prediction of individual gains by adjusting the values and costs associated with resource contribution (Fulk et al. 2004; Ostrom 1990; Vitali et al. 2018; Weill and Ross 2004).

## 3.5 Methods and Results

For the analysis and presentation of our data, we have followed the approach suggested by Venkatesh et al. (2013) to extract the most potential value from mixed-method research. As a result, we first present the method and results of the analysis of the survey data, we follow that with the case studies and thereafter synthesise the findings of both. The synthesis of the results in the discussion is a “bridging” process (Creswell 2018) where we seek to leverage the complementarities between the findings to enrich our empirical and theoretical understanding of scientific data practices.

### *3.5.1 Survey 2016 and 2018*

#### *3.5.1.1 Method and data*

We developed a large-scale global online survey collected in 2016 and 2018 in collaboration with Elsevier, and the academic collaboration of scholars to provide an Open Science Monitor for the European Commission. The survey data allow us to answer RQ1a—whether researchers share data—and RQ2a—how they do it.

The survey of 2016 was sent in June-July 2016 by Elsevier to researchers worldwide in all scientific disciplines. 1,162 researchers responded, which represented a 2.3% response rate. Responses were weighted by the research team to be representative of the researcher population (UNESCO counts of researchers, 2013). The margin of error for 1,162 responses was estimated ± 2.87% at 95% confidence levels (see prior analysis of the survey and full dataset in (Meijer et al. 2017). The full and raw data results from survey 2016 were available at DOI:10.17632/bwrnfb4bvh.1.

The survey of 2018 was sent in October-November 2018 to 40,991 individuals randomly selected from the Scopus author database while being weighted to be representative of the researcher population (UNESCO counts of researchers), to which 1,029 researchers responded (2.5% response rate). Appendix B provides the demographics of the survey respondents. Appendix C presents the full survey questionnaire.

The significant differences between the 2016 and 2018 survey questionnaires are four additional questions that were added in 2018 to assess the consequences of data sharing for scientists in their future collaborations with for-profit entities and other scientists. Finally, other minor modifications were introduced in the questionnaire to improve clarity.

We employ the survey data with a descriptive objective to analyse frequencies, averages, and patterns across researchers to obtain an overview of scientists' willingness to share data if they have done it, and through which means. The survey allows extracting such attitudes and practices by age, country, and discipline, while providing initial trends comparing results from 2016 and 2018.

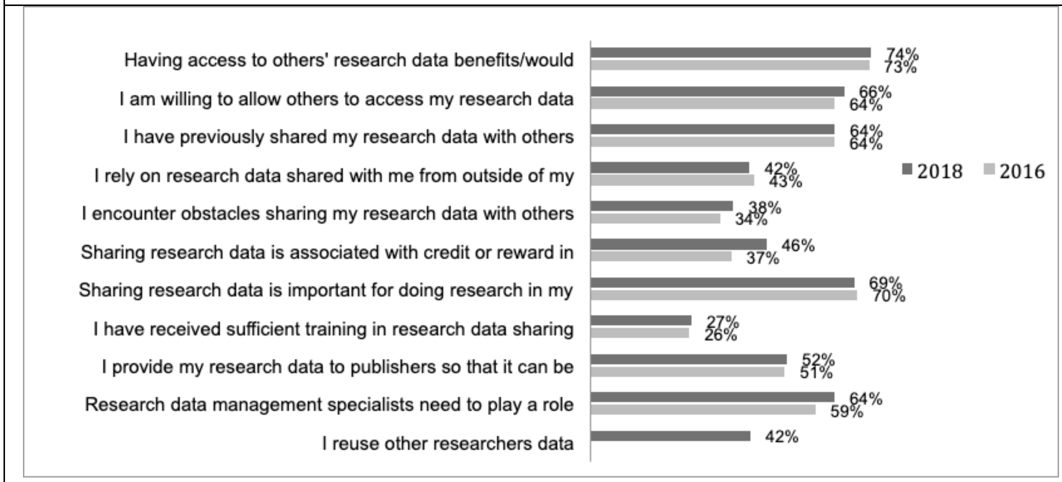
#### *3.5.1.2 Findings of the survey*

***Data sharing practices are steady***

Comparing survey results between 2016 and 2018 reveals that despite widespread support from policymakers and pressure from funding agencies, the number of academic researchers that declare making their data available remains stable, with no growth shown over the past two years (66%) (Figure 2). Although researchers acknowledge the benefits of data sharing, their practices are still limited, with one-third of researchers saying they do not share their data at all. While researchers acknowledge the benefits of sharing unpublished research data (74%), fewer are willing to share data (66%) or have previously shared their data (64%) (Figure 2).

*Figure 2. Researchers′ attitudes on data sharing*

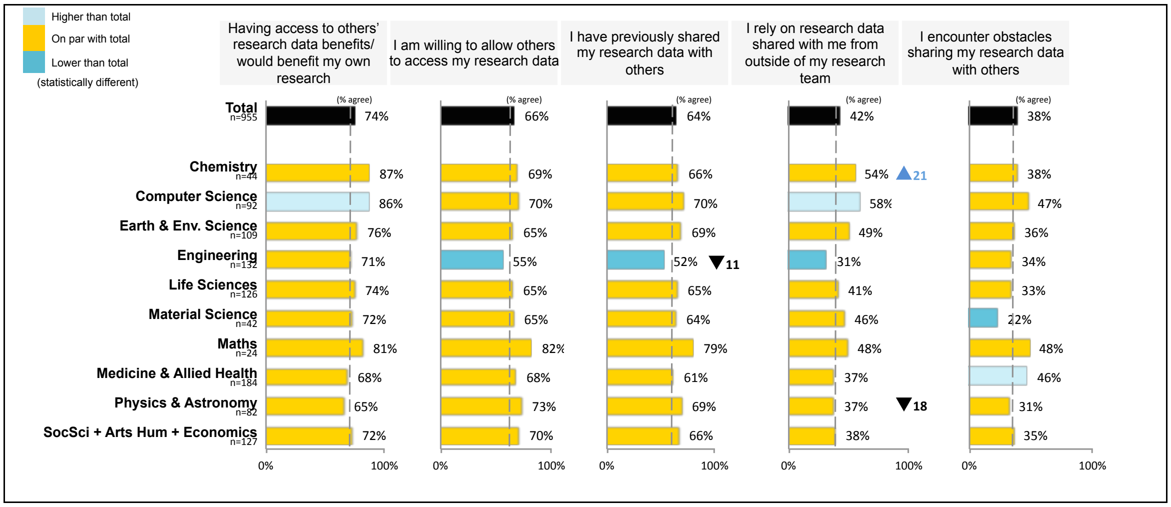
*Note: Corresponding question: Please think about the research data that typically is not published (e.g. not summary charts, tables or images), and indicate how much you agree or disagree with the following statements.*



***Data sharing varies across disciplines***

Data sharing practices are dependent on the field. The survey results show that data sharing activities are highly concentrated in math (79%), computer science (70%) physics and astronomy (69%) and life science (65%) (Figure 3). When we compare physics and astronomy to life science, we see that while 65% of scientists in physics and astronomy say that access to others' data would benefit their research, a larger number coming to 73% is willing to allow others to access their data. On the contrary, in life science, while 74% say they benefit from others' data, the number of scientists willing to allow others to access their data is lower (65%) (Figure 2). The same pattern displayed in life science disciplines of higher perceived benefit compared to the willingness to share is shown in the rest of disciplines except physics and maths.

*Figure 3. Attitudes to data sharing by discipline*



### Discriminatory sharing

The survey data show that most of the data sharing is carried out between collaborators on the same projects (80%), suggesting that researchers adopt a discriminatory approach, sharing data with selected partners on a case by case basis (Figure 4). Most scientists still rely on ad hoc and communicative exchanges to share their data instead of formal data repositories (14%), as "purpose-built, stored and ready for use" data (Edwards et al. 2011) (Figure 4). Although efforts have been made to improve metadata products, the results suggest that we will still see "informal, ad hoc, incomplete and contested processes of communicating about data" (Edwards et al. 2011), as only 14% of researchers share it through data repositories. Overall, a third of researchers say they do not share their data at all.

*Figure 4. Data sharing behaviour of researchers*

*Note: Corresponding question: Have you done any of the following with any or all of the research data that you used or created as part of your <u>last research project</u>? Shared directly with…*

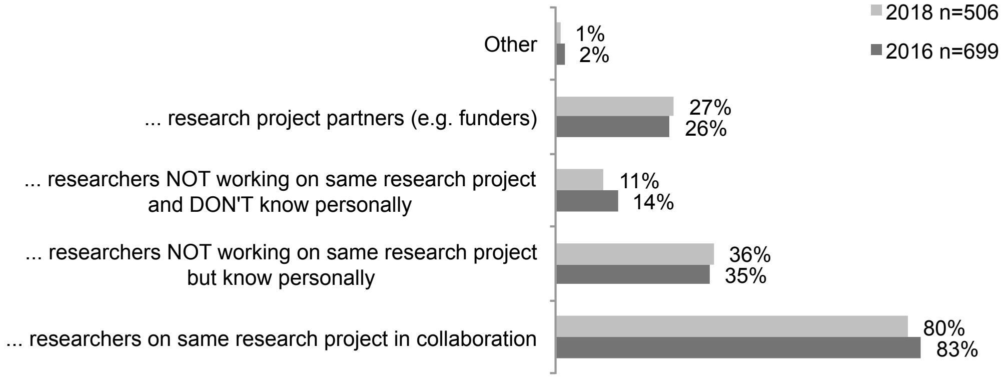
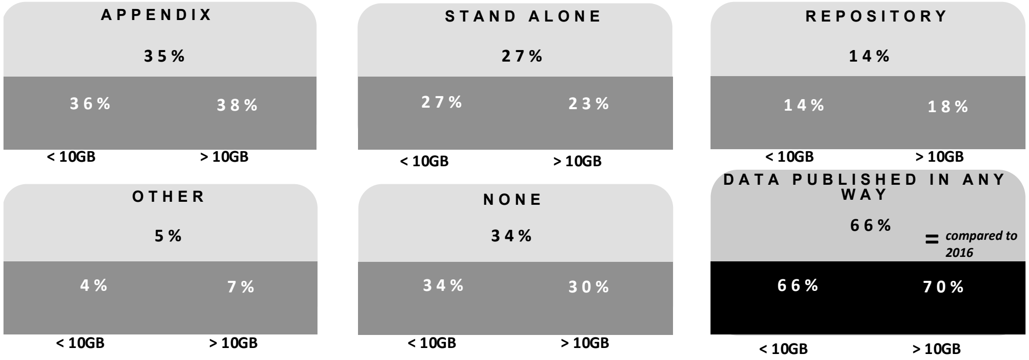


*Figure 5. Preferred ways for data sharing*

*Note: Corresponding question: Have you published the research data that you used or created as part of your last research project in any of the following ways? (See Q1f and Q1c1 in Appendix B)*

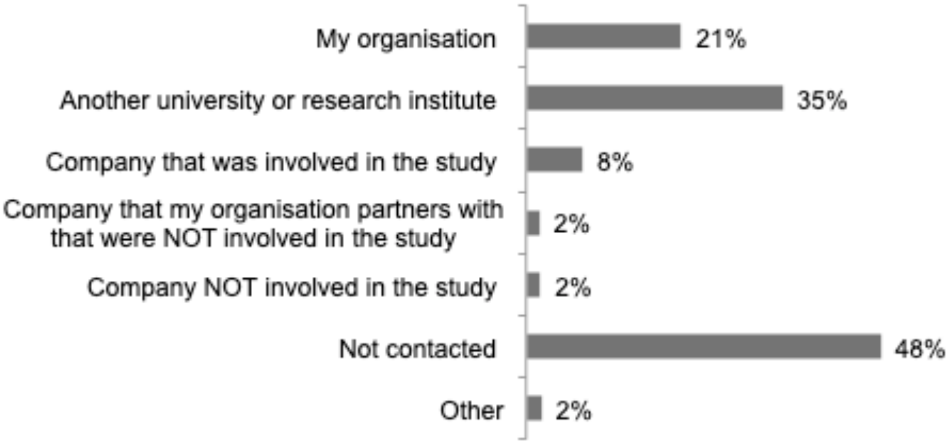


***Data sharing: an interactive process***

Over one-third of researchers were contacted by another university/institute after sharing their data. 10% were contacted by a company, and over one-third of researchers believe that sharing data promoted new collaborations with researchers in their discipline. By being contacted by other researchers, whether in public or for-profit organisations, we infer that others are accessing the data and are trying to understand or reuse it. It also suggests that metadata or the description and contextual information of the data is not enough in some cases or requires clarification with the scientist who generated the data (Figure 5).

*Figure 6. Follow up to data sharing*

*Corresponding question: Thinking about the most recent research project on which you shared data, did individuals outside of the research team contact you concerning the data that you shared? I was contacted by researchers from:*



In sum, what becomes clear from the survey results is that data sharing is a practice that varies significantly across disciplines (and we speculate subfields). Each discipline delineates collectively the tools that fulfill specific requirements for their community. Why are there such disparities across scientific disciplines? Which mechanisms are behind the high data sharing rates of physics or life sciences? And what can we learn from these fields with high scientific data sharing rates?

### 3.5.2 *Case study*

#### *3.5.2.1 Method and Data*

We follow this with a qualitative study of two cases sampled from two of the disciplines displaying high rates of data sharing, molecular biology and high energy physics, to find plausible explanations of factors enabling scientists' data sharing behaviours to enrich and extend the quantitative results from the survey (Creswell 2018). As such, the primary purpose for employing a sequential mixed approach in our study was to acquire complementary explanatory insights about scientific data sharing practices in the two empirical settings while providing opportunities for opening avenues for future research (Venkatesh et al. 2013). This offered us a holistic approach to the phenomenon.

The two case studies are thematically sampled (Creswell 2018) as representative of two different epistemic cultures to account for the cultural dimensions which ground and augment other formal mechanisms that influence how and why researchers share their data. To capture data practices behaviours of HEP and MB, we investigate their practices grounded in two information infrastructures. *Open Targets* is a microbiology consortium created in 2015 by for-profit, non-profit and research entities led by the European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL- EBI). *Reana* is led by the European Organization for Nuclear Research (CERN), established in 2018 as an infrastructure that embeds a set of platforms and services developed by the HEP community to share their data and code and foster reproducibility of scientific results.

Information infrastructures have been defined as "a digital library system based on commonly shared standards and containing information of both local and/or widespread interest" (Kahn and Cerf 1988 pp. 3) … "to augment our ability to search for, correlate, analyse and synthesise available information," (Kahn and Cerf 1988 p. 11.) Adding a social dimension, Constantinides (2012 p. 21) defines them as "efforts to integrate other computer-based and social systems, and to regulate and monitor processes that were previously performed in various, isolated settings." Our decision to focus on information infrastructures (as opposed to less-institutionalised data sharing practices) is based on the fact that the highest data sharing levels are in communities that actively use information infrastructures. As such, we believe that insight into the determinants of the most successful practices can be obtained by studying these infrastructures.

Both infrastructures are based upon data commons, i.e. data commons co-locate data, storage, and computing infrastructures with commonly used services and tools for analysing and sharing data to create interoperable resources for a different base of users (Grossman et al. 2016).

The study of both cases relies on diverse primary and secondary data sources, described in Table 3. Numerous discussions with managers from Open Targets and Reana were an integral part of the *Open Science Monitor* and shared by the European Commission services in separate reports (Pujol Priego and Wareham 2018, 2019).

As part of participation in the two additional EU H2020 funded projects, the authors benefited from extensive conversations with policymakers, research infrastructure managers, data architects, and programmers to discuss data sharing practices and future open research data (CS3MESH4EOSC part of the European Open Science Cloud, and ATTRACT funded by Research Infrastructure Innovation H2020-INFRAINNOV).

*Table 3. Details on Data Collection*

| | **MB - Open Targets** | **HEP-Reana and related platforms** |
|---|---|---|
| **Primary data sources** | 13 interviews with scientists and managerial team of Open Targets | 4 interviews with scientists and managerial team of Reana and related platforms.<br>4 interviews with CERN programmers and data architects. |
| **Observations** | Study visit to Genome Campus for Open Targets Open Days – workshop, working groups and social event (June 2019) | Study visits to CERN (2018, 2019, 2020).<br>Partner in H2020-funded CS3MESH4EOSC, a constituent project of the European Open Science Cloud https://cordis.europa.eu/project/id/863353<br>and ATTRACT https://attract-eu.com/<br>Interviews and discussions with open data-related services at CERN (Zenodo, Open Data Portal, CS3-ScienceMesh) |
| **Secondary data sources** | 41 publications<br>1 tutorial on OT infrastructure<br>3 outreach posts<br>19 release notes<br>6 posts<br>7 websites | Experiments data policy and guidelines:<br>CMS data policy<br>ALICE data policy<br>ATLAS data policy<br>LHCb data policy<br>OPERA data policy<br>CERN open data terms of use<br>22 guidelines in CERN open data portal<br>CERN Analysis Preservation Portal<br>Documentation from data preservation HEP projects<br>Joint declaration and task force documentation on HEP data preservation<br>Reana workshop presentations June 2018<br>12 runnable examples of Reana<br>6 publications<br>6 release notes<br>User guide<br>Administrator guide Developer guide<br>2 blog posts |

***Empirical context 1: Molecular biology and Open Targets***

The sequencing of the human genome (Human Genome Project, HGP) is recognised as “the largest undertaking in the history of biological science” (Chaguturu et al. 2014 p. 35). Not only did it transform biology into a data-driven science as a result of the deluge of new data and computational techniques, but it also opened the debate about research data sharing when Celera, a private undertaking, initially announced their intention to patent "fully-characterised important structures" amounting to 100–300 targets (Leonelli 2012). In March 2000, President Clinton announced that the data on the genome sequence should be made freely available to the entire research community. Some argue that in the post-HGP era, the human genome brought to biology a blueprint on research data sharing that other research communities need to follow. HGP propelled discourse on open research data to the forefront of molecular biology research (Leonelli, 2012) and spawned a new generation of information infrastructures to generate, integrate, and curate the growing data pools with other sources and commonly used tools and analytical methods for the research community (Grossman et al. 2016; Vamathevan et al. 2019). As a result, the discipline has been very active in developing data commons (Pujol Priego and Wareham 2018).

Open Targets (OT) was created in 2015 by the EMBL-EBI, Europe’s flagship laboratory for life science, with the Wellcome Sanger Institute, and pharmaceutical companies (Biogen, Celgene, GSK, Sanofi, Takeda) to accelerate knowledge about the links between genetic targets and disease development. The architecture, data policies, and procedures from researchers participating in OT provide insights about the mechanisms that effectively foster data sharing across the MB research community.

***Empirical context 2: High-Energy Physics and Reana***

Big scientific research infrastructures within High-Energy Physics such as CERN have a long tradition of embracing open data. Large volumes of data generated via expensive, unique, and extensive experiments make data preservation and reuse important. Reana is a reusable and reproducible research data analysis infrastructure created at CERN in 2018 to facilitate data and code reuse. The infrastructure sits on already existing platforms and services provided by CERN to the HEP community such as Zenodo, a free and open data repository, and the CERN open data portal, which are precedents to Reana infrastructure. The infrastructure generalises computational practices used by the HEP community and facilitates the adoption of workflow systems to run and reuse data analysis on remote compute clouds (Simko et al. 2018). CERN generated Reana to allow the different HEP experiments to adhere to FAIR principles and facilitate data sharing and reuse in the community. Reana allows the reuse and reinterpretation of the data shared by helping HEP scientists to structure their input data, their analysis code, containerised environments, and computational workflows to run the analysis on remote clouds (Pujol Priego and Wareham 2019). What makes Reana attractive is that the infrastructure helps to generalise computational practices employed by HEP scientists, thereby systematizing reproducibility. The infrastructure supports a plurality of “container technologies (Docker), workflow engines (CWL, Yadage), shared storage systems (Ceph, EOS) and compute cloud infrastructures (Ku-Kubernetes/OpenStack, HTCondor)” used by the HEP scientific community (Simko et al. 2018, p. 1).

The analysis of HEP and MB scientists around Open Targets and Reana infrastructures gives us an insight into how such culturally different communities are capable of actively sharing and reusing data.

#### *3.5.2.2 Findings of the case studies*

Preliminary observations about HEP and MB communities suggest two different epistemic cultures consistent with Knorr Cetina's work: HEP is more communitarian with MB more individualistic, using Knorr Cetina's terms. When looking at how HEP data flows are organised, we first realise the importance of the entity of “the experiment”. In HEP, few extensive, capital-intensive experiments have been designed and constructed over 20 odd years. For example, CERN currently hosts seven large experiments on the Large Hadron Collider, four of which are elaborate international collaborations (ATLAS, CMS, ALICE, LHCb).

By contrast, MB is organised around the “laboratory” or single institution, and consistent with what Knorr Cetina describes, molecular biologists are shaped by the conviction that they need to compete “for the priority of important findings” (Knorr Cetina, 1999), generating competition within and across labs.

When comparing how HEP and MB ascribe contribution to an individual scientist, we soon realise that in HEP there can be a vast number of authors as the construction and operation the experiments depends on many people, the record being over 5,000 authors on one article (Aad et al. 2015). In MB, although there are also challenges in ascribing results to individual scientists, the experiments are typically far less capital-intensive and permit differentiation in contributions within smaller teams. Finally, it is worth noting that some MB research is closer to commercial organisations (life sciences and pharma), where HEP is traditionally considered basic research with a more extended pathway towards any commercial outcome. Accordingly, we would expect a more competitive culture with less data sharing in MB than HEP. However, despite such differences in their epistemic cultures, both exhibit high levels of data sharing.

### *Open Targets*

OT was set up in 2015 under the umbrella of the EMBL-EBI and Sanger Institute within the collaboration of large pharmaceutical companies. By applying lean user experience (UX) design methods, OT infrastructure was developed to search, assess, and integrate a vast quantity of genetic and biological data to support target-centric and disease-centric inquiries. At present, OT contains more than 27,717 targets, 7,999,050 associations, 13,445 diseases, and 20 data sources (Open Targets, 2020).

OT displays a ***modular*** infrastructure containing different layers, access rights, and data standards that employ different mechanisms for researchers to be able to simultaneously share their data and comply with the norm in the post-HGP community era, while simultaneously allowing them to grasp the competitive benefits of being the generators of the data. The stratified architecture grants different access rights to the data, where data generators are granted access to a hidden layer augmented by a public layer accessible (with different rights) to any researcher willing to reuse the data.

The modular architecture with different access rights combines ***time dilation*** between the generation of the data and the publication of the data in the infrastructure that spans on average two years and could be considered as a considerably long "embargo period".

Finally, the information infrastructure acts as a "boundary organisation" (O'Mahony and Bechky 2008), that is, "structures capable of effectively mediating between disparate constituencies and establishing common ground among the differing interests in play" (Perkmann and Schildt 2015 p. 1134).

The two mechanisms are combined with normative governance rules provided by the infrastructure on data access and reuse, where the ownership and responsibilities over the data are explicit. These two mechanisms fit in a "logic of exchange" that seeks to maximise benefits for the researchers (that is, the potential of the data for being reused and the competitive advantage of data generators) while minimizing the costs of sharing data. This optimisation is completed by providing the protocols and data standards required to minimise the efforts of data reuse and increasing the value of the data aggregated while reducing the uncertainty over who controls and owns the data. The fact that for-profit companies form a significant part of the OT consortium suggests that the mechanisms are effective in balancing incentives to scientists to contribute while mitigating the risks of a competitive loss to other re-users of their data.

### *Reana*

CERN built Reana as an infrastructure for the HEP research community to foster the reuse of the data generated via the large HEP experiments, which built upon data access and preservation policies agreed within the main experiments. While the data policies may differ slightly across experiments, they all stratify the data generated by the HEP community in four main layers: a) data directly related to publications, which include the complete documentation for published results; b) simplified data formats devoted to training exercises within the physics community; c) reconstructed data, simulations, and software analysis to facilitate research analysis; and finally, d) the raw level data and associated software, which permits access to the full potential of the experimental data reuse (Pujol Priego and Wareham 2019). Data sharing is concentrated for data layers (b) and (c) described above. Raw data (d) is not made available to other researchers to reuse for declared pragmatic reasons. For instance, one of the core CERN experiments, CMS (Compact Muon Solenoid) produces on average 1 petabyte (100 gigabytes) of "raw" data per second, and similar data volumes characterise other experiments. As the LHC data policy explains: "*It is practically impossible to make the full raw data-set from scientific endeavours of the scale of high-energy physics easily usable in a meaningful way outside of the collaboration. […]It should be noted that, for these reasons, direct access to the raw data is not even permitted to individuals within the collaboration, and that instead the production of reconstructed data is performed centrally.*"

Experiments also foresee a *time dilation* between the generation of the experimental data and the moment to share it with the external research community. These periods, also referred to as embargo periods, allow the data generators within the experiment to publish.

As explained in the LHC experiment data policy: "*In general data will be retained for the sole use of the collaboration for a period commensurate with the substantial investment in the effort needed to record, reconstruct and analyse those data. After this period, some portion of the data will then be made available externally, with this proportion rising with time. The CB will keep such periods and proportions under review and may reconsider whether they should be varied in the light of experience. In the first instance, access will be granted to portions of the DST data five years after data is taken. The portion of the data which LHCb would normally make available is 50% after five years, rising to 100% after ten years.*"

One of the significant concerns within the HEP community related to data sharing is not credit but more importantly that the reuse may lead to an inflation of incorrect results. Consistent with what researchers claim in a dedicated study on data preservation in HEP (Holzner et al. 2009), "45.0% of the respondents are 'very concerned' or 'gravely concerned' that data re-use may in general lead to an inflation of incorrect results.

Interestingly, experimentalists are by far more concerned (51.3%) than theorists (29.0%)" (p. 7).

Different research teams employ a variety of tools supporting their computational workflows. By analyzing the different scientific pipelines, Reana has abstracted the steps that scientists follow and provides a "simple 'shell script' use case where commands are run sequentially, and each step produces outputs for the next step" (Simko et al. 2018 p. 2). As a result, Reana allows structuring research data analysis in a reusable way making it possible to instantiate computational workflows remotely in the cloud with the support of a set of workflows specifications, storage systems, and container technologies.

*"Our own experience from opening up vast volumes of data is that openness cannot simply be tacked on as an afterthought at the end of the scientific endeavour. Besides, openness alone does not guarantee reproducibility or reusability, so it should not be pursued as a goal in itself. Focusing on data is also not enough: it needs to be accompanied by software, workflow, and explanations, all of which need to be captured throughout the usual iterative and closed research lifecycle, ready for a timely open release with the results"* (Chen et al. 2019).

The main idea behind Reana's infrastructure is to preserve software and data workflows so that they can enhance collaborative scientific work and as a way of grasping the knowledge behind a given analysis during the review process (Dphep Study Group 2009). Such data sharing process and preservation techniques are embedded in the Reana framework and can be translated into new analysis methods for future HEP research. Reana was set up to seek the reuse of experimental data first by the large community of collaborators themselves and then extend it.

The Reana cases describe the challenges for the HEP community to share and reuse their data which implies a shift from in-depth documenting and archiving of analysis towards preservation based on simulation and software containers. For instance, the CMS experiment preserves "the reconstructed data and simulations by keeping available a copy of the data reconstructed with the best available knowledge of the detector performance and conditions for each period of data-taking a virtualised computing environment, compatible with the software version with which the original data can be analysed, is provided and maintained" (Dphep Study Group 2009 p. 7).

Reana acts as a boundary organisation or "interface" to the experiment knowhow so that other researchers outside the experiment can reuse it. Data policies in HEP are decentralised at the experiment level, instead of at the infrastructure level, and Reana builds on top of the data rights and responsibilities agreed within every single experiment.

In Reana, we also find meaningful roles for the mechanisms of modularity (i.e. levels of data with different data access) and time dilation. While normative governance defining data access rights and responsibilities exists, it is not provided at the infrastructure-level but rather at the experiment-level, making the infrastructure respect distinct data policies. Table 4 in the Appendix provides a detailed description of the progression of our empirical analysis towards the theoretical constructs of the two mechanisms, modularity and time dilation, as well as the normative role of the infrastructure/experiments.

*Table 4. Theoretical progression of our analysis*

| **Illustrative examples of empirical observations from data sources** | | **How they are similar** | **How they differ** |
|---|---|---|---|
| **MB- Open Targets** | **HEP- Reana** | **Identification of same theoretical construct** | **Theoretical observation** |
| *"When data is ready, we integrate it into the platform; we need to wait until it is ready and publicise and then we enter it in the platform at that point and release it" OT4.*<br>*"The platform team (in charge of releasing the data) get to see the type of data very early in the process. They have sample data, and they discuss the format. We also have a UX specialist, to understand what the deliverable is and how we manifest it in the platform so that people can use it to make a decision. This is the foundation for data specification. How are we going to receive it? What does it look like, how will it be processed? The discussions are very early on, and we try to get mock-ups very early on, to gather feedback from the consortium partners but also other users, and then we kind of refine them that as we go along (...) It is a moving target, as some of the projects do not know what the data will look like, so we have monthly meetings." OT4* | "New data will enter the portal once the embargo periods for them are over." (CERN Open Data Portal)<br>"The first data release of 2010 data took place in 2014, as a stress-test exercise of the entire preservation, re-use, and access chain. This release was followed by a full analysis of the procedure, which was endorsed by the Collaboration Board in 2015, and regular data releases, accompanied by appropriate simulated data, each approved by the Collaboration Board, are now taking place" (CMS April 2018) | **Time dilation**<br>*(Mechanism 1)* | The embargo period of HEP is around 5 to 10 years, depending on the experiments.<br>In OT, the time dilation between the generation of the data and release in OT is 18 months to two years.<br>After the embargo period in HEP, only a % of the data is agreed to be released.<br>In OT, all the data generated is shared in OT infrastructure. |
| *"So, we have a platform that is public and open to everybody. Then, for the experimental projects, the partners share the data while they are creating it in Google buckets".*[6]<br>*"We have an intranet for the consortium partners. It is an information exchange between* | "Open access to its data by people outside the collaboration can be considered at four levels of increasing complexity." | **Modularity**<br>*(Mechanism 2)* | HEP establishes four layers of data: raw data is not released, while more elaborated versions of data are opened (level 2 in open data portal and reused in Reana; level 1 from publications through HEP library systems). |

[6] A bucket in Google cloud storage is a primary container that holds data. Owners of buckets control access to the data.

|  |  |  |  |
|---|---|---|---|
| *partners (...) The intranet has a link with the platform, and it is used for the general governance of the projects. As we go through the project call processes, there are page proposals to share the details. It is like a one-stop-shop for the whole portfolio of projects." OT4.* |  |  | Raw data from target associations with metadata is released in OT. However, the aggregations with data related to the next steps of the drug discovery process (e.g. proprietary compound libraries) remain closed. |
| *"There is a need to coordinate the integration of data into OT, both from the projects that generate data but also with the data providers such as Chembl and Uniprot and all the data that goes into the platform to keep it up to date. We also work with the developer team that creates some of the features that users will use to visualise the data coming through." OT4* | — | **Boundary organisation**<br>*(Mechanism 3)*<br>*The infrastructure:*<br>*Dissipates uncertainties over data ownership, control, access rights and rights to reuse. Defines and agrees on a clear data policy amongst OT participants.* | - The boundary organisation and what makes the interface that mediates the data flows between researchers and establishes the rules, responsibilities, and drivers in data policies varies in the two cases.<br>- The prominent role of the experiment in HEP, which decides rights and responsibilities across data. These rules prevail across infrastructures, including Reana. The competition over the data is not between scientists but between experiments.<br>- In MB, the different experimental projects need to comply with the data governance and rules of the OT infrastructures, which establish the protocols to avoid unintended spill overs and a regulated process to release the data generated. |
| – | "The data preservation process should follow well-defined policies, defined as soon as possible during the lifetime of the collaborations, and possibly embedded in a global HEP data preservation initiative."<br>"For the widest possible re-use of the data, while protecting the Collaboration's liability and reputation, data will be released under the emerging standard Creative Commons CC0 waiver."[7] | *The experiment:*<br>*Dissipates uncertainties over data ownership, control, access rights and rights to reuse. Defines and agrees on a clear data policy that prevails across infrastructures.* |  |

[7] Creative Commons License: http://creativecommons.org/publicdomain/zero/1.0.

## 3.6 Discussion

### *3.6.1 Implications for theory*

Data sharing is desired by the research community at large. 74% of researchers say that having access to other data would benefit them. Nevertheless, the number of researchers in our survey who have shared their data remains stable from 2016 to 2018, despite all policy activities, funders' efforts, and investments put in place to foster research data sharing. Admittedly, two years may not be indicative of longer trends, but the lack of any meaningful difference suggests that the uptake of data sharing practices is slow. Our analysis further suggests that there is no homogenous explanation. Slow adoption of data sharing comes from an intertwined web of varied cultures and rational pursuits. Where HEP and MB have significantly different epistemic cultures, research infrastructures, and scientific practices, both communities have established information infrastructures with mechanisms designed to mitigate the domain-specific costs and facilitate data sharing and reuse.

In this respect, the cultural explanation usually employed to justify data sharing differences across academic communities is only partially adequate. Both HEP and MB have professional norms characterised by some level of self-interested "exchange logic" which can deter scientists from absorbing the additional costs of data sharing with no apparent benefits.

Our case studies have examined two different information infrastructures that have enacted mechanisms to align scientists' professional incentives with data sharing practices across more individualist and communitarian scientific communities. That MB and HEP have two substantially different epistemic cultures is understandable given the vastly different research infrastructures and scientific practices: the enormous scale of many HEP experiments requires substantial organisations where individual contributions are difficult to account for. MB, by contrast, is conducted in smaller teams with less capitally-intensive research infrastructures, clearly influencing the allocation of academic merit and professional status.

Our analysis of Open Targets and Reana offers insight into how these differences can be accommodated in two different information infrastructures. Both Open Targets and Reana employ modularity, time dilation, and explicit governance to align the private interests of the scientists with the collective interests of their communities. In this respect, there is a consensus that scientific data should be a public good. Scientists, however, need some assurance of the recognition of their scientific endeavours. Towards this, scientific communities can consider adjusting how they allocate professional merit to recognise the cultivation and publication of datasets as a legitimate professional contribution. However, the mere publication of datasets is insufficient to address the challenges of reproducibility and scientific efficiency. Appropriate governance and policies enacted in the information infrastructures can address the needs of metadata as well as the risks of data misuse and liability specific to the scientific community.

In Table 5, we summarise the results and suggest some normative implications that we discuss next in implications for policy and practice.

*Table 5. Summary of findings and normative implications*

| **Research Questions** | **Findings** | **Normative Implications** |
|---|---|---|
| RQ1a: Do researchers share their data? | 66% of researchers say they make their data available. The % remains stable, with no growth shown over the past two years. | Funding and policymaker requirements for data sharing have little effect in the short-term.<br>Alternative approaches (e.g. scientific-community based mechanisms) may be more effective in promoting data sharing practices. |
| RQ1b: How do they share their data? | Data sharing varies significantly across disciplines.<br>Most of the data sharing is carried out between collaborators on the same projects (80%), suggesting that researchers adopt a discriminatory approach by sharing data with selected partners. Only 14% of researchers share theirs through data repositories.<br>Metadata is an interactive process: over one-third of researchers were contacted by another university/institute after sharing their data, and 10% were contacted by a company. | Disparities in data sharing practices suggest that there is no one-size-fits-all in data sharing policies.<br>Knowing with whom you share the data (or delegated mechanisms of trust) is relevant for researchers to share. General repositories may not be the means to enforce data sharing amongst research communities.<br>Releasing data is the beginning but not the end; it leads to interactive exchanges with data re-users. This implies more unexpected effort but also potential new collaborations. |
| RQ2: What mechanisms have emerged to enable researchers to share their data? | Both communitarian and individualistic scientific communities (different epistemic cultures) employ *three* mechanisms (with some variation) to enable data sharing in both scientific communities:<br>-Modularity<br>-Time dilation<br>-Boundary organisation to establish transparent data governance and mediate the identification of the "bona fide" researcher. | Sharing data is not a dichotomous decision, but rather it needs to establish a *degree* towards *what* data you share (modularity), and *when* you share it (time dilation - embargos).<br>Scientific communities can consider adjusting professional norms to recognise data sharing as a legitimate contribution. |

### *3.6.2 Implications for policy and practice*

This study contributes to the current research policy debate that is examining the potential policy interventions to increase data sharing across scientists. Survey data combined with the insights from the two case studies suggest that one size does not fit all, in particular for such a complex phenomenon as an intricate system of incentives and rewards, combined with historical and cultural accounts that shape the diverse research practices.

The insights from the two case studies also guide other disciplines displaying less data sharing practices in the survey such as ours. In particular, our research communities could leverage the potential of research data sharing to increase the transparency and reproducibility of our research practices. Towards this, the establishment of information infrastructures in the social, economic, and managerial sciences can adopt mechanisms such as modularity and time dilation, with the appropriate mechanisms concerning metadata, reuse, and liability that are critical to ensure that data sharing objectives are achieved.

## 3.7 Conclusion

"The Republic of Science is a Society of Explorers. Such a society strives towards an unknown future, which it believes to be accessible and worth achieving. In the case of scientists, the explorers strive towards a hidden reality, for the sake of intellectual satisfaction. And as they satisfy themselves, they enlighten all men and are thus helping society to fulfill its obligation towards intellectual self-improvement" (Polanyi 1962 p. 19). Data sharing is a practice that is intended for the collective benefit of the “society of explorers”. In this respect, scientific communities are far from being united, but display heterogeneous practices and norms in the way science is produced and how merit and status are allocated. The need for greater transparency and reproducibility, combined with advances in ICT, render data sharing a clear choice for scientific policymakers and funders. Yet reasons for its gradual and disparate adoption are less obvious. A delicate system of mechanisms needs to be established to align individual and collective incentives. Moreover, these will differ across scientific communities. The use of modularity, time dilation, and appropriate policies are pivotal in the information infrastructures created by the scientific disciplines currently at the forefront of scientific data sharing. Other academic communities that seek to follow these examples can apply these mechanisms in a manner accordant with their own epistemic cultures and professional practices.

## References

Aad, G. et al. (ATLAS Collaboration, CMS Collaboration) *Phys. Rev. Lett*. 114, 191803 (2015).

Atkins, D. E., Droegemeier, K. K., Feldman, S. I., Garcia-Molina, H., Klein, M. L., Messerschmitt, D. G., Messina, P., Ostriker, J. P., and Wright, M. H. 2003. *Revolutionizing Science and Engineering Through Cyberinfrastructure*.

Baker, M., 2015. "First Results from Psychology's Largest Reproducibility Test," *Nature News*. (https://doi.org/10.1038/nature.2015.17433).

Bhardwaj, A., Bhattacharjee, S., Chavan, A., Deshpande, A., Elmore, A. J., Madden, S., and Parameswaran, A. G. 2014. "DataHub: Collaborative Data Science & Dataset Version Management at Scale," *ArXiv:1409.0798 [Cs]*. (http://arxiv.org/abs/1409.0798).

Boland, R. J., and Tenkasi, R. V. 1995. "Perspective Making and Perspective Taking in Communities of Knowing," *Organization Science* (6:4), pp. 350–372. (https://doi.org/10.1287/orsc.6.4.350).

Borgman, C. L. 2010. *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*, MIT Press.

Borgman, C. L. 2012. "The Conundrum of Sharing Research Data," *Journal of the American Society for Information Science and Technology* (63:6), pp. 1059–1078. (https://doi.org/10.1002/asi.22634).

Borgman, C. L. 2015. *Big Data, Little Data, No Data: Scholarship in the Networked World*, Cambridge, United States: MIT Press. (http://ebookcentral.proquest.com/lib/georgetown/detail.action?docID=3339930).

Bowker, G. C. 1999. *Sorting Things Out: Classification and Its Consequences*, Inside Technology, Cambridge, Mass: MIT Press.

Carillo, M. R., and Papagni, E. 2014. "'Little Science' and 'Big Science': The Institution of 'Open Science' as a Cause of Scientific and Economic Inequalities among Countries," *Economic Modelling* (43), pp. 42–56. (https://doi.org/10.1016/j.econmod.2014.06.021).

Cetina, K. K. 2007. "Culture in Global Knowledge Societies: Knowledge Cultures and Epistemic Cultures," *Interdisciplinary Science Reviews* (32:4), Taylor & Francis Ltd, pp. 361–375. (https://doi.org/10.1179/030801807X163571).

Chaguturu, R., Murad, F., and Murad, F. 2014. *Collaborative Innovation in Drug Discovery: Strategies for Public and Private Partnerships*, Somerset, United States: John Wiley & Sons, Incorporated. (http://ebookcentral.proquest.com/lib/georgetown/detail.action?docID=1662193).

Chen, X., Dallmeier-Tiessen, S., Dasler, R., Feger, S., Fokianos, P., Gonzalez, J. B., Hirvonsalo, H., Kousidis, D., Lavasa, A., Mele, S., Rodriguez, D. R., Šimko, T., Smith, T., Trisovic, A., Trzcinska, A., Tsanaktsidis, I., Zimmermann, M., Cranmer, K., Heinrich, L., Watts, G., Hildreth, M., Lloret Iglesias, L., Lassila-Perini, K., and Neubert, S. 2019. "Open Is Not Enough," *Nature Physics* (15:2), Nature Publishing Group, pp. 113–119. (https://doi.org/10.1038/s41567-018-0342-2).

Collins, F. S. 2003. "The Human Genome Project: Lessons from Large-Scale Biology," *Science* (300:5617), pp. 286–290. (https://doi.org/10.1126/science.1084564).

Connolly, T., B. K. Thorn. 1990. Discretionary databases: Theory, data, and implications. J. Fulk, C. Steinfield, eds. Organizations and Communication Technology. Sage Publications, Newbury Park, CA, 219–233.

Constantinides, P. 2012. *Perspectives and Implications for the Development of Information Infrastructures*, IGI Global.

Constantinides, P., and Barrett, M. 2015. "Information Infrastructure Development and Governance as Collective Action," *Information Systems Research* (26:1), pp. 40–56. (https://doi.org/10.1287/isre.2014.0542).

Consultative Committee for Space Data Systems (2002) *Reference Model for an Open Archival Information System (OAIS). Recommendation for space data system standards.* Available: http://public.ccsds.org/ publications/RefModel.aspx. Accessed 2013 Apr 2.

Cragin, M. H., Palmer, C. L., Carlson, J. R., and Witt, M. 2010. "Data Sharing, Small Science and Institutional Repositories," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* (368:1926), pp. 4023–4038. (https://doi.org/10.1098/rsta.2010.0165).

Creswell, J. W. 2018. *Designing and Conducting Mixed Methods Research*, (Third edition.), Thousand Oaks, California: SAGE.

Demir, I., and Murtagh, M. J. 2013. "Data Sharing across Biobanks: Epistemic Values, Data Mutability and Data Incommensurability," *New Genetics & Society* (32:4), Routledge, pp. 350–365. (https://doi.org/10.1080/14636778.2013.846582).

Dougherty, D. 1992. "Interpretive Barriers to Successful Product Innovation in Large Firms," *Organization Science* (3:2), pp. 179–202. (https://doi.org/10.1287/orsc.3.2.179).

Dphep Study Group. 2009. “Data Preservation in High Energy Physics,” *ArXiv:0912.0255 [Hep-Ex, Physics: Physics]*. (http://arxiv.org/abs/0912.0255).

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

Edwards, P. N. 2019. “Knowledge Infrastructures under Siege : Climate Data as Memory, Truce, and Target,” *Data Politics*, March 13. (https://www.taylorfrancis.com/, accessed March 8, 2020).

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

EIROforum IT working group. 2013. “E-Infrastructure for the 21st Century,” Zenodo, November 8. (https://doi.org/10.5281/zenodo.7592).

Eisenhardt, K. M. 1989. “Building Theories from Case Study Research,” *Academy of Management Review* (14:4), pp. 532–550. (https://doi.org/10.5465/amr.1989.4308385).

European Commission. 2014. “Data Management - H2020 Online Manual.” (https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm, accessed March 10, 2020).

European Commission. 2019a. “Facts and Figures of Open Research Data,” *European Commission - European Commission*. (https://ec.europa.eu/info/research-and-innovation/strategy/goals-research-and-innovation-policy/open-science/open-science-monitor/facts-and-figures-open-research-data_en, accessed April 19, 2019).

European Commission. 2019b. “Cost-Benefit Analysis for FAIR Research Data : Cost of Not Having FAIR Research Data.," Website, Website, January 16. (https://op.europa.eu:443/en/publication-detail/-/publication/d375368c-1a0a-11e9-8d04-01aa75ed71a1).

Fecher, B., Friesike, S., and Hebing, M. 2015. "What Drives Academic Data Sharing?" *PLOS ONE*, p. 25.

Fleck, L. 1979. *Genesis and Development of a Scientific Fact*, Chicago: University of Chicago Press.

Fletcher, J. A., and Zwick, M. 2000. *Simpson’s Paradox Can Emerge from the N-Player Prisoner’s Dilemma: Implications for the Evolution of Altruistic Behavior*, p. 19.

Friesike, S., Widenmayer, B., Gassmann, O., and Schildhauer, T. 2015. "Opening Science: Towards an Agenda of Open Science in Academia and Industry," *The Journal of Technology Transfer* (40:4), pp. 581–601. (https://doi.org/10.1007/s10961-014-9375-6).

Fulk, J., Heino, R., Flanagin, A. J., Monge, P. R., and Bar, F. 2004. "A Test of the Individual Action Model for Organizational Information Commons," *Organization Science* (15:5), pp. 569–585. (https://doi.org/10.1287/orsc.1040.0081).

Gitter, D. M. 2010. "The Challenges of Achieving Open-Source Sharing of Biobank Data," *Biotechnology Law Report* (29:6), Mary Ann Liebert, Inc., publishers, pp. 623–635. (https://doi.org/10.1089/blr.2010.9909).

Gläser, J., Bielick, J., Jungmann, R., Laudel, G., Lettkemann, E., Petschick, G., and Tschida, U. 2015. "Research Cultures as an Explanatory Factor," *Österreichische Zeitschrift Für Soziologie* (40:3), pp. 327–346. (https://doi.org/10.1007/s11614-015-0177-3).

Greco, G. M., and Floridi, L. 2004. "The Tragedy of the Digital Commons," *Ethics and Information Technology* (6:2), pp. 73–81. (https://doi.org/10.1007/s10676-004-2895-2).

Grossman, R. L., Heath, A., Murphy, M., Patterson, M., and Wells, W. 2016. "A Case for Data Commons: Toward Data Science as a Service," *Computing in Science & Engineering* (18:5), pp. 10–20. (https://doi.org/10.1109/MCSE.2016.92).

Haas, P. M. 1992. "Introduction: Epistemic Communities and International Policy Coordination," *International Organization* (46:1), Cambridge University Press, pp. 1–35. (https://doi.org/10.1017/S0020818300001442).

Harley, D., Acord, S. K., and Earl-Novell, S. 2010. *Peer Review in Academic Promotion and Publishing: Its Meaning, Locus, and Future*, Center for Studies in Higher Education. (https://eric.ed.gov/?id=ED512030).

Hess, C., and Ostrom, E. 2003. "Ideas, Artifacts, and Facilities: Information as a Common-Pool Resource," *Law and Contemporary Problems* (66:1/2), pp. 111–145.

Hey, T. 2009. *The Fourth Paradigm: Data-intensive Scientific Discovery*, Redmond, Washington: Microsoft Pr.

Hilgartner, S. 2013. "Constituting Large-Scale Biology: Building a Regime of Governance in the Early Years of the Human Genome Project," *BioSocieties* (8:4), pp. 397–416. (https://doi.org/10.1057/biosoc.2013.31).

Hochachka, W. M., Fink, D., Hutchinson, R. A., Sheldon, D., Wong, W.-K., and Kelling, S. 2012. "Data-Intensive Science Applied to Broad-Scale Citizen Science," *Trends in Ecology & Evolution* (27:2), pp. 130–137. (https://doi.org/10.1016/j.tree.2011.11.006).

Holzner, A., Igo-Kemenes, P., and Mele, S. 2009. “First Results from the PARSE. Insight Project: HEP Survey on Data Preservation, Re-Use and (Open) Access,” *ArXiv:0906.0485 [Hep-Ex, Physics: Physics]*. (http://arxiv.org/abs/0906.0485).

Howison, J., Deelman, E., McLennan, M. J., Ferreira da Silva, R., and Herbsleb, J. D. 2015. “Understanding the Scientific Software Ecosystem and Its Impact: Current and Future Measures,” *Research Evaluation* (24:4), Oxford Academic, pp. 454–470. (https://doi.org/10.1093/reseval/rvv014).

Järvenpää, S. L., and Markus, M. L. 2018. *Data Perspective in Digital Platforms: Three Tales of Genetic Platforms*, presented at the Proceedings of the 51st Hawaii International Conference on System Sciences, p. 10.

Kahn, R. E., and Cerf, V. G. 1988. *An Open Architecture for a Digital Library System and a Plan for Its Development.*, (The digital library project vol. 1: The world of knowbots.), Reston, VA: Corporation for National Research Initiatives, p. 48.

Kallinikos, J., and D Constantinou, I. 2015. "Big Data Revisited: A Rejoinder," *Journal of Information Technology* (30:1), Sage Publications Ltd, pp. 70–74. (https://doi.org/10.1057/jit.2014.36).

Kallinikos, J., and Tempini, N. 2014. “Patient Data as Medical Facts: Social Media Practices as a Foundation for Medical Knowledge Creation,” *Information Systems Research* (25:4), INFORMS: Institute for Operations Research, pp. 817–833. (https://doi.org/10.1287/isre.2014.0544).

Kaplan, F. 2015. “A Map for Big Data Research in Digital Humanities,” *Frontiers in Digital Humanities* (2), Frontiers. (https://doi.org/10.3389/fdigh.2015.00001).

Kellogg, K. C., Orlikowski, W. J., and Yates, J. 2006. “Life in the Trading Zone: Structuring Coordination Across Boundaries in Postbureaucratic Organizations,” *Organization Science; Linthicum* (17:1), pp. 22–44.

Khaladkar, M., Koscielny, G., Hasan, S., Agarwal, P., Dunham, I., Rajpal, D., and Sanseau, P. 2017. “Uncovering Novel Repositioning Opportunities Using the Open Targets Platform,” *Drug Discovery Today* (22:12), pp. 1800–1807. (https://doi.org/10.1016/j.drudis.2017.09.007).

King, G. 2007. “An Introduction to the Dataverse Network as an Infrastructure for Data Sharing,” *Sociological Methods & Research* (36:2), pp. 173–199. (https://doi.org/10.1177/0049124107306660).

Kitchin, R. 2014. *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*, SAGE.

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

Kupferschmidt, K. 2018. “Researcher at the Center of an Epic Fraud Remains an Enigma to Those Who Exposed Him,” *Science*. (https://www.sciencemag.org/news/2018/08/researcher-center-epic-fraud-remains-enigma-those-who-exposed-him).

Lazer, E. 2009. *Resurrecting Pompeii*, Routledge.

Lecarpentier, D., Wittenburg, P., Elbers, W., Michelini, A., Kanso, R., Coveney, P., and Baxter, R. 2013. “EUDAT: A New Cross-Disciplinary Data Infrastructure for Science,” *International Journal of Digital Curation* (8:1), pp. 279–287.

Lyon, L. 2016. “Transparency: The Emerging Third Dimension of Open Science and Open Data,” *LIBER Quarterly* (25:4), pp. 153–171. (https://doi.org/10.18352/lq.10113).

Maunsell, J. 2010. *Announcement Regarding Supplemental Material | Journal of Neuroscience*. (https://www.jneurosci.org/content/30/32/10599.short).

Markus, M. L. 1990. Toward a "critical mass" theory of interactive media. J. Fulk, C. Steinfield, eds. *Organizations and Communication Technology*. Sage Publications, Thousand Oaks, CA, 194–218.

Meijer, I., Berghmans, S., Cousijn, H., Tatum, C., Deakin, G., Plume, A., Rushforth, A., Mulligan, A., de Rijcke, S., Tobin, S., Van Leeuwen, T., and Waltman, L. 2017. *Open Data: The Researcher Perspective*. (https://doi.org/10.17632/bwrnfb4bvh.1).

Melissa Lee, Esteve Almirall, Jonathan Wareham “Open Data and Civic Apps: First-Generation Failures, Second-Generation Improvements” *Communications of the ACM*, January 2016, Vol. 59 No. 1, Pages 82-89

Monge, P. R., Fulk, J., Kalman, M. E., Flanagin, A. J., Parnassa, C., and Rumsey, S. 1998. “Production of Collective Action in Alliance-Based Interorganizational Communication and Information Systems,” *Organization Science* (9:3), pp. 411–433. (https://doi.org/10.1287/orsc.9.3.411).

Mørk, B. E., Aanestad, M., Hanseth, O., and Grisot, M. 2008. “Conflicting Epistemic Cultures and Obstacles for Learning across Communities of Practice,” *Knowledge and Process Management* (15:1), pp. 12–23. (https://doi.org/10.1002/kpm.295).

Nagendra, H., and Ostrom, E. 2012. “Polycentric Governance of Multifunctional Forested Landscapes,” *International Journal of the Commons* (6:2), pp. 104–133.

National Science Board. 2005. “Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century.” (https://apps.dtic.mil/sti/citations/ADA444393).

NIH. 2003. "NIH Data Sharing Policy and Implementation Guidance." (https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm, accessed March 10, 2020).

OECD. 2015. "Making Open Science a Reality," No. 25, *OECD Science, Technology and Industry Policy Papers,* Paris: OECD Publishing. (https://wiki.lib.sun.ac.za/images/0/02/Open-science-oecd.pdf).

Olson, M. 2009. *The Logic of Collective Action*, (Vol. 124), Harvard University Press.

OpenAire, S. 2019. "RDM Costs," *OpenAIRE*. (https://www.openaire.eu/how-to-comply-to-h2020-mandates-rdm-costs, accessed March 10, 2020).

Ostrom, E. 1986. "An Agenda for the Study of Institutions," *Public Choice* (48:1), pp. 3–25.

Ostrom, E. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*, Cambridge University Press.

Parsons, M. A., Duerr, R., and Minster, J.-B. 2010. "Data Citation and Peer Review," *Eos, Transactions American Geophysical Union* (91:34), pp. 297–298. (https://doi.org/10.1029/2010EO340001).

Pasquetto, I. V., Randles, B. M., and Borgman, C. L. 2017. "On the Reuse of Scientific Data," *Data Science Journal* (16), p. 8. (https://doi.org/10.5334/dsj-2017-008).

Perkmann, M., and Schildt, H. 2015. "Open Data Partnerships between Firms and Universities: The Role of Boundary Organizations," *Research Policy* (44:5), pp. 1133–1143. (https://doi.org/10.1016/j.respol.2014.12.006).

Peters, I., Kraker, P., Lex, E., Gumpenberger, C., and Gorraiz, J. I. 2017. "Zenodo in the Spotlight of Traditional and New Metrics," *Frontiers in Research Metrics and Analytics* (2). (https://doi.org/10.3389/frma.2017.00013).

Piwowar, H. A., and Vision, T. J. 2013. "Data Reuse and the Open Data Citation Advantage," *PeerJ* (1), PeerJ Inc., p. e175. (https://doi.org/10.7717/peerj.175).

Piwowar, H. A., Vision, T. J., and Whitlock, M. C. 2011. "Data Archiving Is a Good Investment," *Nature* (473:7347), Nature Publishing Group, p. 285. (https://doi.org/10.1038/473285a).

Plantin, J.-C., Lagoze, C., Edwards, P. N., and Sandvig, C. 2018. "Infrastructure Studies Meet Platform Studies in the Age of Google and Facebook," *New Media & Society* (20:1), pp. 293–310. (https://doi.org/10.1177/1461444816661553).

Polanyi, Michael; Ziman, John; Fuller, Steve. The republic of science: its political and economic theory Minerva, I (1)(1962), 54-73. Minerva, 2000, vol. 38, no 1, p. 1-32.

Pujol Priego, L., and Wareham, J. 2018. *Open Targets: Open Science Monitor Case Study.*, European Commission. (http://publications.europa.eu/publication/manifestation_identifier/PUB_KI0518020ENN).

Pujol Priego, L., and Wareham, J. 2019. *REANA: Reproducible Research Data Analysis Platform : Open Science Monitor Case Study.*, European Commission. (http://publications.europa.eu/publication/manifestation_identifier/PUB_KI0219176ENN).

Raphael, M., Sheehan P. & Vora G. "A controlled trial for reproducibility." *Nature* 579, 190-192 (2020)

Royal Society. 2012. *Science as an Open Enterprise.*, Policy Unit, United Kingdom. (https://royalsociety.org/~/media/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf).

Shapiro, C., and Varian, H. R. 1999. "The Art of Standards Wars," *California Management Review* (41:2), pp. 8–32. (https://doi.org/10.2307/41165984).

Simko, T., Cranmer, K., Crusoe, M. R., Heinrich, L., Khodak, A., Kousidis, D., and Rodriguez, D. 2018. "Search for Computational Workflow Synergies in Reproducible Research Data Analyses in Particle Physics and Life Sciences," in *2018 IEEE 14th International Conference on E-Science (e-Science)*, Amsterdam: IEEE, October, pp. 403–404. (https://doi.org/10.1109/eScience.2018.00123).

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., Pollock, D., and Dorsett, K. 2015. "Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide," *PLOS ONE* (10:8), (P. van den Besselaar, ed.), p. e0134826. (https://doi.org/10.1371/journal.pone.0134826).

Thelwall, M., and Kousha, K. 2016. "Figshare: A Universal Repository for Academic Resource Sharing?," *Online Information Review* (40:3), Emerald Group Publishing Limited, pp. 333–346. (https://doi.org/10.1108/OIR-06-2015-0190).

Uhlir, P. F. & Cohen, D. (2011, March 18). Internal document. Board on Research Data and Information, Policy and Global Affairs Division, National Academy of Sciences

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., and Zhao, S. 2019. "Applications of Machine Learning in Drug Discovery and Development," *Nature Reviews Drug Discovery* (18:6), pp. 463–477. (https://doi.org/10.1038/s41573-019-0024-5).

Vassilakopoulou, P., Skorve, E., and Aanestad, M. 2016. "A commons perspective on genetic data governance: the case of brca data," *Research Papers*. (https://aisel.aisnet.org/ecis2016_rp/136).

Venkatesh, V., Brown, S. A., and Bala, H. 2013. “Bridging the Qualitative-Quantitative Divide: Guidelines for Conducting Mixed Methods Research in Information Systems,” *MIS Quarterly* (37:1), Management Information Systems Research Center, University of Minnesota, pp. 21–54.

Vitali, M., Mathiassen, L., and Rai, A. 2018. “The Sustainability of Polycentric Information Commons,” *MIS Quarterly* (42:2), pp. 607–631. (https://doi.org/10.25300/MISQ/2018/14015).

Wallis, J. C., Rolando, E., and 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), (L. A. Nunes Amaral, ed.), p. e67332. (https://doi.org/10.1371/journal.pone.0067332).

Weill, P., and Ross, J. W. 2004. “It Governance on One Page,” SSRN Scholarly Paper No. ID 664612, SSRN Scholarly Paper, Rochester, NY: Social Science Research Network, November 1. (https://papers.ssrn.com/abstract=664612).

Weinberg, A.M., 1961. Impact of large-scale science in the United States. Science (80). https://doi.org/10.1126/science.134.3473.161

White, H. C., Carrier, S., Thompson, A., Greenberg, J., and Scherle, R. 2008. “The Dryad Data Repository: A Singapore Framework Metadata Architecture in a DSpace Environment,” in *Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications*, DCMI ’08, Berlin, Germany: Dublin Core Metadata Initiative, September 22, pp. 157–162.

Whyte, A., and Pryor, G. 2011. “Open Science in Practice: Researcher Perspectives and Participation,” *International Journal of Digital Curation* (6:1), University of Edinburgh, pp. 199–213. (https://doi.org/10.2218/ijdc.v6i1.182).

Wicherts, J. M., Bakker, M., and Molenaar, D. 2011. “Willingness to Share Research Data Is Related to the Strength of the Evidence and the Quality of Reporting of Statistical Results,” *PLoS ONE* (6:11). (https://doi.org/10.1371/journal.pone.0026828).

# 4

## 4. Opaque Spaces of the Commons: Governing Information Infrastructures in Life Sciences

The article that constitutes this chapter consists in a micro-study of a single case that gives us insight into the different *mechanisms* that help reconcile the main tensions between the two exogenous forces: open science and technology transfer. It empirically investigates Open targets an information infrastructures based upon *data commons* developed by EMBL-EBI and pharmaceutical companies to accelerate the drug discovery process.

## 4.1 Abstract

The sequencing of the human genome is recognized as a major landmark in biomedical research that has facilitated the emerging disciplines of genomics, proteomics, and systems biology. However, the capabilities and economic resources needed to leverage these vast data sources towards a greater understanding of disease mechanisms often exceed the scope of a single organization. In response to this challenge, biopharmaceutical companies have created commons-based information infrastructures. We present the exemplary case of *Open Targets* (OT), a large-scale information infrastructure created by leading organizations in bioinformatics, genomics, and pharmaceuticals that includes for-profit companies, non-profit foundations, and public research organizations. We describe and theorize about the governance conditions of modularity and brokerage that enable the processes of folding and unfolding into concealed or open spaces of work. This fluid dynamic simultaneously enables the benefits of shared investments and protects the private economic interests of its members. It offers a successful model for information infrastructure governance that navigates many of the trade-offs between private and collective interests in common resource pools composed of heterogeneous members with divergent objectives.

**Keywords:** Information infrastructure; collective action; governance; drug discovery

## 4.2 Introduction

The sequencing of the human genome (Human Genome Project, HGP) is recognized as "the largest undertaking in the history of biological science"(Chaguturu et al. 2014, p.35), which brought 1) a deluge of new biological data to be incorporated and assimilated in drug discovery processes; 2) new computational challenges of transforming DNA sequence information into disease-associated protein functions leading to the generation of digital technologies such as bioinformatics, metabolomics or genomics; and more broadly, 3) the discipline of systems biology, which promotes the understanding of how networks in the biological system interact (Au, 2014). Enabled by new computational technologies, the HGP opened the door to abundant data and, consequently, fundamental changes in how scientists understand diseases and biological mechanisms.

Sixteen years since its completion, it is acknowledged that the full potential of the HGP has not been realized for a number of reasons. One explanation is that the technological challenges of exploiting the HGP exceed the capabilities traditionally available to an individual company (Altshuler et al. 2010). In parallel, the cost of drug discovery has been growing (Lee 2015): the total R&D spent worldwide by pharmaceutical and biotechnology firms increased from USD 108 billion (2006) to USD 141 billion (2015) (Evaluate Pharma 2017), while the cost of developing a single medicine is estimated at 2.6 billion, more than double the figure of only a decade ago (Tufts Center for the Study of Drug Development 2015).

As a result, pharmaceutical companies have carefully opened up their boundaries in the early phases of drug discovery by creating pooled, commons-based information infrastructures shared across multiple, often competing, organizations. The primary focus of such information infrastructures has been to generate, integrate, and curate large data pools with commonly used tools and analytical methods for the research community (Grossman et al. 2016, Vamathevan et al. 2019). The main impetus is to avoid redundant investments in early-stage research efforts and to accelerate drug discovery.

Information infrastructures have been defined as "multi-layered entities comprised of technological components, people and institutional arrangements" (Constantanides 2012 p. 25) (Hanseth and Monteiro 1997). A more technology-focused definition is "a digital library system based on commonly shared standards and containing information of both local and/or widespread interest" (Kahn and Cerf 1988 pp. 3) … "to augment our ability to search for, correlate, analyze and synthesize available information," (Kahn and Cerf 1988 p.11.) Adding a social dimension, Constantanides (2012 p.21) defines information infrastructures as "efforts to integrate other computer-based and social systems, and to regulate and monitor processes that were previously performed in various, isolated settings." The logic behind these conceptualizations is now evident, as pharmaceutical companies have developed information infrastructures that integrate shared data, technologies, methods, and the rapidly increasing financial expense of integrating human genome information into drug discovery processes. The goal is to pool resources to achieve the needed depth and scale to validate potential therapeutic *targets* for various diseases in the first two of seven phases in drug discovery, typically occupying three of the twelve total years required on average. In drug discovery, the term "targets" typically refers to proteins that have three-dimensional structures to which specific molecules can bind to provoke some physiological effect.

The development of these information infrastructures has been based upon the concept of the commons, which refers to a set of resources that are collectively owned and shared among a community (Ostrom 1990). Commons contain public and private property over which different agents have certain rights. By creating such information infrastructures, organizations commit to revealing resources (data, methods, technologies) and forgo IP related to drug targets (Mishra et al. 2016). More concretely, companies agree to *postpone the time to patent* in the drug discovery process, accepting that targets will no longer be patent objects, and the data, technologies, and knowledge related to targets are to be shared openly in an information infrastructure.

Such *time dilation* at the moment of patenting offers a novel practice of sharing in the early phase of drug discovery that poses both benefits and challenges: while limited sharing can lead to redundant investments and delays in scientific discovery, excessive sharing can lead to unintended spillovers that damage the firm's competitive position. In other words, the development of the infrastructure (i.e., the conceptualization and implementation) must adequately negotiate a conundrum: it must be sufficiently open that organizations benefit from sharing but adequately restrictive at certain points to protect private economic interests in the competitive race towards therapeutic drug development. Ultimately, participants in these information infrastructures face a *collective action problem*: how do pharmaceutical companies develop an infrastructure that grants access to the data, methods, and technology, which we refer to as *collective resources,* critical to upstream drug development while preserving downstream commercial opportunities? Balancing these economic and scientific incentives remains a challenge in an IP-intensive industry, which has historically been characterized by high levels of secrecy, even between clinical teams within a company (Allarakhia 2014, Mittleman et al. 2013). To address this challenge, the development of the information infrastructure requires well-considered *governance* that enables the sharing of pooled resources, which we term openness, while offering the protections, or closure, needed to realize an economic return on firms' investments in later phases.

The term "collective action" refers to joint action by a number of agents to achieve and distribute some gain through coordination or cooperation (Hardin, 1982). Research into collective action problems was initiated with Mancur Olson's (1965) now classic *Logic of Collective Action* and later popularized with Gareth Hardin's (1968) thesis on the "tragedy of the commons." Collective action research has been adopted by information infrastructure scholars to understand the challenges of information infrastructure development that align individual and collective interests from the different organizations involved. Tightly coupled with this literature is the concept of *governance*, which refers to the rules that underlie the social activities that are integral to order relationships, responsibilities, and expectations of contributors (Ostrom 1990, Mindel et al. 2018, Weill and Ross 2004). The debate on commons governance distinguishes between bottom-up, decentralized or polycentric approaches and top-down or centralized structures in assigning rights, responsibilities, and privileges in various types of collective resource systems (Constantinides and Barrett 2015, Mindel et al. 2018). The historical discourse has been polarized in two rival schools of thought: one represented by Garret Hardin and the second by Elinor Ostrom. Hardin's 'tragedy of the commons' describes the cautionary tale of how collective resources will eventually decline if governed in a decentralized manner characterized by a high level of openness. By contrast, the 2009 Nobel Prize winner in Economics Elinor Ostrom documents in her book *Governing the Commons* (1990) how communities manage collective resources without top-down regulations.

When adopting the commons discourse to understand information infrastructures, the proponents of the top-down approach argue that infrastructure development needs to have clear governance defining who makes what decisions e.g., (Weill and Ross 2004). A contrasting scholarship argues that information infrastructures cannot be governed in a top-down, centrally controlled manner due to the complex dynamics required to cultivate a constantly growing base of users with diverse needs (Hanseth and Lyytinen 2010, Sahay and Aanestad 2009). This literature suggests that information infrastructures need to be governed through their design, suggesting a set of design principles and rules that acknowledge "pivotal relationships between technical and social elements, and their dynamic interactions" (Hanseth and Lyytinen 2010 p.15); that is, through careful design choices, information infrastructures can be self-organized.

In this paper, while we acknowledge the relevance of bottom-up governance approaches to information infrastructures, we argue that they also need to be effectively governed with certain restrictions to award the appropriate economic protections for companies to realize a return on their substantial investments.

Thus, the key research question that we seek to answer is as follows: *how do organizations develop commons-based information infrastructures that govern access to collective resources while simultaneously protecting the members' private interests?*

Following calls in the recent literature (Mindel et al. 2018) and requests to "cover the interplay with institutions, goods, and the social practice" (Von Krogh et al. 2012, p. 670), our objective is to advance beyond the usual rivalry between bottom-up and top-down governance and the binary distinction between public and private goods in information infrastructure research. We seek to understand and develop theory about alternative governance forms that allow dynamic navigation through the needs of sharing collective resources (openness) with the appropriate levers of restriction and confidentiality (closure) to enable the pursuit of competitive interests.

Among other successful results, OT researchers have effectively contributed to accelerating the development of targeted cancer treatments by discovering "thousands of genes essential for cancer's survival and ranked which ones show the most promise as drug targets for developing new treatments" by employing a novel computational framework that integrates "multiple lines of evidence to assign each gene a target priority score" (Behan et al. 2019 p. 511).

Combining insights from the extant literature and findings from our in-depth study, we develop a generalizable model of information infrastructure governance based on the commons that enables both the protection of competitive knowledge and a degree of openness to make the collaboration effective. Our study seeks to better understand the collective action challenge of granting participating organizations access to collective resources while reconciling both collective and self-interests. While our empirical context is an information infrastructure in the life sciences, we believe that our findings can inform information infrastructure development in a wider range of domains. Our theoretical development integrates ideas from three main areas: 1) information infrastructure governance scholarship, e.g., (Constantanides 2012, Constantinides and Barrett 2015, Hanseth 2001, Hanseth and Lyytinen 2010, Hanseth and Monteiro 1997); 2) collective action theory, e.g., (Hardin 1968, Hess and Ostrom 2003, Ostrom 1990); and 3) the literature describing the implications of the sequencing of the human genome and recent computational research methods to the collaborative dynamics of the pharmaceutical industry, e.g., (Allarakhia 2014, Au 2014, Choudhury et al. 2014, Collins 2003, Hood and Rowen 2013, Vamathevan et al. 2019).

The remainder of the paper is structured as follows. In the next section, we develop our theoretical underpinnings regarding information infrastructure development and collective action theory. This is followed by a discussion of the research context and methods employed for conducting the empirical study. We then draw on our theoretical approach in presenting the results of our analysis on OT. Finally, we develop contributions to collective action challenges and information infrastructure development and suggest avenues for future research.

## 4.3 A Paradigm Shift in Science and Technology for Drug Discovery

*Drug discovery* is defined as “the process of creating chemical or biological molecules that have the potential to be developed as therapeutic agents, typically because they generate a desired biological effect in an appropriate testing or assay system against a particular molecular (drug) target” (Weigelt, 2009, p. 941). The basic approach in drug discovery consists of developing drugs that will alter the disease state by modulating (i.e., as an agonist or antagonist) the activity of a molecular target (Vamathevan et al. 2019).

Despite some idiosyncrasies that may change from company to company, the standard *drug discovery and development process* can be divided into the following interdependent processes (Chaguturu et al. 2014): 1) target identification, which focuses on discovering the molecular targets (normally proteins) that play a fundamental role in disease; this phase is devoted to uncovering causal associations between a target and disease, which requires demonstrating that the modulation of a target has an effect on (i.e., modulates) a disease state; 2) target validation, which is the confirmation of the molecular target being associated with the disease identified by employing physiologically relevant ex vivo and in vivo models; 3) lead or hit identification, which consists of identifying multiple pharmacological molecules active against potential targets; 4) the lead optimization process consists of optimizing the "function of how tightly the molecules interact with the target in order to improve selectivity and the degree to which a dose of a drug produces the desired effect against a specific target"(Chaguturu et al. 2014, p.34) while evaluating the safety of leading molecules; 5) preclinical trials to identify the best candidate molecules to test; and 6) clinical trials (phases I, II, and III) test potency, metabolism, toxicity and other variables critical to regulatory approval.

In essence, once a target has been confirmed to be associated with a disease, companies need to identify for further development the pharmacological molecules that can affect the target. At this point, the lead leaves the discovery phase and enters preclinical development (Zanders 2011). The transition from the discovery phase (related to targets) to preclinical development (lead development) is a major one: scientists leave 'blue sky' research and enter into a heavily regulated process of developing and marketing a medication to be sold to a global market. The steps to be taken from the discovery phase onwards have a diverse set of legal and financial implications for the company at hand. While the work on the discovery phase, which involves in targets until the lead molecule is chosen to proceed to development, is experimental, the following steps devoted to development require the manufacture and formulation of the compounds, which comprise highly regulated procedures with very well-defined processes that need to be consistently performed. Fundamentally, such consecutive phases need to prove that the drug candidate can effectively bind and modulate the target in a safe manner. Towards this end, company teams need to examine the pharmacodynamics (i.e., the effect of the drug on the body), pharmacokinetics (i.e., the effect of the body on the drug), and safety pharmacology or toxicology, in other words, any undesirable effects (Zanders 2011).

Historically, academia conducted basic research in biology, deciphering new disease targets and relevant pathways with potential therapeutic value, while biopharmaceutical companies pursued closed research in search of therapeutic targets. Since the completion of the HGP, genomics technologies in drug discovery have shifted the problem from the "identification and creation of novel small-molecule drugs against known targets (chemistry) to the biological characterization and functional validation of large numbers of unknown drug targets (biology) at the molecular, cellular and system levels" (Hopkins et al. 2007, p. 371). The effort consists of decoding the disease-associated mechanisms that are generated by single or multiple genes and understanding their interaction with environmental factors.

While the genome is a significant determinant of how diseases originate and evolve, environmental factors often play an essential role, and in many cases, these two factors are intertwined (i.e., a particular genotype may change the risk of an environmentally induced disease) (Katsila et al. 2016).

As a result, firms have shifted their focus towards the *early stages* of the drug discovery process, reducing investments in later stages to improve the success rate of drugs entering the development pipeline. The significant economic cost of this novel research form requires collaborative investments to generate large shared data resources. In essence, we have witnessed a paradigm shift since the completion of human genome sequencing and the consequent technological advances, the features of which we summarize in Table 1.

*Table 1. A Paradigm Shift in Drug Discovery (inspired by Au, 2014)*

| | **Before** | **After** | **Source** |
|---|---|---|---|
| **Time (average)** | 8 – 10 years | 10 - 15 years | (e.g., Au 2014; Lee, 2015; Schuhmacher et al. 2016) |
| **Cost (average)** | $800 million | $1.65 billion | (e.g., Tufts Center for the Study of Drug Development, 2015) |
| **Intellectual property** | Patent protection | Shorter patent protection and numerous drugs going off a patent-patent cliff | (e.g., (Markus et al. 2006) Lee, 2015; Lesser and Hefner, 2017) |
| **Paradigm Shift** | | | |
| **Drug discovery Methodology** | "Trial and Error" – start with 1000 compounds and narrow it down | More targeted pathways that demand a better understanding of biology | (e.g., Lesser and Hefner 2017) |
| **Focus** | Focus on later stages of drug development | Increased focus on the basic science of drug discovery | (e.g., Lesser and Hefner, 2017) |
| **Technologies implemented for drug discovery** | The standard strategy was to internalize public-domain data and to build (or license) internal platforms to manage and integrate them with internal data | Innovative computational biology approaches require the development of new technologies that extract value from increasingly comprehensive public-domain data sources. A shift from 'proprietary data' to 'proprietary understanding of data' | (e.g., Au 2014; Barnes et al. 2009; Loging et al. 2007; Schrattenholz and Soskic 2008) |
| **Publication of clinical trial data** | Restricted access due to IP protection reasons. Patient data confidential unless subpoenaed by a court order | Companies opening up clinical trial data for research to increase a better understanding of disease progression | (e.g., Au 2014; Pogorelc 2014) |
| **The scope of industrial-academic collaborations** | Focus on specific targets. Partnership agreements are typically small in scope | The broader focus of the collaboration expanding across one or more indications, therapeutic areas, or operational capabilities | (e.g., Bianchi et al., 2011; Chaguturu et al. 2014; Hunter and Stephens, 2010; Salah and McCulloch, 2011 |
| **Type of industrial-academic collaboration** | Typically involving two parties and using a structure (a "sponsor" and "partner" model) that distributes control, risks, and rewards | Typically involving three or more parties including biopharmaceutical companies, academia, non-profit contributing resources. Shared control and decision-making, thus increasing potential risks and rewards | (e.g., Lesser and Hefner, 2017) |

## 4.4 Theoretical underpinnings

### *4.4.1 A Collective Action Approach to Information Infrastructures*

Information infrastructures such as data and code repositories for scientific or health information have long been differentiated from transactional information systems (e.g., ERP) for being integrators of widely distributed and previously siloed information spaces (Constantanides 2012, Constantinides and Barrett 2015, Hanseth and Monteiro 1997). The key attributes of an information infrastructure described by such a body of literature are described in Table 2.

*Table 2. Summary of Key Aspects of an Information Infrastructure (II)*

| **II attributes** | **Definition** | **Sources** |
|---|---|---|
| *Shared and open* | An information infrastructure is shared by and open to a large user base and technological components. | e.g., (Byrd and Turner 2001, Constantanides 2012, Hanseth 2001) |
| *Reusable and modular* | An information infrastructure is modular in that it has the ability to add, modify and remove technological components with little effect on its features and process of other components. The modular attribute of an infrastructure leads to the subprinciples of decomposition, recombination and reusability of its components. | e.g. (Byrd and Turner 2001, Chung et al. 2003, Duncan 1995) |
| *Built on installed base* | An information infrastructure is not developed from scratch but on the existing installed base, which constantly evolves in different layers. | e.g., (Grisot et al. 2014, Hanseth 2001, Weill and Broadbent 1998) |
| *Enabling* | An information infrastructure has a supporting or enabling function. It is not designed to automate something that already exists or to support one way of working or a specific application but to support the emergence of new activities. | e.g., (Hanseth 2001) |
| *Embodied in standards* | An information infrastructure and its components are embodied in different standards (e.g., coding schemes, terminologies) that need to be agreed upon to facilitate the interoperability and | e.g., (Ciborra and Andreu 2001, Hanseth et al. 2006, Hanseth and Monteiro 1997) |

| | | |
|---|---|---|
| | connection between the different components and their further reusability. “Standards are a necessary constituting element” of the collection of information infrastructure connections (Hanseth 2001, pp.57). Standards allow a shared pattern of use among a diverse range of user organizations. | |
| *Heterogeneous* | An information infrastructure is a heterogeneous collage of people, systems and processes. | e.g., (Ciborra and Andreu 2001, Hanseth et al. 2006, Hanseth and Monteiro 1997) |

The process of integrating information spaces is characterized by being complex due to a number of challenges that it needs to face (Constantinides and Barrett 2015). Challenges include combining heterogeneous interests and resources from the different organizations and (Hanseth and Lyytinen 2010, Hanseth and Monteiro 1997) agreeing on a set of standards (Bowker 1999, Hanseth 2001, Hanseth and Monteiro 1997, Star and Ruhleder 1996), which “are a necessary constituting element” (Hanseth 2001, p.57) “to the collection of information infrastructure connections” (Constantanides 2012 p.26). An additional challenge is defining an appropriate *governance* that facilitates the integration and sustainability of the information infrastructure (Constantinides and Barrett 2015). Decisions are made across a wide range of aspects, including infrastructure architecture and its components, procurement and operation, type of information, standards, access and user rights, applications, processes, and resource investment e.g., (Weill and Ross 2004).

Some researchers have argued that information infrastructures need to be governed in a centrally controlled manner to solve conflicts of interests, e.g., (Markus et al. 2006, Vincent and Camp 2004), while other scholars have argued that centralized control is insufficient to address the dynamics of a perpetually changing base of heterogeneous users (Hanseth and Lyytinen 2010, Henfridsson and Bygstad 2013, Sahay and Aanestad 2009, Yoo et al. 2012).

A *top-down* versus *bottom-up* approach for governing information infrastructures translates the centuries long debate among economists, sociologists, ecologists, and political scientists regarding how to govern different types of collective resource systems called *commons* (Mindel et al. 2018). Both streams of literature are preoccupied with how large-scale resource pools can be made openly accessible to a large population of users while maintaining an equilibrium between, often competing, private and public interests. The central dilemma is that the inappropriate governance of a commons may disincentivize agents to contribute to the common resource pool and jeopardize its overall sustainability (Rolland & Monteiro, 2002).

The historical discourse has been polarized into two rival schools of thought: one represented by Garret Hardin and the second by Elinor Ostrom. In 1968, Hardin popularized the 'tragedy of the commons,' which became a leading paradigm in political science as an argument explaining what will happen to openly accessible resources if strong top-down institutions do not set limits on individual freedoms (Hardin 1968), showing how uncontrolled individual self-interested pursuits may sabotage the common good (Greco and Floridi 2004). The tragedy of the commons is an instantiation of the prisoner's dilemma, specifically, an *n*-person prisoner's dilemma where the rational pursuit of each agent's individual self-interest leads to suboptimal management of common recourses (Greco and Floridi 2004, Fletcher and Zwick 2000, Ostrom 1986). According to Hardin's logic, decentralized online information systems would not succeed due to their high openness levels (Mindel et al. 2018). There are abundant studies providing empirical support for Hardin's argument, e.g., (Ma and Agarwal 2007, Moon and Sproull 2008, Ransbotham and Kane 2011, Stewart and Gosain 2006).

As an alternative, Ostrom (1990) argued that the logic behind the tragedy of the commons is simplistic and problematic. In the 1980s, Ostrom and her school of thought collected and analyzed more than 5,000 empirical field studies from around the world to scrutinize and identify the structural characteristics of open resource systems: the attributes and practices of their users and rules and the reported outcomes. This research identified many well-functioning open resource systems that work in the absence of strong, centralized authority (Nagendra and Ostrom 2012, Ostrom 1990). She observed that the 'most resilient governance arrangements were those that dynamically managed boundary setting and mutual accountability through a high degree of inclusivity in decision-making' (Mindel et al. 2018). The main idea was that by increasing the number of decision-makers (polycentricity), an individual's commitment to the open resource system is reinforced, mitigating the need for central governance. The concept of polycentricity was first developed by (Polanyi 1951) to describe the free independent exercise by scientists unconstrained by the intervention of a central management authority (Aligica and Tarko 2012). Ten years later, polycentricity was adapted by Ostrom (et al. 1961) as an alternative to centralization. The resulting framework became the foundation for common-pool resource governance research, also known as collective-action research.

The literature, to date, has paid little theoretical attention to whether centralized and polycentric governance of an information infrastructure are discrete, static alternatives or, rather, a fluid and manageable characteristic that could be dynamically governed through certain architectural characteristics of the infrastructure. Just how such movement between openness and closure is made possible in an information infrastructure (what this balancing involves, who manages it, how it happens and how it actually impacts the collective action problem while addressing both collective and private self-interests) is an under-researched area that we endeavor to explore and theorize about. Such an exploration requires an understanding of the type of collective resources under focus.

### 4.4.2 Commons Goods in Information Infrastructures

Samuelson (Samuelson 1954 pp. 387-389) classified goods as either private or public, placing great emphasis on *exclusion*. Goods for which the use by other individuals was excluded were labeled private, in contrast to goods for which all individuals were included (i.e., public goods). Another dimension was introduced into the schema by adding *subtractability* (also referred to as *rivalry*), where the use of a good by one person subtracts from the availability of the good to others. Across this two-dimensional classification of goods, research has been developed to identify the varying degree of exclusion and subtractability. Embracing these two dimensions (exclusion & subtractability), collective action research transcended the dual classification of public versus private goods (Monge et al. 1998), which led to the approach to goods as commons, which exhibit properties of both private and public goods (Ostrom 1990).

Despite their parallel characteristics, Hardin's and Ostrom's 'commons' (called common pool resources by Ostrom) have important differences with the goods involved in discussions of information infrastructures. Hardin and Ostrom theorize about physical resource systems with tangible natural or man-made resources. The application of collective action research to theorize about information infrastructures conceptualizes a '*good'* as the functionalities that the information system affords and the collective interests and resources of the users. Compared to the classical theorization around natural collective action goods (e.g., forests, fisheries, pastures), in information infrastructures, goods are "sociotechnically interdependent on the heterogeneity of interests and resources of a distributed user base" (Constantinides and Barrett 2015 pp.44, Markus et al. 2006). As a result, information infrastructure research maintains that due to the distributed and interdependent nature of such goods, 'the level of the good at any given time will depend on the average rate of collective resources contributed' (Monge et al. 1998 p.417)

While the physical nature of the resources in Ostrom's and Hardin's theories makes the resource unit subtractable, in information infrastructures, the resources are in a digitalized form, so one person's use of information does not directly imply subtraction from another person's ability to use it; resources do not face the social dilemma of overconsumption (Constantinides and Barrett 2015, Mindel et al. 2018). The characteristics of exclusion and subtractability are not "givens" as in natural resources, but "they can be fabricated and technologically contingent"(Vassilakopoulou et al. 2016 pp.4.) Thus, information infrastructures are always subject to negotiations regarding the extent to which they remain open and shared by a wide and growing base of users (Hanseth 2001, Star and Ruhleder 1996), how they regulate their use through IP regimes e.g., (Benkler 2006), standards (Monteiro 1998), and the governance structures to manage their use e.g., (Weill and Broadbent 1998, Weill and Ross 2004).

### *4.4.3 Folding and Unfolding Processes in Information Infrastructures*

In addition to information infrastructure scholarship and collective action theory, we now introduce two constructs to complete our conceptual foundations. In our theorizing, we borrow from Shaikh and Vaast (2016) the concepts of *folding* and *unfolding*. Shaikh and Vaast (Shaikh and Vaast 2016) employ these terms to describe how open-source developers create opaque spaces of work (the fold) and transition dynamically back to open spaces. Folding refers to the process by which developers create such private workspaces (the fold), and unfolding is the process by which the output of the fold (e.g., code, bug fixes) is released back into the open. Folding and unfolding processes balance the need in open-source software development for complete openness in the development process and sharing the source code with the need for moments to work in opacity. It is "a folding from what happens outside" (Shaikh and Vaast 2016 p.827) into a more restricted or hidden space that creates a territory for reflective organizing.

Originally inspired by Deleuze's and Kavanagh and Araujo's ideas (Deleuze and Strauss 1991, Kavanagh and Araujo 1995), the creation of the fold or hidden territories is temporary and return that which is inside back to the outside or a wider environment after a period of time. Folds are enabled by digital technologies and represent a virtual space for restricted exchanges and possess a fluid nature 'where change is the norm' (Shaikh and Vaast 2016 p.827). Unfolding is the natural occurrence after the fold when releasing the output (e.g., code) of the discussion.

We will adapt the concept of folding and unfolding in our theorizing to illustrate how information infrastructures dynamically manage pharmaceutical companies' need for openness and closure.

## 4.5 Research context and methods

We conducted a longitudinal, in-depth case study of OT with the goal of providing theoretical insights into how to govern data commons, allowing firms to reveal their data and resources while competing in later stages (Yin, 1984). Case-based exploratory methods are suitable for investigating poorly understood phenomena (Eisenhardt 1989).

OT offered a powerful opportunity for theory generation. Based on a set of preliminary interviews with managers at a variety of pharmaceutical companies (n=5) and life science laboratories (n=3), we selected OT following three main criteria: its *fit, distinctiveness*, and *revelatory* nature (Eisenhardt 1989; Siggelkow 2007; Yin 2003). First, OT is an information infrastructure that has achieved extraordinary success in organizing the disclosure of data, technology, and methods, leading to the identification of 2,540 potential new indications for 791 current drug targets. Second, the case is *distinctive* in the sense that the OT includes Europe´s leading pharma and life science organizations and flagship scientific research infrastructures. Finally, we view OT as highly *revelatory* of a successful example of information infrastructure governance that enables data and knowledge sharing (Altshuler et al., 2010) to accelerate target identification and validation while protecting organizations' assets in the later stages of the drug development process. In this section, we describe the research context and our data collection and analysis.

### *4.5.1 Research Context*

OT is constituted by a group of organizations: EBI-EMBL—Europe's flagship laboratory for life science—the Wellcome Sanger Institute, GSK, Biogen, Takeda, Celgene, and Sanofi. The organizations collaborate to generate target-centered data on human physiology and systems biology in pursuit of cutting-edge experimentation, which they openly share and integrate with publicly available data in the OT infrastructure.

The methods used by OT include a combination of large-scale genomic experiments with scientific statistical and computational techniques to identify and validate causality between targets, pathways, and diseases (Open Targets, 2018). OT employs advances in cutting-edge genetic methods to support researchers in the first step of exploring new drugs, concretely helping them to identify "where to start" (Open Targets, 2018). By applying lean user experience (UX) design methods, OT members developed an infrastructure that searches, assesses and integrates a vast quantity of pubic and proprietary genetic and biological data to support target-centric and disease-centric inquiries.

According to the last update available (November 2019), OT contains more than 27,069 targets, 6,336,307 associations, 13,579 diseases, and 20 data sources (Open Targets, 2018). OT collective work has resulted, among other highlighted results, in the identification of 2,540 potential new indications for 791 current drug targets. A total of 1,366 of these 2,540 indications are for Orphanet rare diseases where the target is a known drug target for a common disease (Khaldakar et al., 2017). OT has suggested potential drug-repositioning opportunities for 14 rare diseases, and according to Pharmaprojects[8], which gathers

[8] Pharmaprojects website: http://www.pharmaprojects.com

worldwide drug development pipeline data, 6% of all new target-disease pairs uncovered in OT are in drug development, which is a conservative estimate given that only indications with exact matches were considered (Khaldakar et al., 2017). Drug repositioning is a strategy in drug development that seeks to expand the indication space for a successful drug or find a new indication for a drug that was not successful in clinical trials. While the traditional approach to drug development takes from 10 to 15 years, the repositioning strategy takes an average of 6.5 years (Khaladkar et al. 2017).

### *4.5.2 Data Collection and Sources*

Our study relies on a diverse set of primary and secondary data to provide richness and enhance the validity of our findings (Alvesson 2009, Klein and Myers 1999). Primary data included 25 semi-structured interviews conducted in two phases from 2017 to 2019 and direct observations from one study visit at the Welcome Genome Campus for the OT open days (June 2019). Our objective was to interview a representative cross-section of OT, including academics, company members, the operational team and external users (non-members) of the OT infrastructure. The interview process was concluded when no significant additional insights were obtained from the data, and theoretical saturation was achieved.

Secondary sources were also an essential data source. As OT activity has been widely publicized in media outlets, it was possible to collect substantial and relevant information from published sources. We combined these data with the publications resulting from OT (i.e., 41 research publications), together with tutorials about how to use OT infrastructure, blog posts, release notes, webinars, workshop presentations, the question and answer (Q&A) section of the OT website, and OT information contained in the seven partners' websites and annual reports. These secondary sources appear to be very useful, as they allowed us to perform crosschecks using multiple sources. The combination of our primary data with secondary data analysis allowed us to build our theoretical inferences from the case.

Finally, we supplemented our data with peer-reviewed publications on the topic of target identification and validation in the drug discovery process. We drew upon these sources to better grasp the technical work involved in OT to identify the target-disease associations and understand the type of data and methods used to find the best evidence of an effective and safe target. These academic sources served the mutually relevant but separate purpose of acquainting us much more deeply with relevant bioinformatics and biomedical backgrounds. This gave us contextual knowledge to make better sense of our primary data, both the interviews and observations (Lok and de Rond 2012). Table 4 in Appendixes provides a detailed description of the data collection, sources and their use in the analysis.

### *4.5.3 Data Analysis*

We performed a two-stage inductive analysis. The first stage, conducted between September 2017 and November 2018, was exploratory. We obtained primary data from 10 in-depth semi-structured interviews and informal conversations with research scientists and engineers from academic organizations and the managers of companies participating in information infrastructures in the life science sector. This phase was also devoted to reading abundant material available online about emerging information infrastructures after the HGP in the life science sector.

After this first exploratory phase, in a second stage, we completed 15 interviews in a second round with a cross-section of representative organizations participating in OT. The major themes in our interview protocol are summarized in Appendix B. Interviews were, on average, 45–60 min long, and the questions focused on OT governance, technical characteristics of the infrastructure, the role of the OT operational team, and the competitive and cooperative dynamics sustaining the infrastructure. Interviews were anonymized, and we organized and analyzed data for salient themes. We compared transcripts to identify themes in initial interviews to then explore and contrast these themes in subsequent interviews. Themes were coded by one of he co-authors and they were iteratively discussed with the other co-author, especially when the categorization was unclear to reach an agreement. We performed the interviews and their analysis in several iterations, and thus those earlier transcripts informed and incorporated information emerging from later interviews. In our results (section 5), we present interview excerpts from the study, with alphanumeric key identifiers (corresponding to table 3) representing quoted interviewees. Table 4 in appendix provides a detailed description of the progression of our empirical analysis towards the theoretical constructs.

*Table 3. Details on Data Collection and Use in the Analysis*

| **Source of data** | **Type of data** | **Description** | **Identifiers** | **Use in the analysis** |
|---|---|---|---|---|
| *Interviews* | *First Round*<br>n=10 | Research scientists and engineers from academic organizations in information commons in the pharmaceutical sector (n=3) | R1<br>R2<br>R3 | To gather data and an overall understanding of the logic behind developing common information infrastructures |
| | | Managers of companies participating in information infrastructures in the pharmaceutical sector (n=6) | M1 M2<br>M3 M4<br>M5 M6 | |
| | | Operational team members involved in information infrastructures | O1 | |
| | *Second Round*<br>n=15 | Research scientists and engineers from academic organizations partnered with OT (n=4) | ROT1<br>ROT2<br>ROT3<br>ROT4 | To gather data on technical attributes of OT infrastructure, governance and organizational processes. |
| | | Managers of companies in OT (n=3) | MT1<br>MT2 MT3 | |
| | | Managers of companies not participating in OT but using the platform and their outcomes (n=3) | NOT1<br>NOT2<br>NOT3 | |
| | | Managers of OT – operational team (n=5) | OT1 OT2<br>OT3 OT4 OT5 | |
| *Observations* | *Visit to OT* | Observation to OT Open Days – workshop, working groups and social event (June 2019) | | To gain additional understandings of how the OT operational team facilitates the work of OT partners |
| *Secondary data* | Publications | 41 publications | | To gather data and obtain an overall understanding of all OT infrastructure, components, usages, and governance and major outcomes of the collaboration. |
| | Tutorials | 1 tutorial on OT infrastructure | | |
| | Blog | 3 outreach posts | | |
| | Release notes | 19 release notes | | |
| | Q&A | 6 posts | | |
| | OT partner websites | 7 websites | | |

## 4.6 Results

### *4.6.1 Open Targets Information Infrastructure*

OT was created in 2015 by a nucleus of academic institutions and pharmaceutical companies that sought to mitigate the attrition rates of firms' pipelines and increase the probability of a successful drug going through the process. The organizations decided to achieve this via the development of an information infrastructure to 1) *integrate* comprehensive datasets from myriad public databases, such as UniProt, ChEMBL, NHGRI-EBI GWAS, EuropePMC, and Cancer Gene Census, and share computational techniques to calculate, rank and score gene-disease associations and 2) *generate* new data, methods and tools via joint experimental research projects, the results of which would later be combined with large public datasets to support data-driven target prioritization.

The integration of the datasets followed a federated approach to develop summary information about the data, which takes the form of evidence objects supplied by the source database or by the OT team from parsing other databases. The idea was not to store all relevant data because the databases are already uniquely tailored to many of the specialized data sources and often evolve quickly. The infrastructure interface works as an open access "Google" – a type of search engine that extensively searches, assesses and integrates the vast quantity of genetic and biological data available – supporting two main paths: target-centric and disease-centric inquiries. An OT user can search for a target and is presented with visualizations of the evidence for associations with specific diseases clustered into broad therapeutic areas, allowing in-depth investigations of the evidence and user-defined lists of associations. In the second path, the user enters the name of a disease and asks which targets can be associated with this disease. The output is a visualized summary of the genetic targets associated with that disease and the underlying evidence available (Koscielny et al. 2017). OT also integrates third-party visualizations, which include visualizations of biological pathways developed by Reactome, a graphical display of RNA baseline expression developed by Expression Atlas, a visualization of the different protein features developed by UniProt or a three-dimensional protein structure display for targets[9]. In Appendix we offer an example of a search result in the OT infrastructure.

The analysis of our data reveals a governance form in OT that afforded two processes: *folding* and *unfolding*, which allowed organizations to dynamically navigate from open towards opaque and closed workspaces to protect companies' economic interests. The conditions creating and dissipating the fold are as follows: 1) a modular infrastructure containing different layers, access rights, and data standards supporting the systems'

[9] WebGL-based viewer for proteins and other macromolecular structures: http://dx.doi.org/10.5281/zenodo.20980

interoperability (so-called 'technical attributes' of the information infrastructure) and 2) a brokerage exercise by a trusted third party, the OT operational team, which behaves as an independent entity and as a liaison to the different companies to manage the boundaries between open, opaque and closed spaces of work (i.e., the 'organizational attributes' of the information infrastructure). We turn first to the description of the technical and organizational attributes (the conditions) to allow an appropriate description of the processes of folding and unfolding.

### *4.6.2 The Technical Attributes of Open Targets: A Modular Infrastructure with Multiple Layers*

OT designed different layers in the information infrastructures: a *public* layer, where any user can add data and tools through a federated approach coordinated by an OT operational team; a *consortium* layer, where only a selected number of partners can join through a negotiation process; and finally, each partner can privately include their proprietary data by integrating OT with their internal information systems (the *private* layer).

1) First, regarding the **public layer**, the information shared is accessible to any organization but belongs to the organization that contributed it to the public OT domain. The results are openly shared in the OT infrastructure and are aggregated and temporally delayed formats.
"*We are not keeping the data for ourselves. We generate the data on the Wellcome Trust campus, and the data is made available. It takes a little bit of time, it has to be in the right format, but all data is available, and we need to write a publication before disclosing the data, which takes time.*" MT2 claims, "*The platform has an average of 1,100 visitors per week.*" As OT4 further explains, "*We contribute not only with the data but also with the processes, documentation and the code that runs the platform.*"

2) Regarding the **consortium layer**, OT members have access to the data that generate through collaborative experimental projects, together with all contextual information relevant for extracting insights from large-scale experimentation. The information is accessible to any organization in the consortium and belongs to the organization contributing to the project and generating the data. As MOT1 describes,
"*Accessing the raw data is not that easy. It is not easy to interpret. The data is hard to deal with, and we need to be part of the consortium to be part of the experimental projects that generate the data, to have access to the metadata, and be close to the academic partners in the consortium to exploit the data.*" As OT4 describes, *"Everybody that joins OT understands the premise for being here. There is limited access without a doubt. Everybody*

*understands that until there is a formal publication after the project there is no disclosure." OT4*

The process of becoming a member of OT is highly regulated to safeguard productive and sustained collaboration: a) all members have to reach consensus on accepting the candidate organization; b) only one organization per year is allowed to enter the consortium; c) significant in-cash and in-kind investment is required (a commitment that the organization's teams will devote resources to consortium activity); and d) the decision entails a rigorous evaluation, spanning several months of negotiation, to determine whether the candidate organization is committed to sharing its resources and capabilities for a sustained period of time and to ensure strong alignment between the goals of the candidate and the consortium. As OT1 describes,

"*To become a member, you need to share the vision of the consortium [open disclosure] and agree on investing around two-digit million euros, and all partners have to accept your membership*." As MT2 explains, "*We have to be careful; this is why we have only one company join every year. It takes time to integrate a company into the consortium*."

The process of incorporating a new member can be proactive (i.e., reaching out to organizations that OT members want to bring into the consortium) or reactive (i.e., addressing the requests of organizations asking to join the consortium):

"*There are companies that contact us and express their interest in being part of the consortium, and there are other companies that we actively reach out to*" (MT1).

Agreeing with the OT 'philosophy' of sharing resources across partners and the results with the broader research community is not natural for pharmaceutical companies. As MOT2 explains,

"*Our head of R&D thought that it was better to put the investment in working with others on targets than trying to be the only ones knowing about the targets* (…) *An important issue is that for this step of target identification, genetics and genomics data is extremely relevant, which is a seed that is changing extremely rapidly and is a seed where most of the data advances are provided by academic advances. The way to exploit this information and our expertise was working with leading academic centers and joining forces with other companies."* As he further explains, "*Our head of R&D came from academia and decided that in the specific activities of early discovery and early biology, there is so much going on that he realized we would not be able to compete. Typically, to compete requires considerable investment, and we did not know if such investment was going to pay off. It is a long time and a very risky investment."*

3) **The private layer** to individual company members: in-house, firms can integrate OT with their proprietary data on compound libraries and preclinical and clinical trial data to further

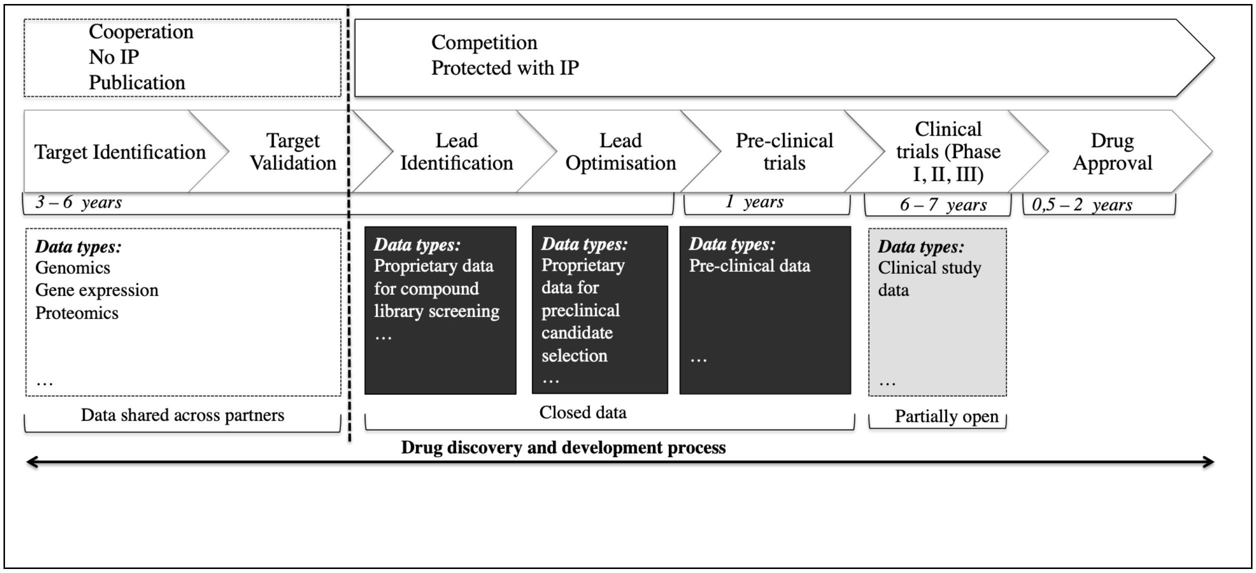
develop a potential drug for the identified target. Only company members can access the information, which is exclusive to the organization (i.e., a private good).

"*We have other features that are private, that we do not share with others. Those hidden features allow me, for instance, to work with my compound library on the platform, which I do not share with other OT partners*" (MT2). "*We also can ask for new features in the platform. We can also implement the platform in-house, within the company, integrate it with our systems, and you receive support from the other members of the consortium" (OT1).*

In addition to the demarcation of multiple layers based on *access rights or visibility*, the modularity of the infrastructure emphasizes a decomposition of loosely coupled knowledge domains in the drug discovery process (Henfridsson et al. 2014). This decomposition is based on 1) ***temporal latency*** and 2) ***domain separation of data and knowledge***.

First, ***temporal latencies*** refer to the periods between the generation of a particular datum and its mandated release to the public commons (Contreras 2010). Second, regarding ***knowledge domain separation***, e.g., (Sakakibara 2002), firms agreed in knowledge domains across the drug discovery process where they would reveal their data, methods, and technologies while separating them from domains where they would compete. Generally, firms agreed where to draw the line between what is considered precompetitive and competitive knowledge (figure 1).

*Figure 1. Descriptive Visualization of Data and Knowledge Shared within Drug Discovery*



In practical terms, firms decomposed the type of data and knowledge required in each phase of drug discovery into different components. They agreed to disclose data related to phases I and II of the drug discovery process (target identification and validation) while not revealing the knowledge and proprietary data that they use to identify the multiple molecules active against the potential targets or other information useful in later stages of the drug development process (i.e., phase 3 onwards) (figure 1). Dr. Rolf Apweiler, former director of OT, explains this as follows: "*the identification of a promising new target is precompetitive and should be shared, with the subsequent steps moving into the competitive arena*" (figure 1). This distinction is fluid: "*The definitions of precompetitive data may change over time, and boundaries of the intellectual property are becoming increasingly fluid*." (OT1)

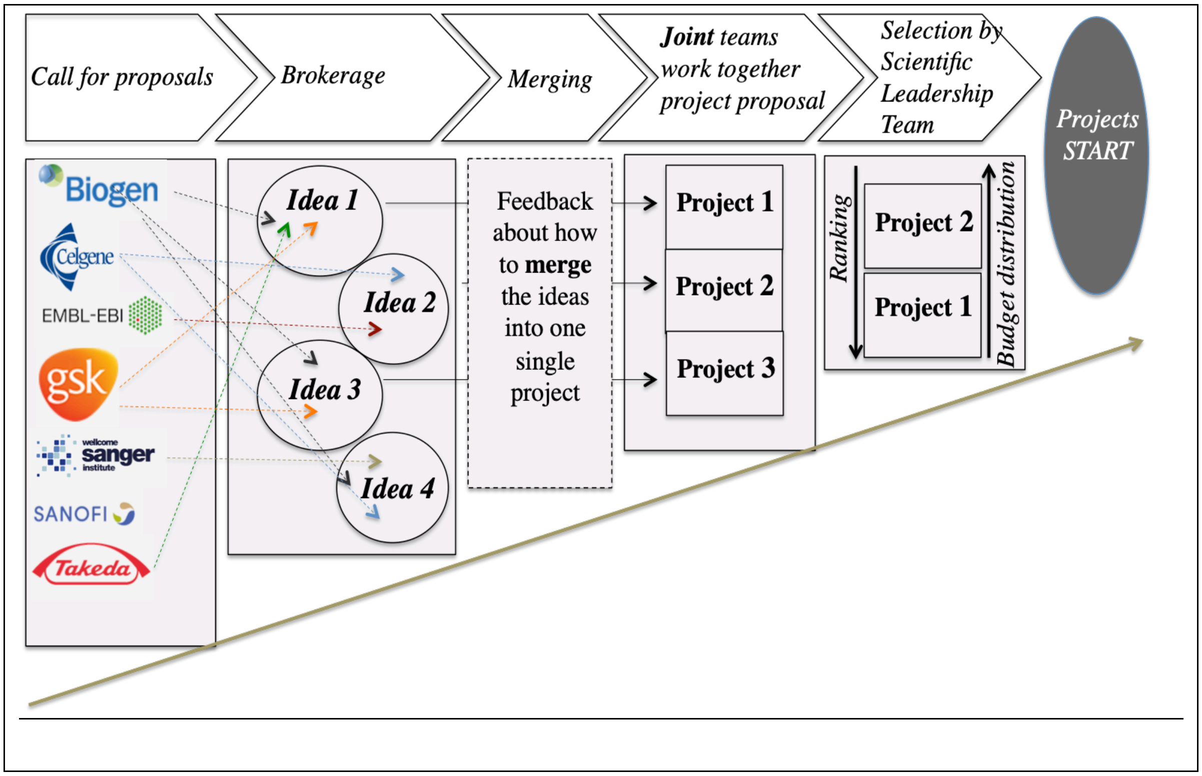
### 5.3. The Organizational Attributes in Open Targets: A Brokerage Exercise

Integrating information spaces from the different organizations in OT required a *brokerage* exercise by a trusted third party. The organizations involved in OT agree on the selection of a team – the OT executive and operational team (henceforth, the operational team) – that includes individuals with highly sophisticated scientific and managerial expertise who operate as a separate entity. As OT3 describes, "*we compare ourselves to a startup.*"

The OT team acts as a 'broker' across the different organizations, enabling the pooling of dispersed information across the organizations and supporting transitions among public, shared, and private workspaces. Non-disclosure agreements and complementary legal boundaries are implemented bilaterally between the OT team and each organization. The operational team also helps the different organizations agree on a set of standards and data protocols that enable not only interoperability between different layers but also boundaries between layers (public, consortium and private). As OT4 describes, "*Any access to data from the experimental project is provisioned through a person on the OT operational team. We have a gatekeeper for that, so that any person from our team asking for data needs to go through this person.*"

The operational team also helps match individual research initiatives across companies in pursuit of a joint research agenda. Organizations agree on a set of projects that will generate experimental data, which will initially be owned by the organizations generating it but will be made available in the public layer of the OT infrastructure after two years. An example of a brokerage activity implemented by the OT operational team is described by *MT1*: "*The process starts with a call for proposals that we distribute across all the companies, and the academic partners do the same as well, and we ask for an expression of interest, which consists of a one-page idea. We receive various ideas from our colleagues worldwide, we look at them [referring to the OT management team], and we see if the idea makes sense in what it tries to achieve. Then we do some matchmaking between companies, putting together similar ideas and proposals coming from X, Y and Z [a reference to specific OT companies], and we build teams combining the three firms. We merge the ideas that we agreed made sense into single projects. Based on the feedback that we provide, those teams come together after the matchmaking and write a full project proposal, which is around five pages. The projects are then reviewed by the Scientific Leadership team composed of representatives of* *each partner. Based on their ranking, we give a cut-off according to the budget available. Then, a certain number of projects go ahead and start. Typically, projects last 2-3 years"* (see figure 2).

*Figure 2. Shaping the Research Agenda, Match-making and Merging*



### 4.6.3 *Folding and Unfolding Processes: Navigating Dynamically among the Private, Shared and Public Layers*

A detailed analysis of our data gave us insights into what happens at the boundaries of each layer (public, consortium, private) and how the different organizations move across layers in the modular infrastructure. We characterize this process of fluidly transitioning through layers as *folding* and *unfolding,* employing Shaikh and Vaast's terminology (Shaikh and Vaast 2016). We use the term 'folding' to refer to the process by which organizations in OT create 'private pockets of interactions' inside their organizations (i.e., the movement of going private) and limit the openness of the information infrastructure. We employ the term 'unfolding' to describe the process by which organizations in OT release private information from private interactions into the shared or public layer of the infrastructure (figure 3). In other words, we consider the 'open' characteristics traditionally discussed in the information infrastructure literature (Constantanides 2012, Hanseth 2001) as a fluid and manageable attribute of an information infrastructure.

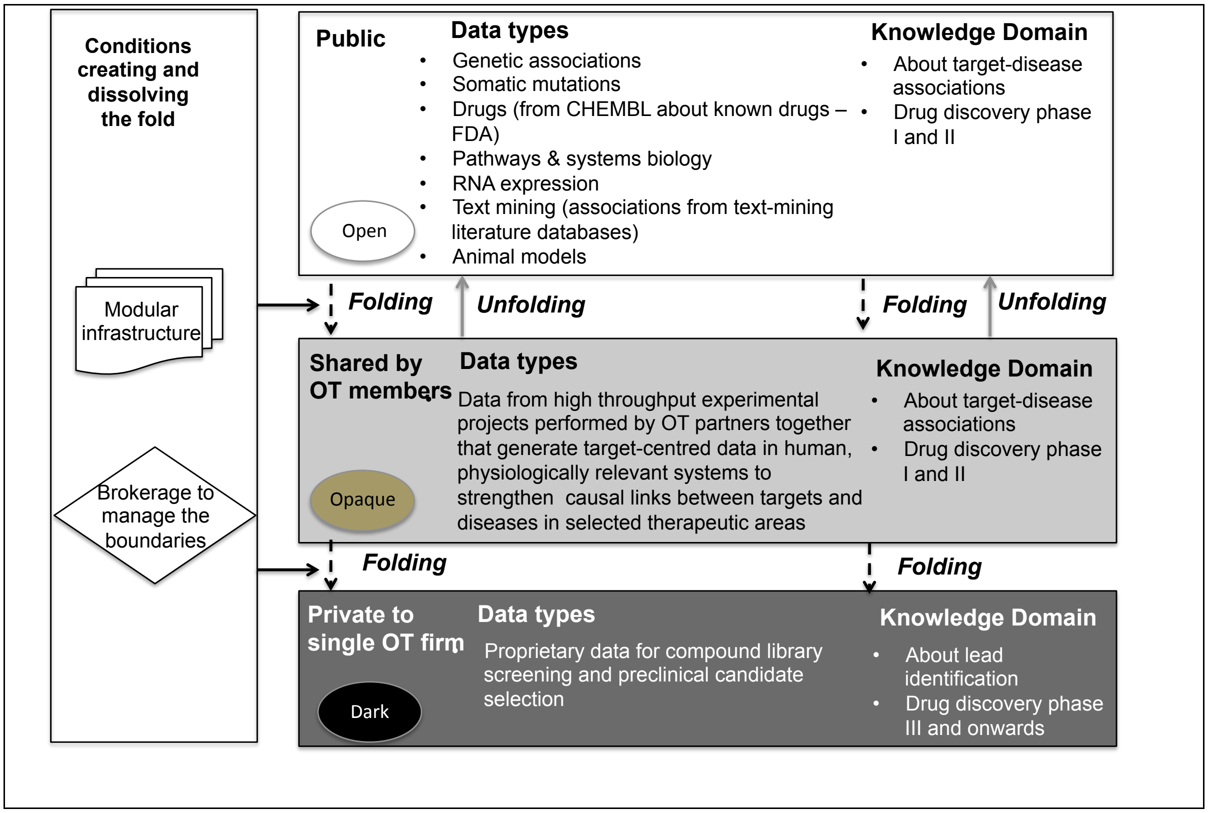
1) *Folding* refers to the process by which OT organizations create private workspaces: either a) including the OT consortium (i.e., opaque), or b) totally *closed* within a single organization (i.e., dark). Folding towards the consortium occurs when organizations seek to generate experimental data and robust evidence about targets and disease associations in collaborative projects. The outcomes are released after an average of two years into the public domain. In the second case of folding towards a single organization, companies periodically return to their drug development departments to exploit target-disease associations for their own drug discovery pipeline, that is, identifying potential molecules in their private compound libraries that may more satisfactorily modulate such targets.

2) *Unfolding* refers to the process by which OT organizations bring the output of the fold (e.g., experimental data, method, or technological tool generated in an OT project) back into the open.

These two processes allow information infrastructures to balance the principle of openness with the organization's needs for opacity and closure following competitive and market logic. The process of folding towards the consortium layer gives organizations "*access to data since minute one, while the public will only have it after a publication and appropriate curation, which takes on average of two years, and this makes a huge difference,*" explains OT1. "*This gives us an advantage compared to others outside OT*," MT1 confirms. Additionally, MT1 cites the reason for folding into the shared consortium space: "*Sharing scientific data in the public domain is a necessary but insufficient requirement for being able to reuse such data for drug development. You need, and this is what you pay for, close interaction with the scientists who generate data to understand what the data says, how it was generated, and how to interpret it.*"

Folding towards a completely closed space inside the company occurs when organizations need to integrate proprietary data and knowledge to reuse the information obtained from targets in the shared space in its pipeline. As OT1 explains, "*We know about other companies that are using the platform, and it is beneficial to them. Absolutely, that is happening. What they cannot do if they are not part of the consortium is to direct, integrate proprietary domains or data going to the platform, prioritize regarding functionalities of the platform, or determine the types of functions they want in the platform*" (see figure 3).

*Figure 3. Folding and Unfolding Processes*


Conditions creating and dissolving the fold
Modular infrastructure
Brokerage to manage the boundaries
Public
Data types
• Genetic associations
• Somatic mutations
• Drugs (from CHEMBL about known drugs – FDA)
• Pathways & systems biology
• RNA expression
• Text mining (associations from text-mining literature databases)
• Animal models
Knowledge Domain
• About target-disease associations
• Drug discovery phase I and II
Open
Folding
Unfolding
Folding
Unfolding
Shared by OT members
Data types
Data from high throughput experimental projects performed by OT partners together that generate target-centred data in human, physiologically relevant systems to strengthen causal links between targets and diseases in selected therapeutic areas
Knowledge Domain
• About target-disease associations
• Drug discovery phase I and II
Opaque
Folding
Folding
Private to single OT firm
Data types
Proprietary data for compound library screening and preclinical candidate selection
Knowledge Domain
• About lead identification
• Drug discovery phase III and onwards
Dark

## 4.7 Discussion

It has been suggested that the massive increases in genomic and proteomic data have had a profound impact on the structure of the pharma and life sciences industries (Chaguturu et al. 2014, Zanders 2011). Specifically, the clear demarcations that once existed between a) drug discovery and drug development and b) precompetitive and competitive domains are now becoming more fluid. This is a consequence of the rapidly increasing costs of drug development and the high costs of curating very large bioinformatic datasets and novel computational methods; the economic costs of developing and operating these information infrastructures are simply too great to be borne by any single organization. Our case study of OT is an outstanding exemplar of a life sciences information infrastructure constituted by some of the world's leading research, non-profit, and for-profit organizations. Despite their different objectives, they have developed a successful governance form to balance the tensions of sharing and collaboration inherent in a business context characterized by secrecy and vigilant patent protection.

Given that OT exhibits both private and public good properties (Ostrom 1990), our analysis has identified the collective action challenges of aligning private and collective interests and an appropriate governance form for this information infrastructure. We theorize towards a governance approach that affords two processes, i.e., folding and unfolding, that allows dynamic navigation towards open, opaque and dark workspaces that protect the members' economic interests.

The governance form provides a continuum between openness and closure, which are enabled by two elements. The first element is the technical attributes of the information infrastructure, which is modular with private, consortium, and public layers. Modularity not only relates to access rights or visibility based on the organizational level but also employs temporal levers (i.e., delays) to affect this. In combination, these attributes enable the decomposition and redefinition of drug discovery domains that have historically been highly segregated (i.e., target identification, target validation, lead identification, lead optimization.) As a consequence, the clear boundary between what had previously been considered precompetitive and competitive is now more fluid, manifesting a structural change in the industry. The second element that enables the folding and unfolding process is the function of a broker. The broker acts as a trusted third party, enabling members to effectively transition through the various levels of disclosure and collaboration based upon need. The broker first negotiates the entry of new members into the OT shared layer (the consortium) to determine whether the candidate organization is committed to sharing its resources for a sustained period of time and whether its private interests are aligned with OT's collective goals. Later, through bilateral agreements with each organization the broker facilitates the integration of the OT's shared work layers into the companies' private workspaces to facilitate competitive drug development. Additional tasks under the broker's role include the pooling of disperse resources into the information infrastructure, mediation to agree on data standards and protocols, and support for the integration and compatibility of data across the infrastructure.

### *4.7.1 Implications for theory*

Our theoretical developments integrate ideas from information infrastructure scholarship, collective action theory, and the conceptual foundations of the folding and unfolding constructs imported from Shaikh and Vaast (Shaikh and Vaast 2016). These constructs extend previous work on secluded workspaces and related concepts such as the 'structural folds' of Vaan et al. (2015) and Vedres and Stark (2010) and the 'relational spaces' of Kellogg (2009). Our research differs in that our folds do not theorize about individual behavior. Rather, we take the organization as the focal entity and theorize about the process of folding and unfolding under the governance of an information infrastructure. The temporal aspect of the previous work of (Shaikh and Vaast 2016) also differs in terms of context: the research and development processes in drug discovery are longer than those in open-source development, and as a result, the temporality of the opaque spaces is longer.

Applying this theoretical perspective to the empirical case, we sought to understand the dynamic governance of an information infrastructure that overcomes the challenge of simultaneously aligning individual and collective interests. This research contributes to the current scholarship in the information infrastructure literature by theorizing the open attribute of infrastructure as a manageable and fluid attribute that allows moving from private to open workspaces with common goods that can be reused by any potential organization. Openness and closure are both enabling and constraining attributes of an information infrastructure depending on the context at hand. Our findings describe a governance approach that makes the two compatible, describing the ongoing navigation between open and private workspaces that allows organizations to optimize their private and collective interests. Our research suggests a formula to overcome the historical social dilemmas of collective action (i.e., free-riding and overconsumption) in the sense that contributors have incentives to invest in the commons because the governance approach allows them to benefit via the use of specific levers (i.e., access time, integration with proprietary data, rights to patent fully developed drugs.)

Finally, we extend the classic debate between top-down and bottom-up governance models: our case bridges the two with an alternative approach in which a broker, a trusted third party, is assigned coordination and arbitration tasks to orchestrate and mediate the flows of data and knowledge. According to our case data, folding and unfolding processes were allowed not only by the technical attributes of the modular infrastructure but also through an organizational component whereby an operational team with bilateral trust helped move the organizations from one layer to another. This governance form offers a more nuanced perspective to the classic portrayal of top-down or bottom-up governance approaches as discrete, static alternatives.

### *4.7.2 Implications for Practice*

By many measures, OT represents an extreme case of an information infrastructure: the partners include several of the world's most accomplished research and scientific institutions together with some of the most highly capitalized companies in the pharmaceutical sector. In this sense, it represents an exceptional case in which the ethos of scientific knowledge as a social good and profit-seeking business investment intertwine. While it is unreasonable to expect that all information infrastructures will operate with such acute opposition between public and private interests, most collaborative infrastructures and platforms with heterogeneous contributors will have divergent or competing objectives in some form

(Wareham et al. 2014). We suggest that the governance mechanisms that permit a dynamic transitions among varying layers of openness and closure, both technical (i.e., modularity) and organizational (i.e., brokerage), are in some sense generalizable to alternative forms of information infrastructures characterized by divergent motives and institutional objectives (European Commission 2019, Hey 2009).

A leitmotif in our analysis is that the general increase in novel data-intensive computational methods in drug discovery and development is not unique to the pharmaceutical sector. Rather, it is a tendency we expect to observe across other scientifically intensive industries. As this evolution continues, we expect to see a number of industries transformed in response to increased reliance upon computationally intensive research and development methods. It can also be expected that the movement to combine resources and share costs towards what might be considered precompetitive or public-good outcomes will increase. While preliminary, our analysis of OT offers a model of how one might map similar structural shifts in related industries.

### 7. Conclusion

The pharmaceutical industry is notorious for its reliance on patent protection and secrecy. Open Targets has been celebrated as a model for other industry organizations and policymakers due to its form of governance that enables the participation of for-profit pharma companies in a shared information infrastructure alongside non-profit foundations and public research institutions; the benefits of shared investments and public-good outcomes can be realized while simultaneously protecting the private economic interests of the OT members. This is attributable to a governance form that allows fluid navigation from openly shared and private workspaces. Our case focuses on modularity and brokerage as general conditions that make members' private and collective interests compatible. As computationally intensive research and development methods pervade other industries, we can expect a commensurate increase in required investment levels, rendering them prohibitive to individual organizations. As commons-based information infrastructures emerge in response, the governance described in this case offer a model of how to successfully navigate the trade-offs between private and collective interests.

## References

Aligica PD, Tarko V (2012) Polycentricity: From Polanyi to Ostrom, and Beyond. *Governance* 25(2):237–262.

Allarakhia M (2014) The successes and challenges of open-source biopharmaceutical innovation. *Expert Opinion on Drug Discovery* 9(5):459–465.

Altshuler JS, Balogh E, Barker AD, Eck SL, Friend SH, Ginsburg GS, Herbst RS, Nass SJ, Streeter CM, Wagner JA (2010) Opening Up to Precompetitive Collaboration. *Science Translational Medicine* 2(52):52cm26-52cm26.

Alvesson M (2009) *Reflexive methodology: new vistas for qualitative research* 2nd ed. (SAGE, Los Angeles).

Au R (2014) The paradigm shift to an "open" model in drug development. *Applied & Translational Genomics* 3(4):86–89.

Behan FM, Iorio F, Picco G, Gonçalves E, Beaver CM, Migliardi G, Santos R, et al. (2019) Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens. *Nature* 568(7753):511–516.

Benkler Y (2006) *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. New Haven, CT: Yale University Press.

Bowker GC (1999) *Sorting things out: classification and its consequences* (MIT Press, Cambridge, Mass).

Burt RS (2000) The Network Structure Of Social Capital. *Research in Organizational Behavior* 22:345–423.

Byrd TA, Turner DE (2001) An Exploratory Analysis of the Value of the Skills of IT Personnel: Their Relationship to IS Infrastructure and Competitive Advantage. *Decision Sciences* 32(1):21–54.

Chaguturu R, Murad F, Murad F (2014) *Collaborative Innovation in Drug Discovery: Strategies for Public and Private Partnerships* (John Wiley & Sons, Incorporated, Somerset, United States).

Choudhury S, Fishman JR, McGowan ML, Juengst ET (2014) Big data, open science and the brain: lessons learned from genomics. *Frontiers in Human Neuroscience* 8.

Chung SH, Rainer Jr, Lewis BR (2003) The Impact of Information Technology Infrastructure Flexibility on Strategic Alignment and Application Implementations. *CAIS* 11.

Ciborra CU, Andreu R (2001) Sharing knowledge across boundaries. *Journal of Information Technology* 16(2):73–81.

Collins FS (2003) The Human Genome Project: Lessons from Large-Scale Biology. *Science* 300(5617):286–290.

Constantanides P (2012) *Perspectives and Implications for the Development of Information Infrastructures* (IGI Global).

Constantinides P, Barrett M (2015) Information Infrastructure Development and Governance as Collective Action. *Information Systems Research* 26(1):40–56.

Contreras JL (2010) Prepublication Data Release, Latency, and Genome Commons. *Science* 329(5990):393–394.

Dean JW, Sharfman MP (1996) Does Decision Process Matter? A Study Of Strategic Decision-making Effectiveness. *Academy of Management Journal* 39(2):368–392.

Deleuze G, Strauss J (1991) The Fold. *Yale French Studies* (80):227–247.

Duncan NB (1995) Capturing Flexibility of Information Technology Infrastructure: A Study of Resource Characteristics and Their Measure. *Journal of Management Information Systems* 12(2):37–57.

Eisenhardt KM (1989) Building Theories from Case Study Research. *Academy of Management Review* 14(4):532–550.

European Commission (2019) Facts and figures of open research data. *European Commission - European Commission*. Retrieved (April 19, 2019), https://ec.europa.eu/info/research-and-innovation/strategy/goals-research-and-innovation-policy/open-science/open-science-monitor/facts-and-figures-open-research-data_en.

Evaluate Pharma (2017) *World Preview 2017* Retrieved (April 19, 2019), https://info.evaluategroup.com/rs/607-YGS-364/images/WP17.pdf

Fletcher JA, Zwick M (2000) Simpson's Paradox Can Emerge from the N-Player Prisoner's Dilemma: Implications for the Evolution of Altruistic Behavior. In Proceedings of The World Congress of the Systems Sciences and ISSS 2000, Allen, J.K. and Wilby, J.M. eds, Toronto, Canada: International Society for the Systems Sciences.

Greco GM, Floridi L (2004) The tragedy of the digital commons. *Ethics and Information Technology* 6(2):73–81.

Grisot M, Hanseth O, Thorseng A (2014) Innovation Of, In, On Infrastructures: Articulating the Role of Architecture in Information Infrastructure Evolution. *Journal of the Association for Information Systems* 2014, 15(4): 197-219

Grossman RL, Heath A, Murphy M, Patterson M, Wells W (2016) A Case for Data Commons: Toward Data Science as a Service. *Computing in Science & Engineering* 18(5):10–20.

Hanseth O (2001) GatewaysmJust as Important as Standards: How the Internet Won the"Religious War" over Standards in Scandinavia. *Knowledge, Technology and Policy* 14(3): 71–89.

Hanseth O, Jacucci E, Grisot M, Aanestad M (2006) Reflexive Standardization: Side Effects and Complexity in Standard Making. *MIS Quarterly* 30:563–581.

Hanseth O, Lyytinen K (2010) Design Theory for Dynamic Complexity in Information Infrastructures: The Case of Building Internet. *Journal of Information Technology* 25(1):1–19.

Hanseth O, Monteiro E (1997) Inscribing behaviour in information infrastructure standards. *Accounting, Management and Information Technologies* 7(4):183–211.

Hardin G (1968) The Tragedy of the Commons. *Science* 162(3859):1243–1248.

Hardin R (1982) *Collective action* (Published for Resources for the Future by the Johns Hopkins University Press, Baltimore).

Henderson RM, Clark KB (1990) Architectural Innovation: The Reconfiguration of Existing Product Technologies and the Failure of Established Firms. *Administrative Science Quarterly* 35(1):9–30.

Henfridsson O, Bygstad B (2013) The Generative Mechanisms of Digital Infrastructure Evolution. *MIS Quarterly* 37(3):907–931.

Henfridsson O, Mathiassen L, Svahn F (2014) Managing Technological Change in the Digital Age: The Role of Architectural Frames. *Journal of Information Technology* 29(1):27–43.

Hess C, Ostrom E (2003) Ideas, Artifacts, and Facilities: Information as a Common-Pool Resource. *Law and Contemporary Problems* 66(1/2):111–145.

Hey T (2009) *The Fourth Paradigm: Data-intensive Scientific Discovery* (Microsoft Pr, Redmond, Washington).

Hood L, Rowen L (2013) The Human Genome Project: big science transforms biology and medicine. *Genome Medicine* 5(9):79.

Hopkins MM, Martin PA, Nightingale P, Kraft A, Mahdi S (2007) The myth of the biotech revolution: An assessment of technological, clinical and organisational change. *Research Policy* 36(4):566–589.

Kahn RE, Cerf VG (1988) An open architecture for a digital library system and a plan for its development. The digital library project vol. 1: The world of knowbots.:48.

Katsila T, Spyroulias GA, Patrinos GP, Matsoukas MT (2016) Computational approaches in target identification and drug discovery. *Computational and Structural Biotechnology Journal* 14:177–184.

Kavanagh D, Araujo L (1995) Chronigami: Folding and unfolding time. *Accounting, Management and Information Technologies* 5(2):103–121.

Khaladkar M, Koscielny G, Hasan S, Agarwal P, Dunham I, Rajpal D, Sanseau P (2017) Uncovering novel repositioning opportunities using the Open Targets platform. *Drug Discovery Today* 22(12):1800–1807.

Klein HK, Myers MD (1999) A Set of Principles for Conducting and Evaluating Interpretive Field Studies in Information Systems. *MIS Quarterly* 23(1):67.

Koscielny G, An P, Carvalho-Silva D, Cham JA, Fumis L, Gasparyan R, Hasan S, et al. (2017) Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Research* 45(D1):D985–D994.

Lee WH (2015) Open Access Target Validation Is a More Efficient Way to Accelerate Drug Discovery. *PLOS Biology* 13(6):e1002164.

Loging W, Harland L, Williams-Jones B (2007) High-throughput electronic biology: mining information for drug discovery. *Nature Reviews Drug Discovery* 6(3):220–230.

Lok J, de Rond M (2012) On the Plasticity of Institutions: Containing and Restoring Practice Breakdowns at the Cambridge University Boat Club. *Academy of Management Journal* 56(1):185–207.

Ma M, Agarwal R (2007) Through a Glass Darkly: Information Technology Design, Identity Verification, and Knowledge Contribution in Online Communities. *Information Systems Research* 18(1):42–67.

Markus ML, Steinfield CW, Wigand RT (2006) Industry-Wide Information Systems Standardization as Collective Action: The Case of the U.S. Residential Mortgage Industry. *MIS Quarterly* 30:439–465.

Mindel V, Mathiassen L, Rai A (2018) The Sustainability of Polycentric Information Commons. *MISQ* 42(2):607–631.

Monge PR, Fulk J, Kalman ME, Flanagin AJ, Parnassa C, Rumsey S (1998) Production of Collective Action in Alliance-Based Interorganizational Communication and Information Systems. *Organization Science* 9(3):411–433.

Monteiro E (1998) Scaling Information Infrastructure: The Case of Next-Generation IP in the Internet. *The Information Society* 14(3):229–245.

Moon JY, Sproull LS (2008) The Role of Feedback in Managing the Internet-Based

Volunteer Work Force. *Information Systems Research* 19(4):494–515.

Nagendra H, Ostrom E (2012) Polycentric governance of multifunctional forested landscapes. *International Journal of the Commons* 6(2):104–133.

Obstfeld D, Borgatti SP, Davis J (2014) Brokerage as a Process: Decoupling Third Party Action from Social Network Structure. Brass DJ, Labianca G (JOE), Mehra A, Halgin DS, Borgatti SP, eds. *Research in the Sociology of Organizations*. (Emerald Group Publishing Limited), 135–159.

Ostrom E (1986) An Agenda for the Study of Institutions. *Public Choice* 48(1):3–25.

Ostrom E (1990) *Governing the Commons: The Evolution of Institutions for Collective Action* (Cambridge University Press).

Pogorelc D (2014) To Spark Cancer Discoveries, Several Big Pharma Companies are Sharing Idle Clinical Trial Data. *MedCity* (April 8) http://medcitynews. com/2014/04/sanofi-pfizer-jj-astrazeneca-share-clinical-trial-data/.

Polanyi M (1951) *The logic of liberty: reflections and rejoinders* (University of Chicago Press, Chicago).

Ransbotham S, Kane GC (Jerry) (2011) Membership Turnover and Collaboration Success in Online Communities: Explaining Rises and Falls from Grace in Wikipedia. *MIS Quarterly* 35(3):613–627.

Sahay S, Aanestad M (2009) Configurable Politics and Asymmetric Integration: Health e-Infrastructures in India. *Journal of the Association for Information Systems* 10(5):399–414.

Sakakibara M (2002) Formation of R&D consortia: industry and company effects. *Strategic Management Journal* 23(11):1033–1050.

Sambamurthy V, Zmud RW (2000) Research Commentary: The Organizing Logic for an Enterprise's IT Activities in the Digital Era—A Prognosis of Practice and a Call for Research. *Information Systems Research* 11(2):105–114.

Samuelson PA (1954) The Pure Theory of Public Expenditure. *The Review of Economics and Statistics* 36(4):387–389.

Schrattenholz A, Soskic V (2008) What Does Systems Biology Mean for Drug Development? *Current Medicinal Chemistry* 15(15):1520–1528.

Schuhmacher A, Gassmann O, Hinder M (2016) Changing R&D models in research-based pharmaceutical companies. *Journal of Translational Medicine* 4(1): 105.

Shaikh M, Vaast E (2016) Folding and Unfolding: Balancing Openness and Transparency in

Open Source Communities. *Information Systems Research* 27(4):813–833.

Siggelkow N (2007) Persuasion With Case Studies. *Academy of Management Journal* 50(1):20–24.

Star SL, Ruhleder K (1996) Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces. *Information Systems Research* 7(1):111–134.

Stewart KJ, Gosain S (2006) The Impact of Ideology on Effectiveness in Open Source Software Development Teams. *MIS Quarterly* 30(2):291–314.

Stovel K, Shaw L (2012) Brokerage. *Annual Review of Sociology* 38(1):139–158.

Tufts Center for the Study of Drug Development (2015) Personalized Medicine Gains Traction but Still Faces Multiple Challenges, Impact Report. (June).

Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, Li B, et al. (2019) Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery* 18(6):463–477.

Vassilakopoulou P, Skorve E, Aanestad M (2016) A commons perspective on genetic data governance: the case of BRCA data. In *European Conference On Information System (ECIS)*. Association For Information System (AIS).

Vincent C, Camp J (2004) Looking to the Internet for models of governance. *Ethics Information Technol*ogy 6(3):161–173.

Wareham J, Fox PB, Giner JLC (2014) Technology Ecosystem Governance. *Organization Science*. 25(4):1195-1215.

Weill P, Broadbent M (1998) *Leveraging the New Infrastructure: How Market Leaders Capitalize on Information Technology* (Harvard Business Press).

Weill P, Ross JW (2004) *It Governance on One Page* (Social Science Research Network, Rochester, NY).

West J, Cutting-Decelle AF, Mignon S (2016) Standardization and Coopetition: A Study of Pharmaceutical Industry Consortia. *EURAS Proceedings*:18.

Yin RK (2003) *Case Study Research: Design and Methods* (SAGE).

Yoo Y, Boland RJ, Lyytinen K, Majchrzak A (2012) Organizing for Innovation in the Digitized World | Organization Science. *Organization Science*. 23(5): 1398-1408.

Zanders ED (2011) *The science and business of drug discovery: demystifying the jargon* (Springer, New York).

## Appendix A

*Table 4. Construct definitions*

| Construct definition | Source |
|---|---|
| **Openness:** access to information. Related to the act of sharing information, it is a key featured exhibited by information infrastructures | e.g., (Hanseth and Lyytinen 2010, Shaikh and Vaast 2016) |
| **Information infrastructure**: "multi-layered entities comprised of technological components, people and institutional arrangements" | (Constantanides 2012 p. 25) |
| **Commons**: A set of resources that are collectively owned and shared among a community. Commons contain public and private property over which different agents have certain rights. | (Ostrom 1990) |
| **Collective action**: refers to joint action by a number of agents to achieve and distribute some gain through coordination or cooperation | (Hardin 1982) |
| **Drug discovery**: "the process of creating chemical or biological molecules that have the potential to be developed as therapeutic agents, typically because they generate a desired biological effect in an appropriate testing or assay system against a particular molecular (drug) target". | (Weigelt, 2009, p. 941) |
| **Governance**: the rules that underlie the social activities, which are integral to order relationships, responsibilities, and expectations of contributors. | e.g., (Mindel et al. 2018, Weill and Ross 2004) |
| **Modularity:** An information infrastructure is modular in that it has the ability to add, modify and remove technological components with little effect on its features and the processes of other components. The modular attribute of infrastructure leads to the subprinciples of decomposition, recombination and reusability of infrastructure components | e.g., (Byrd and Turner 2001, Chung et al. 2003, Duncan 1995, Henfridsson et al. 2014) |
| **Brokerage:** the process of connecting actors in systems of social, economic, or political relations to facilitate access to valued resources | e.g., (Burt 2000, Obstfeld et al. 2014, Stovel and Shaw 2012) |
| **Folding:** the process by which organizations create 'private pockets of interactions' inside their organizations (i.e., the movement of going private) and limit the openness of the information infrastructure | (Shaikh and Vaast 2016) |
| **Unfolding:** the process by which organizations release private information from private interactions into the shared or public layer of the infrastructure | (Shaikh and Vaast 2016) |

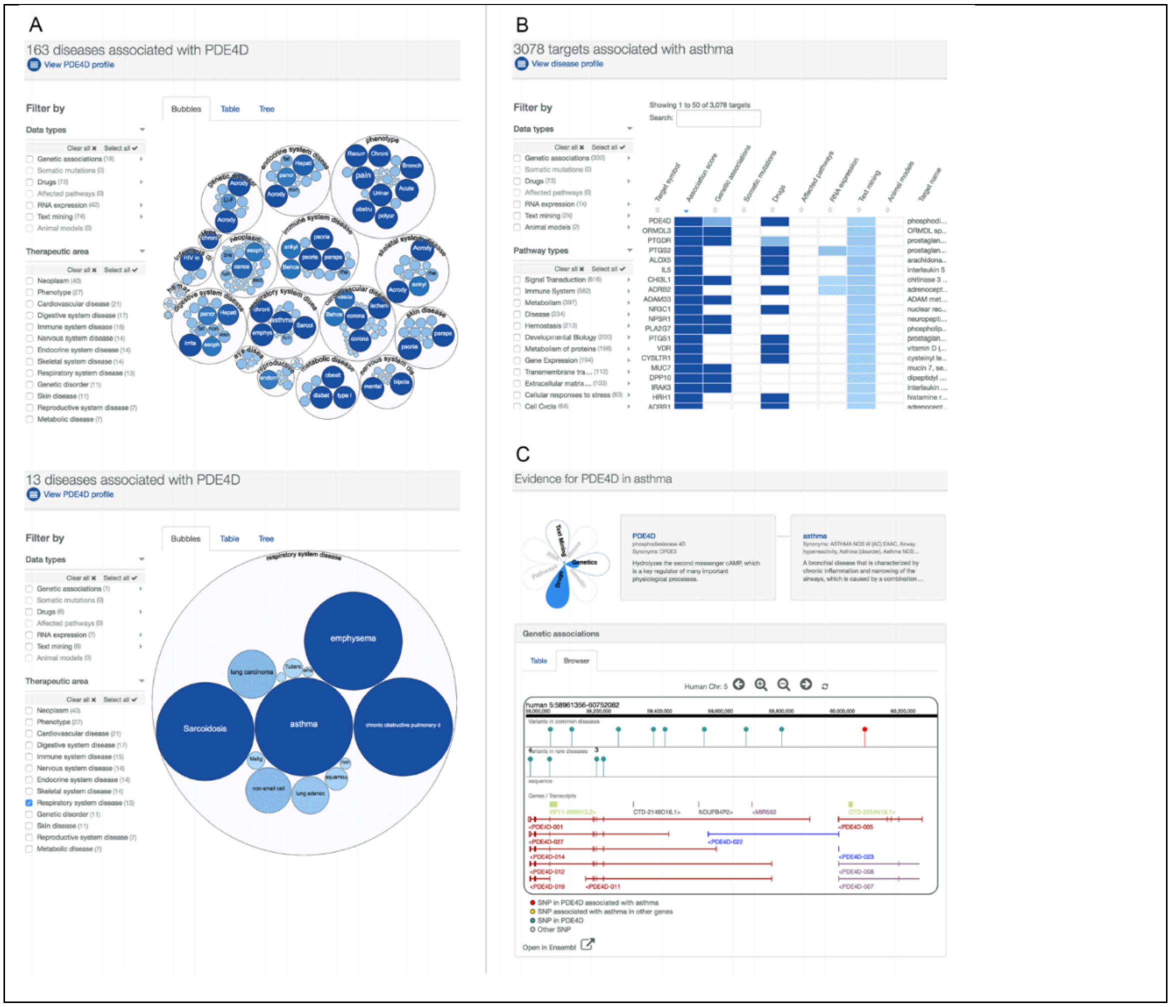
# Appendix B

Example of Interview Guide

| **About the organization, role and responsibilities** |
|---|
| • What does your organization do? |
| • What is your role at the organization? |
| • Which activities have you developed in OT? |
| **Initial engagement** |
| • How did organization become aware of and initially get involved in OT? |
| • Why did you want to join OT? |
| • Why did your organization agree not to patent targets? What has this decision meant for your organization? |
| • How was the process of taking part in OT? |
| • Which processes did you undertake within your organization to join OT? |
| **About OT infrastructure** |
| • What are the components, functions and applications of OT? |
| • What is the difference between OT public and OT access for partners? |
| • How are the data, technologies, and methods generated in the collaboration within OT integrated into the OT infrastructure? |
| • How are the data, technologies, and methods generated in the collaboration within OT integrated into your organization? |
| • Which processes did you follow to make such integration possible outside of OT collaboration and within your organization? |
| **About OT collaboration** |
| • What do you share in OT projects? |
| • What do you not want to share in OT projects? |
| • How do you control what your team does not share with OT in keeping with what the organization does not want to share? In other words, what processes do you follow so that the information shared is only relevant for targets and not for competitive phases in drug discovery? |
| • How do you use the knowledge from OT collaboration in the subsequent drug discovery process? How do you reuse the data, technologies or methods? |
| • What have been the positive and negative effects of collaborating with your competitors? |
| **The role of OT operational team in the collaboration** |
| • How is the OT operational team selected? |
| • How does the OT team help select projects? |
| • How does the OT team manage the process of publishing the data, methods, technologies and any output in the infrastructure? |
| • How does the OT team help you to integrate such outputs with your internal processes to proceed in the drug discovery process? |

# Appendix C

*Figure 4. Example of Visualization of Search Results in OT Infrastructure (Koscielny et al. 2017)*



PDE4D
phosphodiesterase 4D
Synonyms: DPDE3
Hydrolyzes the second messenger cAMP, which is a key regulator of many important physiological processes.

asthma
Synonyms: ASTHMA NOS W (AC) EXAC, Airway hyperreactivity, Asthma (disorder), Asthma NOS…
A bronchial disease that is characterized by chronic inflammation and narrowing of the airways, which is caused by a combination …

## Appendix D

*Table 5. Data Analysis and Theoretical Constructs*

| **Illustrative examples of empirical observations and excerpts from the interviews** | **Theoretical observations** | **Theoretical constructs** | **Category in theorization** |
|---|---|---|---|
| *“Partners (referring to company partners) take the public instance of OT and they replicate that in a private instance, and then they will inject their own private data (…) They want their internal teams to make decisions with their private data but also knowing what already existing data tell us and how our data integrate with that data” OT4*<br>*“Sometimes, their private instances (referring to companies) are managed entirely internally by them, but other times they involve a third party, a vendor, to assist them in the maintenance and update. They try to align with our releases, we release five times a year, every two months or so, and they take the last data releases and refresh [them]with any additional features that we release, and they replicate that internally.” OT5*<br>*“Our development team would sit with the company members or the intermediary vendor and work through how we run the pipelines and how we can get their proprietary data in. And then, we go through their different requirements, so we work and see how we can hand something over to them that can be simply configured for them when they take it inside” OT5* | **Process of folding**<br>- Integrate consortium data or public data with company proprietary data<br>- Collaboration with OT team to migrate the shared or open data to proprietary systems in a bilateral collaboration<br>- Non-disclosure agreement with OT team to help integrate public or shared data with proprietary data.<br>- Agreement on data standards and metadata compatible with the structure of private datasets to enable locking<br>- Protocol agreement on data flows to guarantee control points before releasing or keeping data private. | Folding | Governance processes |
| *“When data is ready, we integrate it into the platform; we need to wait until it is ready and publicize and then we enter it in the platform at that point and release it” OT4.*<br>*“The platform team (in charge of releasing the data) get to see the type of data very early in the process. They have sample data, and they discuss the format. We also have a UX specialist, to understand what the deliverable is and how we manifest it in the platform so that people can use it to make a decision. This is the foundation for data specification, really. How are we going to receive it, what does it looks like, how will it be processed? The discussions are very early on, and we try to get mock ups very early on, to gather feedback from the consortium partners but also from other users, and then we kind of refine them that as we go along (…) It is a moving target, as some of the projects do not know what the data* | **Process of unfolding**<br>- Public release of data in OT infrastructure originating from OT consortium projects announced through OT dissemination networks.<br>- Data curation process by OT team before releasing the data<br>- Formal approval by OT scientific lead committee represented by all organizations of data, methods, and tool release | Unfolding | |

| | | | |
|---|---|---|---|
| *will look like, so we have monthly meetings." OT4* | | | |
| *" We have other features that are private, that we do not share with others. Those hidden features allow me, for instance, to work with my compound library on the platform, which I do not share with other OT partners" (MT2).*<br>*"We also can ask for new features in the platform. We can also implement the platform in-house, within the company, integrate it with our systems, and you receive support from the other members of the consortium" (OT1).*<br>*"So, we have the platform that is public and open to everybody. Then, for the experimental projects, the partners share the data while they are creating it in google buckets*[10] *"*<br>*"We have an intranet for the consortium partners. It is information exchange between partners (...) The intranet has a link with the platform, and it is used for general governance of the projects. As we go through the project call processes, there are page proposals to share the details. It is like a one-stop shop for the whole portfolio of projects" OT4.* | **Access rights**<br>- Public: anyone can access the data<br>- Shared: only consortium partners can access the data<br>- Private: only company members can access the data | Modularity | Conditions creating the fold–unfold: Technical attributes of information infrastructures |
| *"Everybody that joins OT understands the premise for being here. There is limited access without a doubt. Everybody understands that until there is a formal publication after the project there is no disclosure."* | **Time latency**<br>- Two years on average from the generation of the datum until it is released in the OT public layer. | | |
| *"Companies usually have a flag for the data that is private and what is public in their private instances because for them it is very important to differentiate that" OT4* | **Domain or boundary demarcation**<br>- Delimitation of data and knowledge boundaries that correspond to public, shared and private layers. | | |
| *"We never get the whole, the full set of data. And sometimes we only get dummy data, which is fine because we only need to see the structure of the data. And likewise with the consortium members (referring to when they emulate Open Target in private instances) they do need our help and say, ok I have my private instance, now how do I inject my private data, but they do not want to share their data with us, so we talk about it in the context; we talk about the structure of their data, so to some extent it is blinded" OT4.* | **Interoperability and data standards**<br>Agreement among OT partners on the structure and labeling of the data to make them integrable with public and private layers. | | |

[10] A bucket in google cloud storage is a basic container that holds data. Owners of buckets control access to the data.

| | | | |
|---|---|---|---|
| *“Any access to data on an experimental project is provisioned through a person on the operations team at OT. We do have a gatekeeper for that. Any person from our team asking for data needs to go through them” OT5*<br>*“There is a need to coordinate the integration of data into OT, both from the projects that generate data but also with the data providers such as Chembl and Uniprot and all the data that goes into the platform to keep it up to date. We also work with the developer team that creates some of the features that users will use to visualize the data coming through.” OT4* | **Trusted third party**<br>-Selection of an executive and operational team that behaves as a separate entity financially reporting to the governance board representing all organizations.<br>- Coordination and mediation role when conflicts of interest arise among company members | Brokerage | Conditions creating the fold–unfold: organizational attributes of information infrastructures |
| *“To become a member, you need to share the vision of the consortium [open disclosure] and agree on investing around at least two-digit million euros, and all partners have to accept your membership." OT1*<br>*"We have to be careful; this is why we have only one company join every year. It takes time to integrate a company into the consortium." MT2*<br>*“EMBL EBI, Wellcome Trust [the academic partners] are part of this philosophy to make raw data available. They are academic institutions, and this is easy for them. However, the question of openness is a question for companies. Not all companies agree to be open and share their knowledge and information about a discovery. This is a major issue. Some companies agree to open up their knowledge about a discovery, but other companies do not join because they do not believe in this strategy. They do not want to share" (MT2)* | **Bilateral negotiations**<br>- Bilateral negotiations before entering OT consortium to agree on the conditions of the collaboration (i.e., resources to be invested and policies about data releases and open collaboration among partners) | | |
| “*We implement brokerage and matchmaking exercises to create the working teams and projects"* (OT1*)* | **Pooling common resources**<br>- Identification of the disperse resources that are being invested in the different company teams subject to become a joint collaborative project. Coordination of the process through calls for proposals and open collaboration days to identify such common ground. | | |
| *“We have regular meetings with the consortium members. They come to us with challenges on data integration, and suggestions, and then we feed it back to the developers, and try to have something improved for them (...) We get also suggestions on the features or data that we should get from partners.”OT4*<br>*“The requirements from the different companies are framed differently. We have meetings on a quarterly basis. We prioritize the requirements. There is a roadmap” OT4* | **Bilateral support for integration**:<br>- Bilateral NDA agreements with the OT operational team to support integration of shared and public data to private systems. | | |

# 5

## 5. Serendipity in Big Science Infrastructures

The article that constitutes this chapter aims at understanding the second vector: technology transfer, responding to the second sub-goal (1.2.) of our PhD investigation. The study seeks to understand the nature of the serendipitous process behind transferring big science technologies to commercial applications by empirically investigating the European initiative: ATTRACT

## 5.1 Abstract

This paper explores how policy can promote the application of scientific research beyond its original purview. We analyze ATTRACT[11], a novel policy instrument in the European Commission's Horizon 2020 program, aiming to harness the detection and imaging technologies of the leading European research infrastructures towards entrepreneurship. In this initiative, 170 projects were funded with €100,000 for each to develop a proof-of-concept commercial application within one year. Leveraging the unique dataset from the projects funded under ATTRACT, our study finds different serendipity modes compared to the previously proposed typologies, as follows: (1) building on previous research, (2) combining different technologies, (3) applying a technology into a different field, and (4) using artificial intelligence or machine learning. This study contributes to the emerging literature on serendipity by showing the potential of policy interventions to enable individuals and organizations to find unexpected commercial applications of big science research.

**Keywords:** big science, serendipity, high-imaging technology

## 5.2 Introduction

Some of the most pervasive technologies in society today, such as the internet, medical diagnostics and treatments, and information and communication technologies, result from leveraging the research generated by big science infrastructures to areas beyond their direct scientific purview. While the potential of big science to create social, cultural, and economic impacts is acknowledged, uncertainty remains on how these big science infrastructures can deliberately find novel applications outside of their immediate scopes of research. Moreover, there are also questions regarding the extent to which policymakers can play an active role in enabling these research centers to find novel uses for their research that were previously unanticipated. Exploring these questions, this paper examines a novel policy response by the European Union to promote the commercialization of technologies from some of Europe's most impactful research infrastructures.

[11] The members of ATTRACT are as follows: the European Organization for Nuclear Research (CERN), European Molecular Biology Laboratory (EMBL), European Southern Observatory (ESO), European Synchrotron Radiation Facility (ESRF), European X-Ray Free Electron Laser Facility (European XFEL), and the Institut Laue-Langevin (ILL), Aalto University, ESADE Business School, and the European Industrial Research
Management Association (EIRMA).

The term serendipity has been evoked to describe various unintended discoveries, typically with some beneficial outcomes. For example, Fleming's discovery of penicillin is often cited as a serendipitous discovery with tremendous social value. The definition of serendipity, however, can be ambiguous. The Merriam Webster dictionary defines serendipity as "the faculty or phenomenon of finding valuable or agreeable things not sought for" (Merriam-Webster, 2020), while the Oxford dictionary defines it as "the occurrence and development of events by chance in a happy or beneficial way" (Oxford University Press, 2019). In the management and innovation literature, creating conditions that foster serendipity is considered desirable for managers and policy-makers (Yaqub, 2018).

On the surface, the argument that serendipity can play a positive role in scientific processes and policy has its immediate value as ex-post, anecdotal narratives with limited normative value. However, this misconception comes from interpreting serendipity as mere happenstance instead of resulting from deliberate effort (de Rond, 2014). A systematic analysis of serendipity is useful because it offers a more nuanced understanding of its antecedents and mechanisms (e.g., Yaqub, 2018; Garud 2018). By identifying the formative conditions of serendipity, the design of mechanisms to realize the peripheral benefits of scientific research infrastructures can be improved; in effect, one could attempt to systematize serendipity. However, to date, most of the research has been speculative or based on small-sample, anecdotal evidence from previous scientific discoveries.

This study examines the ATTRACT project, a €20M-funded initiative within the Horizon 2020 Framework Program that aims to systematize the discovery of breakthrough applications of imaging and detection technologies from the leading European science research infrastructures. Recognizing that the full potential of these detection and imaging technologies is unknown, ATTRACT was formulated with the understanding that capturing the value of big science will require both stimulating exploration and the simultaneous fostering of commercial development through risk absorption and support. Accordingly, ATTRACT supported 170 projects with seed-funding grants of €100,000 each to leverage their various technologies towards sustainable businesses and greater economic returns for the European economy.

Analyzing how the large research infrastructures can find new impactful uses for their science is highly relevant. Given their extreme sophistication and required investment levels, research infrastructures are normally funded by taxpayers via national ministries or funding agencies – often in pan-national consortia. As such, it bears upon policymakers to seek mechanisms to optimize the potential socioeconomic value of these public investments. ATTRACT brings six of the largest European scientific research infrastructures, which are also members of the EIROforum, together; they are as follows: European Organization for

Nuclear Research (CERN), European Molecular Biology Laboratory (EMBL), European Southern Observatory (ESO), European Synchrotron Radiation Facility (ESRF), European X-ray Free Electron Laser Facility (European XFEL), and the Institut Laue-Langevin (ILL). These organizations work in diverse domains, such as nuclear, particle, and condensed matter physics; life sciences; molecular biology; astronomy; materials science; structural biology; and chemistry.

The 170 projects funded under ATTRACT provide a unique dataset to examine the processes towards serendipity. In this analysis, we find the following modes: (1) building on previous research, (2) combining different technologies, (3) applying technology into a different field and (4) using artificial intelligence or machine learning. We validate the previous typologies of serendipity and extend these notions by describing new categories. Unlike the previous studies that examined serendipity ex-ante, this study explores purposeful actions carried out in the pursuit of serendipity. Moreover, we explore how the intentional nature of the policy intervention by ATTRACT can help in finding new, previously unexplored applications of research technologies.

The article proceeds by reviewing the history of big science, the polemics of its underlying social value, and the mechanisms and measures that policymakers use to stimulate the application of science towards social and economic impacts. We describe the literature on serendipity, summarizing the extant literature and the main research questions. We present the ATTRACT project and explore how it attempts to systematize serendipity. Contributing to the serendipity literature, we summarize the 170 projects funded under the call and examine the various modes used to discover previously unanticipated applications. We conclude with observations concerning serendipity and describe trajectories for future initiatives concerning big science and socioeconomic value.

## 5.3 Background: big science and social impact

In the following, we explore the history of big science and the issues related to its impact on society.

### *5.3.1 Definition and History*

Big science infrastructures are defined by Florio and Sirtori (2016) as institutions with a) high capital intensity, b) long-lasting facilities or networks, c) operating in monopoly or oligopoly conditions affected by externalities, d) who produce social benefits via the generation of new knowledge (either pure or applied). As argued by Giudice (2012), the evolution of big science began early in the twentieth century with examples such as the factory-like conditions where Heike Kamerlingh Onnes made seminal discoveries on superfluidity and superconductivity in the early 1900s, or the Wilson Observatory, completed in 1917 and made famous by Edwin Hubble. What began to characterize research as big science was how it differed from the ideal of the lone genius in the laboratory with simple table-top experiments.

This new model of scientific exploration was fully institutionalized by Ernest Orlando Lawrence at the University of California, Berkeley with the development of the cyclotron, which is a device for accelerating nuclear particles to very high velocities to bombard, disintegrate and form completely new elements and radioactive isotopes. While the first cyclotron was merely a simple 4-inch device that could be held in the human hand, over time, larger versions that could achieve greater energy levels were created. With each subsequent generation of the cyclotron, a larger number of physicists, engineers, and chemists were needed for construction, operation, and maintenance. More importantly, he advanced a form of team-based, collaborative science that contrasted with the isolated model of 'smaller science'[12] (Hiltzik, 2016) and later matured into large research teams with hundreds of scientists and engineers. This new type of industrialized science eventually propagated to other American and European universities and was facetiously called the 'Cyclotron Republic' by Lawrence's numerous admirers and rivals (Hiltzik, 2016).

The cyclotron also provides an early example of how big science research can have alternative applications for socioeconomic impact. A serendipitous by-product of Lawrence's lab was the production of radioactive isotopes useful for cancer treatment (Hiltzik, 2016). With the help of his brother John Lawrence, a medical doctor who became the director of the university's Medical Physics Laboratory, Ernest was able to recraft the cyclotron's narrative to court funders intrigued by the potential of important isotopes. In a Faustian spirit, the laboratory metaphorically produced oncology-focused isotopes by day, while discretely conducting basic research by night, and while many on the team bemoaned the fact that commitments to medical research hindered advancement in fundamental physics, this shrewd strategy enabled Lawrence to fund his constantly moving targets of higher energy levels that required more sophisticated hardware, complex operating organizations, and generated unprecedented costs. This tactic further institutionalized the future relationship between big science and big funders, be they philanthropies, national ministries of defense or energy, or increasingly, supranational-coalitions (Crease et al., 2016).

[12] Quoting Luis Alvarez in Hiltzik (2015): There were no doors inside the Rad Lab. 'Its central focus was the cyclotron, on which everyone worked and which belonged to everyone equally (though perhaps more to Ernest). Everyone was free to borrow or use everyone else's equipment or, more commonly, to plan a joint experiment'. The team approach to physics, Alvarez judged, was 'Lawrence's greatest invention'. (Hiltzik 2015:129–30).

The rise of big science, however, is often associated with the Manhattan Project and the numerous technological innovations that were enhanced during WWII, such as radar and wireless communication. Motivated primarily by military and global political concerns, technological superiority was considered a central element of geopolitical competition (Galison and Hevly, 1992). This superiority was not limited to military research, although the defense industry was certainly a central protagonist. Espoused in the famous report of Vannevar Bush (1945), *Science: The Endless Frontier*, basic research was not only good for fundamental science but generated applied engineering and technologies that translated into products, spin-offs, jobs, and overall economic prosperity that benefited all social classes. The 'Bush legacy' (Wilson, 1991) was further catalyzed by the successful leap-start of the Soviet space program, an event that galvanized the American public to approve the astronomical funding levels of the American space program while having little concern for its scientific merit. With the perceived technological gap between the USA and the USSR, the Soviet space program was considered a severe existential threat that, similar to the Manhattan Project, could only be remedied by massive investments in basic, applied, and ever-bigger science (Giudice, 2012).

Currently, with the cold war decades in the past, the role of big science in society has transformed. The perception of grand existential geopolitical threats has turned into a more disperse narrative. As a result, investments in big science motivated by national security or geopolitical stability have decreased. This decrease has weakened the sacrosanct link between nuclear physics, weapons research, and geopolitical security and, as a consequence, has reduced the primacy of fundamental physics (Galison and Hevly, 1992; Hiltzik, 2016). Moreover, the tenacious success of the Standard Model has left aspiring physicists scrambling for new avenues to conduct physics, leading them to astrophysics and cosmology, as well as more distant fields, such as biology and life sciences (Galison, 2016).

In addition, the nature of big science infrastructures has become more heterogeneous. Today, traditional particle accelerators and nuclear reactors work alongside synchrotron radiation, neutron scattering, and free electron laser facilities, where the empirical scope has widened to materials science, chemistry, energy, condensed matter physics, nanoscience, biology, biotechnology and pharmacology (Doing, 2018; Heinze and Hallonsten, 2017). Finally, big science infrastructures are no longer constrained by national security mandates. These infrastructures must now compete in a global scientific market with increased mobility, transparency, and competition. As such, they are often in positions where they need to justify their utility and efficiency across diverse scientific communities and policymakers (Hallonsten, 2014; Heidler and Hallonsten, 2015).

### *5.3.2 Impact Assessment of Big Science*

The previously described changes have transformed the political context in which big science operates. An important early figure looking into the new challenges faced by big science was Alvin Weinberg, director for the Oak Ridge National Laboratory, where uranium was enriched for the atomic bomb in its early years. In his important articles in *Science* (1961) and *Minerva* (1964 and 1963), he voiced his concerns that big science had become a bloated self-serving institution of bureaucracy and complacency, disconnected from more basic human and social needs (Crease et al., 2016). At the same time, the softening of geopolitical ethos did not free big science from excessive political influence (Hellström and Jacob, 2012; Weinberg, 1964, 1963, 1961). In contrast, since public budgets require substantial political support, there were concerns that champions may be tempted to sell and defend their visions with a certain level of sensationalism (Scudellari, 2017). Moreover, there were worries that the business of blockbuster science could undermine the more serious and less sensational work of normal science (Hellström and Jacob, 2012). Weinberg then wanted to establish some criteria for which investments in big science could be evaluated against alternative social priorities (Hellström and Jacob 2012).

An obvious point of departure is to evaluate the scientific productivity levels of big science infrastructures, which are typically evidenced through citation and patent counts. While quantitative evaluation of these measures is easy, they are also considered very imperfect proxies of scientific value, as well as poor indicators of the many peripheral benefits of big science infrastructures (OECD, 2003; Schopper, 2016). As an example, Bianco et al. (2017) argue that the International Space Station, which has cost over $100 billion to build and $2 billion a year to operate, has, as of 2017, only produced 34 refereed articles and 4 patents. Given their long cycle times, publication and patent counts favor more mature infrastructures and are often used as post hoc justifications of sunk-cost investments.

Broadening the scope beyond scientific impact, the normal focus for researchers attempting to evaluate the value of big scientific research infrastructures is on the impacts of direct spending on high-tech procurement with subsequent multiplier effects (Autio et al., 2003; Castelnovo et al., 2018). For instance, aggregating numerous studies of CERN, Schopper (2016) estimates that for every euro spent on high-tech products, an additional 4.2 euros are generated in supporting industries. Beyond the impacts on immediate suppliers, another narrative used to justify investments in scientific research infrastructures are technology spinoffs, with their corresponding or assumed economic growth, job creation, and tax revenue (Aschhoff and Sofka, 2009). Here, NASA may be the most prolific example, boasting over 2,000 spinoffs since 19764 (NASA Spinoff)[13]. Like the early cyclotrons at

[13] https://spinoff.nasa.gov/database/

Berkeley, the value of spinoffs is that they often commercialize technologies in applications outside of a laboratory's principal scientific purview, demonstrating how major research infrastructures can generate impacts beneficial to society without detriment to its main mission (OECD, 2014).

An important characteristic of technology spinoffs as a metric of social value is that the benefits are assumed to accrue to society well beyond the immediate scientific community, and this assumption is important in justifying the investments to taxpayers. However, estimating the indirect, or even direct, economic impacts becomes even more problematic when the technological derivatives are not protected by patents, trademarks, or citations (Schopper, 2016), as is often the case. Given that the political mandate of many research infrastructures is to generate scientific knowledge towards greater social value (Hammett, 1941), the decision not to protect technologies with property rights is frequent and explicit. These practices are consistent with the ethos of open science and open innovation movements (Chesbrough, 2003; European Commission, 2016a), as well as specific mandates from funding agencies to make publicly funded research data accessible, with research results published in open access platforms and FAIR data principles (European Commission, 2012). The most famous and recent case was the World Wide Web (specifically, HTTP, URL, HTML), i.e., when Tim Berners-Lee convinced CERN's managers in 1993 to place it in the public domain and make the IP freely available to everyone. By accepting this case, CERN effectively agreed not to draw revenues or economic value from it. In the words of a CERN senior scientific officer: 'In the case of a conflict between revenue generation and dissemination, dissemination takes precedence' (World Intellectual Property Organization, 2010). For a technology with this level of impact, any quantification of its socio-economic value almost approaches the surreal.

Researchers have attempted to derive more holistic models by conceptually defining the alternative social benefits of research infrastructures (Autio et al., 1996). For example, Florio et al. (2016) derive a model that is based on the following six main dimensions: 1) impact on firms due to technological spillovers produced by access to new knowledge and learning by doing; 2) benefits to employees and students through increases in human capital; 3) the social value of scientific publications for scientists; 4) cultural benefits through outreach activities; 5) additional services provided to consumers; and 6) the value of the scientific discovery.

An earlier, complementary perspective was offered by Autio et al. (2004) who derived a number of propositions related to the positive value that a big science infrastructure can have on its ecosystem of suppliers. These include pushing the frontiers of technology and engineering standards, reducing uncertainty surrounding standards and technology investments, sharing their capacity to manage highly complex projects, aggregating highly diverse and specialized knowledge domains towards radical learning and novel combinations, access to international networks, prestige and reputation, network formation, an exceptional scale and a scope that supports extreme prototyping and testing.

Overall, the indicators are not perfect in terms of assessing the impacts of research infrastructures since they can be insufficient proxies of what they are measuring (e.g., citations), suffer from time-lag effects (Schopper, 2016), and can be myopic in capturing the value provided (spillover effects, human capital formation, or cultural value). As argued in Boisot et al. (2011), the more that a research infrastructure deals with fundamental research, the greater the uncertainty surrounding the future value of the outputs. The lack of reliable data, or well-understood causality, means that more holistic conceptualizations are excessively difficult to quantify and can lead to politically oriented narratives.

In summary, the previous discussion leads to the following conclusions: For research at the forefront of science, a variety of big science organizations have been created with facilities, infrastructures, and instrumentation with unprecedented technical sophistication. With questions on how limited public resources are allocated, concerns have arisen on the social and economic value of big science and how to effectively measure these impacts. Despite these worries, big science infrastructures have a consistent track record in terms of finding alternative applications for their technologies that have tangible impacts on society. While it is common for big science to find serendipitous value in areas previously unanticipated, there is a limited amount of rigorous empirical research on the nature of serendipity and how it can be proactively cultivated. We, therefore, review the literature on serendipity and its mechanisms in the following section.

## 5.4 Serendipity

Serendipity refers to a broad, multifaceted phenomenon related to the unanticipated discovery of something beneficial. As it has been used in various contexts, we trace its various conceptualizations over time. Moreover, we describe the current understanding of how serendipity can be fostered.

### *5.4.1 Definitions and Typologies of Serendipity*

The term serendipity was coined by writer Horace Walpole in 1754, who was inspired by the Persian fairy tale, *Three Princes of Serendip* (Cunha et al., 2010; Rosenman, 2001). He refers to serendipity as an unexpected discovery found from the combination of accident and sagacity (Rosenman, 2001). Sagacity refers to having perception and sound judgment, or in other words, a prepared mind. As such, instead of being merely interchangeable with the words luck, happenstance or providence, serendipity is better seen as a capability requiring the focus of attention (de Rond, 2014). An equivalent formulation can be seen in the context of entrepreneurial opportunity, where serendipity has been seen as the combination of directed search, favorable accidents and prior knowledge (Dew, 2009). By stripping away the random and sometimes mystical aspects of serendipity, it becomes a concept that can be subject to rigorous evaluation, allowing an examination of its triggers, antecedents and mechanisms.

A methodical attempt to understand serendipity was initiated by Robert Merton in the 1950s, which eventually resulted in a book dedicated to serendipity in 2004 (Merton and Barber, 2004). Yaqub (2018) conducted a systematic review of Merton's archives to identify four specific archetypes of serendipity. Mainly focusing on scientific discoveries, he organizes these according to a) whether there is a targeted line of inquiry; and b) the type of solution discovered. Yaqub (2018) defines *Walpolian* serendipity as a targeted line of inquiry that leads to discoveries that researchers were not in search of (solution to a different problem). *Mertonian* serendipity happens where the desired solution is achieved via an unexpected route (targeted problem – different path). *Bushian* serendipity is where untargeted exploratory research leads to a solution for a well-known problem. Finally, *Stephanian* serendipity is where untargeted research finds an unsought solution that may find a future application.

However, even earlier than Yaqub (2018), de Rond (2014) describes a different framework for the structure of serendipity. While he also organizes serendipity in a 2x2 matrix, he divides it differently according to a) whether the solution was the intended target and b) whether the original research design was causal to the solution. In his work, de Rond evokes the term *pseudo-serendipity* to describe when the solutions are intended in the first place, compared to (only) serendipity, where the solutions are completely unanticipated.

One key difference between the two is that de Rond (2014) already assumes that there is an intended target for serendipity to occur, while Yaqub (2018) also permits untargeted search in his framework. Nonetheless, we can see some equivalence between their categories. For instance, while not exactly the same, pseudo-serendipity corresponds to the Mertonian formulation of serendipity, while de Rond's serendipity is equivalent to the Walpolian formulation. Another difference is that whether the discovered solution is a consequence of random variation or deliberate design is not adequately captured by Yaqub's recent typology.

The role of design in serendipity is further emphasized in the work of Garud et al (2018). Taking insights from the evolutionary biology literature, they introduce the term 'exaptation' to the innovation literature to refer to the "emergence of functionalities for scientific discoveries that were unanticipated ex-ante." They identify two forms of exaptation, as follows: *franklins* and *miltons*. *Franklins* refer to the supplementary usage of existing structures in areas in which they were not originally intended for use (e.g., using coins as screwdrivers). *Miltons* refer to discoveries without a currently known function. A widely known image to illustrate miltons is that of spandrels, i.e., the triangular space unintentionally created by the shape of arches, which were later used as a blank canvas for painting (Bahar, 2018).

In contrast to the previous formulations of serendipity, Fink et al. (2017) propose another perspective altogether, with is based on the crossovers of interdependent components. In an experimental study, they show that components early on do not have much benefit, as their utility depends on the existence of other components. However, as the innovation process continues and other components appear, the potential of this original set of components can suddenly manifest. This moment, which seems to come out of nowhere, is what is perceived as serendipity. Accordingly, they explain that serendipity is not only a matter of happenstance but is a result of the components' delayed fruition, which occurs from the existence of other important components.

Finally, it is also important to note another field where the term serendipity has also gained ground, as it gives insights into what differentiates serendipity from other similar concepts. In the field of information systems, serendipity has become an important metric in recommender systems (Kotkov et al., 2016). Recommender systems seek to predict what rating a user would give to a certain product so that new products can be recommended. These systems have been the backbones powering widely used services such as Netflix, Spotify, and YouTube. In such systems, serendipity means that users do not only receive results that are relevant but results that are significantly different from the user's previously rated items. This component of surprise is what seems to define serendipity in this context.

### *5.4.2 Realizing Serendipity*

Aside from attempting to find better definitions of serendipity and understanding its nature, there has also been much progress made on the various factors or mechanisms that can lead to serendipity. McCay-Peet and Toms (2015) propose a process model for how individuals discover and perceive serendipitous events. Their model consists of the following components: Trigger, Connection, Follow-up, Valuable Outcome and an Unexpected Thread. The trigger refers to environmental cues sparking the interest of the individual. This trigger is then connected by the individual to their previous knowledge and experiences. Individuals then follow-up on these triggers to obtain a valuable outcome. The surprise occurs from noticing the unexpected thread present from the previous processes.

The conditions that promote serendipity have also been explored. For instance, the strategies that individuals can pursue to increase the likelihood of serendipity include "varying their routines, being observant, making mental space, relaxing their boundaries, drawing on previous experiences, looking for patterns and seizing opportunities" (Makri et al., 2014). Yaqub (2018) also describes four mechanisms that we summarize as (1) examining deviations from theory, (2) activating previously acquired knowledge and experiences from individuals, (3) tolerating errors and following up on such occurrences, and (4) leveraging networks. In the organizational context, Cunha et al. (2010) identify some conditions related to serendipity, including boundary spanning, mindfulness, social networks, teamwork, free space for creativity and opportunities for playing with ideas.

Artificial intelligence has also been used to find novel solutions to various challenges. Computational methods can aid in the search for interesting information, enabling the discovery of new knowledge domains that have been previously unexplored (Arvo, 1999; Beale, 2007). In drug discovery, for instance, it has been used to repurpose drugs to new therapeutic areas (Mak and Pichika, 2019). As progress in the field increases, artificial systems that "catalyze, evaluate and leverage serendipitous occurrences themselves" are also increasingly explored (Corneli et al., 2014).

While serendipity at the personal and organizational level has been emerging, the literature on how serendipity can be actively pursued at a macro-level is still limited. Garud et al. (2018) describe arrangements to induce exaptation of science, as follows: exaptive pools, exaptive events, and exaptive forums. Exaptive pools refer to the maintenance of scientific discoveries such as through patent and publication databases. These ideas, however, remain decoupled until they are activated by exaptive events, such as technology fairs and workshops. These possibilities can be further developed and contextualized through exaptive forums, where actors become increasingly entangled.

In summary, the extant literature on serendipity has mostly been speculative or based upon small-sample, anecdotal examples of scientific discoveries. Moreover, the previous studies mainly focus on the individual scientists, lacking understanding of how serendipity can be induced at a more macro-level. As such, questions remain on how serendipity can be cultivated towards finding market applications for science and how it can be cultivated, for instance, with the help of policy. To move the serendipity literature forward, there is a need for studies based on empirical evidence, preferably using quasi-experimental conditions. By examining the novel policy response ATTRACT, this study puts forward a rigorous empirical examination of serendipity.

## 5.5 ATTRACT

The ATTRACT project is a €20M-funded initiative within the Horizon 2020 Framework Program that aims to systematize the discovery of breakthrough applications of research from the leading European big science infrastructures. In the following section, we describe its underlying philosophy, aims and results to date.

### *5.5.1 Philosophy*

The assertion that the products of scientific research centers can have value outside of their intended scientific purview is not new.[14] It was demonstrated clearly by Lawrence's early cyclotrons in oncology, and the idea was perhaps best institutionalized as an important policy driver by Bush, who advocated large investments in untargeted scientific research as a source of serendipitous discoveries or solutions (Bush, 1945; Yaqub, 2018). In a more liberal interpretation, the Bush legacy favors large investments in research for its unknowable scientific value, as well as numerous unknown benefits that accrue as socio-economic derivatives (education, spin-offs, job creation, etc.)

During the last three decades, policy-makers have increasingly emphasized policies to accelerate innovation and economic growth (Edler and Fagerberg, 2017). Three main types of approaches have been developed. The *mission-oriented* approach aims to support solutions to challenges that are part of an explicit political agenda. Here, policy-makers tend to anchor innovation policies in grand societal challenges, such as national defense, climate change, or other sustainable development goals (Galison, 2016; Galison and Hevly, 1992; Mazzucato, 2016; Mazzucato and Semieniuk, 2017; Mowery, 2012). *Invention-oriented* approaches aim to stimulate the supply of inventions as derivatives of scientific discovery while leaving any commercial exploitation to the market (Bush, 1945; Wilson, 1991). This was the most widely adopted approach championed post-war by Bush, as policy-makers sought to advance science and technology as broad drivers of geopolitical policy (Galison, 2016; Galison and Hevly, 1992). Finally, recent decades have seen *system-oriented* approaches that seek to foster interactions among the different actors taking part in the innovation ecosystem (Borrás and Laatsit, 2019; Lundvall, 2010; Lundvall and Borrás, 2009).

Within these main orientations, a wide range of policy instruments have been deployed in Europe to stimulate innovation (European Commission 2016), and different typologies have been suggested to understand them (e.g., Borrás and Edquist, 2013; Edler and Georghiou, 2007). The most widely accepted view considers instruments such as those focusing either on technology push or market pull. Technology push (supply-side) policies stimulate framework conditions and opportunities for innovation to thrive, including measures to support R&D collaboration, network formation, and incentives to attract highly skilled labor to focal regions and sectors. For example, in Europe, the Future and Emerging Technologies (FET) program has allocated €2.7 billion to pursue breakthrough ideas through unexplored

[14] Detailed information can be found at https://attract-eu.com.

collaborations of multidisciplinary scientific and cutting-edge engineering teams, which is indicative of the invention-oriented approach mentioned earlier.

Market pull (demand-side) interventions have been emphasized with greater frequency in the most recent literature (Edler and Georghiou, 2007; European Commission, 2016b; Rolfstam, 2009). This perception recognizes that the derivatives of basic scientific research have limited value if specific market-pull mechanisms are not in place to facilitate their entry to the market (Scherer, 1982; Schmookler, 1962). Demand-side policy instruments include measures to foster investments by private capital (brokering, tech-transfer, IP, subsidies, etc.) or, alternatively, pre-commercial procurement to nurture financial liquidity, investment, and operational scale in start-ups and SMEs (Edler and Fagerberg, 2017; Rolfstam, 2009). However, instruments that simultaneously stimulate both the supply-side and demand-side dynamics, especially for early-stage, high-risk technologies, are less common (Cunningham et al., 2013; European Commission, 2016b).

The challenge of bridging the supply and demand sides of the innovation cycle is not an exclusive concern of innovation policies. It is also a well-known challenge in entrepreneurship research, where it is frequently metaphorized as the valley of death (VoD) (Beard et al., 2009; Hudson and Khazragui, 2013). This metaphor describes the difficult phase in product development and commercialization where many viable products or start-ups do not survive for a variety of reasons. Typically, these include excessive and unforeseen costs for research, prototyping, testing and manufacturing, limited product development budgets, ineffective coordination and expertise, sub-critical market exposure, and the inability to obtain sufficient internal or external funding to bring the product or start-up to a revenue-generating state (Frank et al., 1996).

A substantial amount of research has focused on the various mechanisms that can be marshaled towards mitigating the VoD phenomenon, which include the following: innovation intermediaries (Islam, 2017); scientific parks; technology clusters and living labs (Almirall and Wareham, 2011); industry associations (Markham et al., 2010); business incubators and accelerators; technology brokers and tech-transfer functions (Beard et al., 2009); regional, national, and pan-national funding instruments, such as Horizon 2020, EIT and ERC of Europe, and NIH, NSF of the US (Hudson and Khazragui, 2013). Finally, particularly in the medical and life sciences fields, there has been a growth in initiatives in translational research (Butler, 2008). No single VoD scenario is applicable to all technologies. For technologies with high technology readiness levels (TRL) (Banke, 2010), the VoD is potentially less fatal, particularly for incremental innovations with probable market uptake. This condition is typically addressed by risk mitigation functions performed by private investment and venture capital. However, technologies with low TRLs require more extensive interventions, typically with both risk absorption (seed funding and early industry involvement) and risk mitigation (public/private investment mechanisms). It is important to note that TRLs are highly context dependent; i.e., the technology may be very mature and tested in its original application at the scientific research installation (high TRL), but immature in a larger system of commercialization when used in a different sector or market (low TRL) (Héder, 2017).

### *5.5.2 Purpose, design and results to date*

The main aim of ATTRACT is to harness and direct exploration towards breakthrough innovation opportunities in detection and imaging technologies, while also offering space for serendipity to stumble onto unforeseen applications. As such, there are no 'intended' technological applications or desired outcomes. Rather, the ATTRACT governance is designed to generate as many options and variety in the applications as possible. That acknowledged, there are some obvious areas where detection and imaging technologies can be employed towards substantial, if not paradigmatic, advances in other domains. Frost and Sullivan argue that imaging and detection technologies will have core functions in almost all technically sophisticated commercial products and will constitute an annual market of over $100 billion in their own right (Frost & Sullivan, 2015). These domains include medical device and imaging technology, biotechnology, energy, advanced manufacturing, automation, microelectronics, materials and coatings, environment and sustainability, and information and communication technology.

On many dimensions, ATTRACT has been designed to directly address the ineffectual transition – or disconnection – between the technology-push instruments (applied in the early phases) and the market-pull instruments (the later entry of private capital) (Auerswald and Branscomb, 2003; Wolfe et al., 2014). In this respect, ATTRACT is distinctive from recent instruments, such as FET, given that the focal actors include both research infrastructures and industrial players, and equal protagonism is given to both the supply and demand sides. This is enabled by the pre-existing relationships between research infrastructures and their industrial suppliers; that is, the highly specialized SMEs that have contributed to the engineering, construction, and operation of some of the world's most sophisticated technologies. Thus, the industrial relevance and operational feasibility of the projects are verified from the start. Specifically, for projects involving European research facilities and industrial organizations, the most immediate use of their technologies is guaranteed. In this sense, a first 'internal market' is assured. This 'internal market' paves the way for industry to target other applications and new commercial opportunities (i.e., the feasibility of the pilot technologies has been prototyped and tested in the real and demanding working conditions of big science facilities).

The completion of ATTRACT phase I is expected to lead to insights and findings that inform modifications and extensions to the design of ATTRACT phase II and related innovation policy initiatives. ATTRACT phase II will aim to take a select group of 10-20 validated projects from ATTRACT phase I and scale them towards technology readiness levels 5-8. ATTRACT phase II is specifically designed to address the intermediate or secondary phases of the valley of death phenomenon, which requires greater scalability, maturity, and support. Funding for ATTRACT phase II is currently being negotiated with the relevant funding bodies and is subject to receiving grants. However, the current estimates suggest a total funding of €35 million. In addition, emphasis will be placed on the transition to public sources of equity-based capital (e.g., the European Investment Fund and the European Investment Bank), as well as private capital sources, such as early and late-stage venture capital and private equity.

Table 1 highlights the main attributes of ATTRACT and how they are positioned relative to traditional EU funding instruments and private capital investments.

*Table 1. Comparison between ATTRACT and other funding instruments*

| | **ATTRACT** | **EU range public funding instruments**[1] | **Private instrument** |
|---|---|---|---|
| **Approach for crossing the valley of death** | Considers that breakthrough technologies need two steps of risk absorption and risk mitigation. | Assumes that only one step is needed – normally risk mitigation (projects are funded on equal footing).[2] | Focuses on relatively low-risk technologies with no need for risk absorption. |
| **Risk absorption (reduce large TRL gap)** | Public seed funding to foster ideas with breakthrough potential (100k EUR). ATTRACT2 aims to continue with public scale funding for selected projects (2-4M EUR). | | |
| **Risk mitigation (close TRL gap)** | Public/private investment mechanisms. [3] | Public/private investment mechanisms. | Angel, Venture capital funding. |
| **Pre-competitive market** | Ensured in projects with participation of research infrastructures. | Not ensured and depending on a project-by-project case. | Not ensured. |
| **Scaling up** | Late-stage VC funding instruments, private equity, IPOs, etc. | | |

[1] We are referring to EU funding programs such as Horizon 2020. We do not consider national public funding programs.

[2] Exceptions exist, such as the SME instrument https://ec.europa.eu/programmes/horizon2020/en/h2020-section/smeinstrument. Nevertheless, they differ from ATTRACT because a project needs to apply for seed funding, and subsequently, for scale funding. In ATTRACT, the transition between seed and scale is streamlined.

[3] http://www.eif.org/; http://www.eib.org/en/index.htm

As of the writing of this paper, ATTRACT has implemented the following steps:

1. An open call was launched to solicit project proposals (1,211 submitted) for leveraging detection and imaging technologies towards potentially commercially sustainable products or services. While not exclusive, the emphasis was on concepts at technology readiness levels 2-4. The call solicited proposals leveraging the following four main technology groups: a) sensors; b) data acquisition systems and computing; c) software and integration; and d) front- and back-end electronics.
2. All submissions were assessed on technical merit and innovation-potential. Specifically, the evaluation dimensions included the project definition, scope, and technological feasibility, state-of-the-art, scientific/engineering merit, industrial potential, commercial feasibility, and social value.
3. 170 projects were awarded €100,000 for the development of a proof-of-concept or prototype with an application outside of the original purview of the technology, over a period of one year.

### *5.5.3 The 170 Funded ATTRACT Projects*

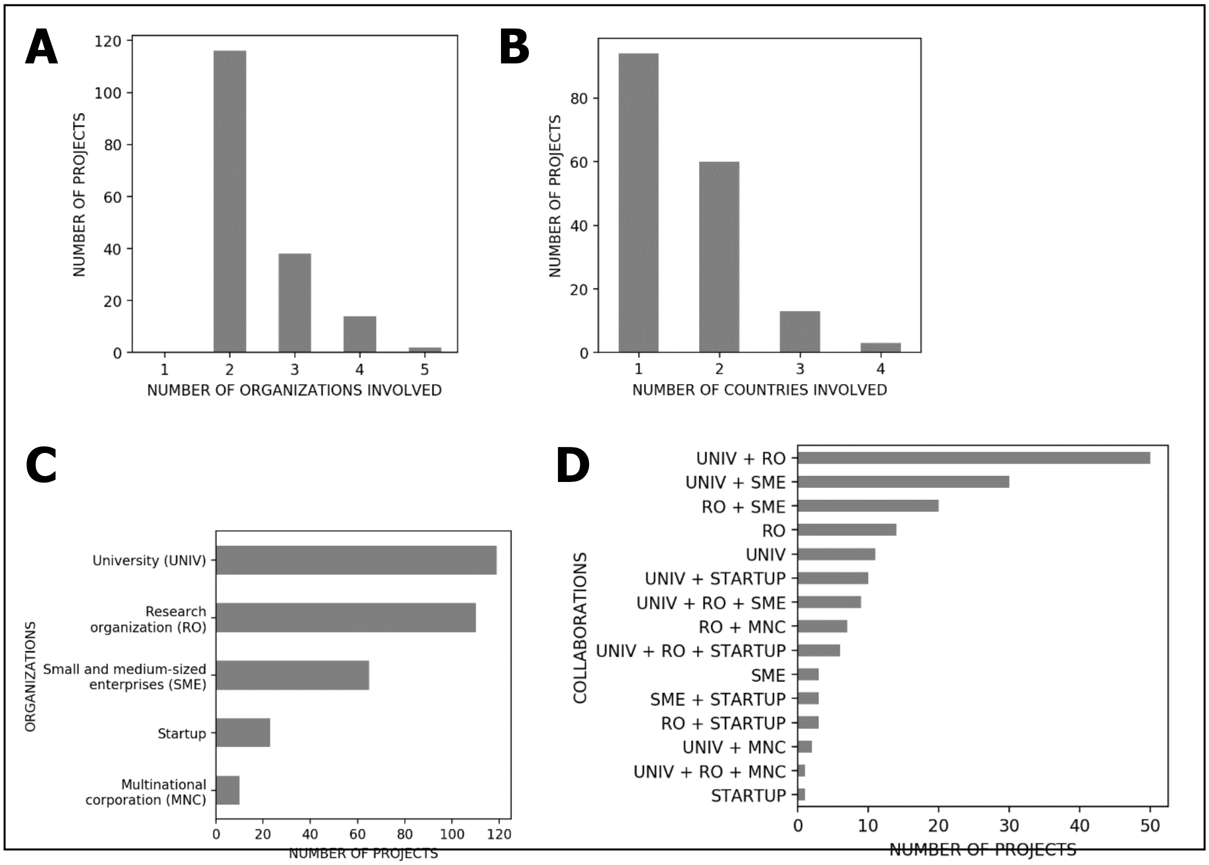
The call was open from 1 August to 31 October 2018. In that period, 1,211 proposals were received. The top 10 countries submitting applications were as follows: Italy (261); Spain (230); Switzerland (108); France (96); the United Kingdom (81); Germany (67); Finland (65); the Netherlands (59); Portugal (33); and Austria (26). From these submissions, 170 projects were selected for funding.

To analyze these different projects, we carried out the following: We collected the text proposal of the 170 funded projects for analysis. Each proposal submitted contained a maximum of 3,000 words, including the following parts: a) summary; b) project description; c) technology description and external benchmarks; d) envisioned innovation potential (scientific and/or industrial), as well as envisioned social value; e) project implementation, budget, deliverables, and dissemination plan. The proposals of these 170 funded projects were read by the authors and three master's students for evaluation.

Three master's students with backgrounds in biomedical engineering, mechanical engineering/physics and entrepreneurship evaluated each project independently. They coded for the following project characteristics: technology readiness level (scale of 1 to 9), scope of market application (specific, specific but easily expandable, or general), location in the value chain (upstream or downstream), technology novelty (scale of 1 to 5), technology relevance to the market (scale of 1 to 5) and credibility of budget and milestones (scale of 1 to 5). The variables were used based on extant definitions in the literature (i.e. TRL and MRL). In the event that there were no extant definitions, new categories were induced from the phenomenon (serendipity). After analyzing each project separately by the three independent coders, their findings were integrated. In cases where the codes were not consistent, discussions were held to reach agreement. The coding was then validated in an additional round of coding by the authors and then tabulated. As such, each project was evaluated and coded by a minimum of three independent evaluators. Three physicists (two co-author of the study) and a venture capital expert oversaw the coding process and validate results.  The results are presented in the following paragraphs.

The ATTRACT project call required the participation of a minimum of two collaborating organizations. While the majority of projects were the result of two organizations collaborating, as many as five organizations can be seen collaborating in a single project (Figure 1A).

*Figure 1. Summary of Organizations Involved in ATTRACT projects.*



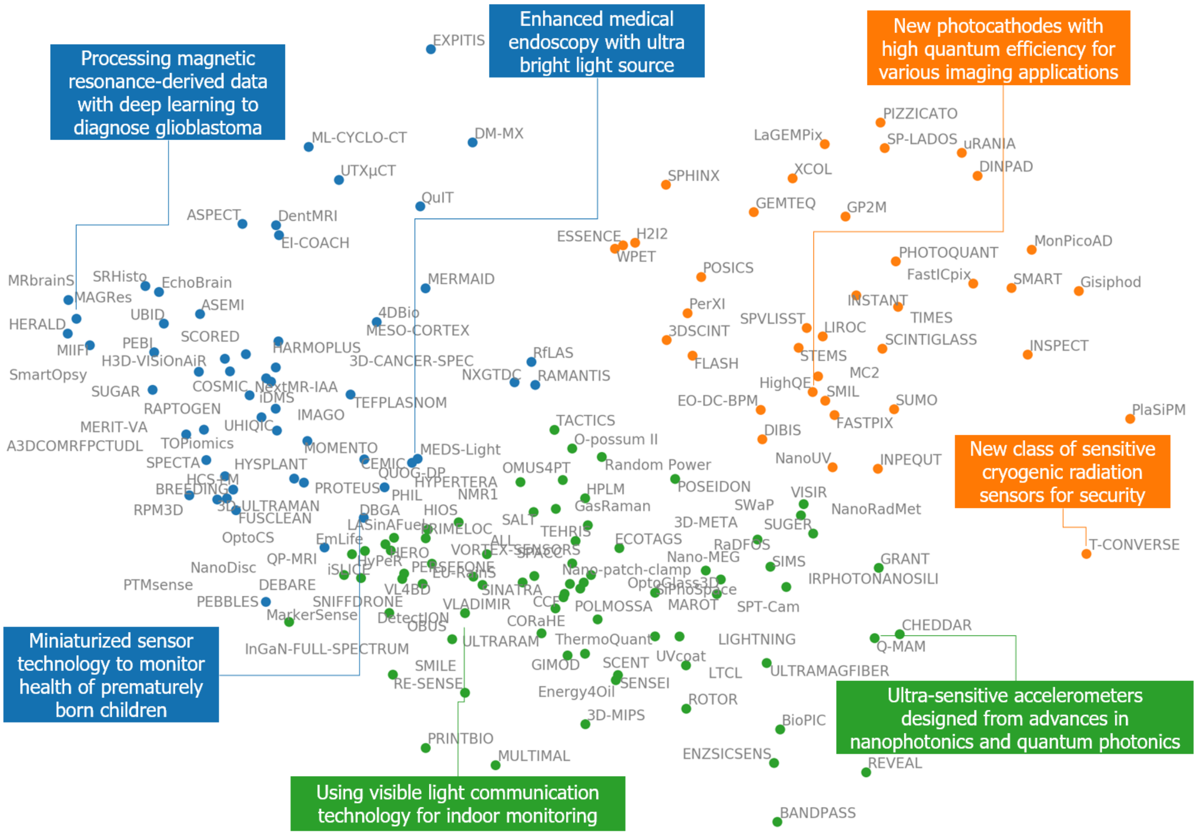
A shows the number of organizations collaborating across projects. B shows the number of countries collaborating per project. C shows the types of organizations involved across all ATTRACT projects. D demonstrates the various combinations of organizations collaborating in a project

Exploring the countries represented in each project funded in ATTRACT, Figure 1B shows that the majority of projects involve collaborations between organizations located in the same country. Such arrangements allow the partners to closely interact and meet frequently as they work on their projects. Interestingly, almost half of the projects (45%) involve international collaboration. Especially when projects require highly specialized, scarcely available expertise among partners, it is necessary for collaborations to occur across borders.

As seen in Figure 1C, the majority of projects involve research organizations (ROs) or universities. Aligning with the goals of ATTRACT, many projects also involve input from industrial partners, including startups, small and medium-sized enterprises (SMEs) or multinational corporations (MNCs). The most represented configuration involves collaborations between universities and research organizations (Figure 1D). These research organizations typically have expertise in spinning out technologies. Aside from this configuration, industry-academia collaborations are extremely common, most notably between universities and SMEs and ROs and SMEs.

We visualize the 170 projects in Figure 2 through automated processing of the textual data from the proposals. As the showcased projects reveal, ATTRACT covers wide ground in the domains of technologies sourced and targeted application areas. There is a huge cluster of projects applying big science research to impact the field of healthcare, such as through better diagnostics and treatments (blue cluster). Aside from this cluster, there are many more projects applying the imaging and detection technologies of big science to various commercial applications, such as consumer electronics, environmental monitoring, and security (green cluster). Finally, we see efforts to further improve the technologies themselves, with the immediate market of serving the big science infrastructures (orange cluster).

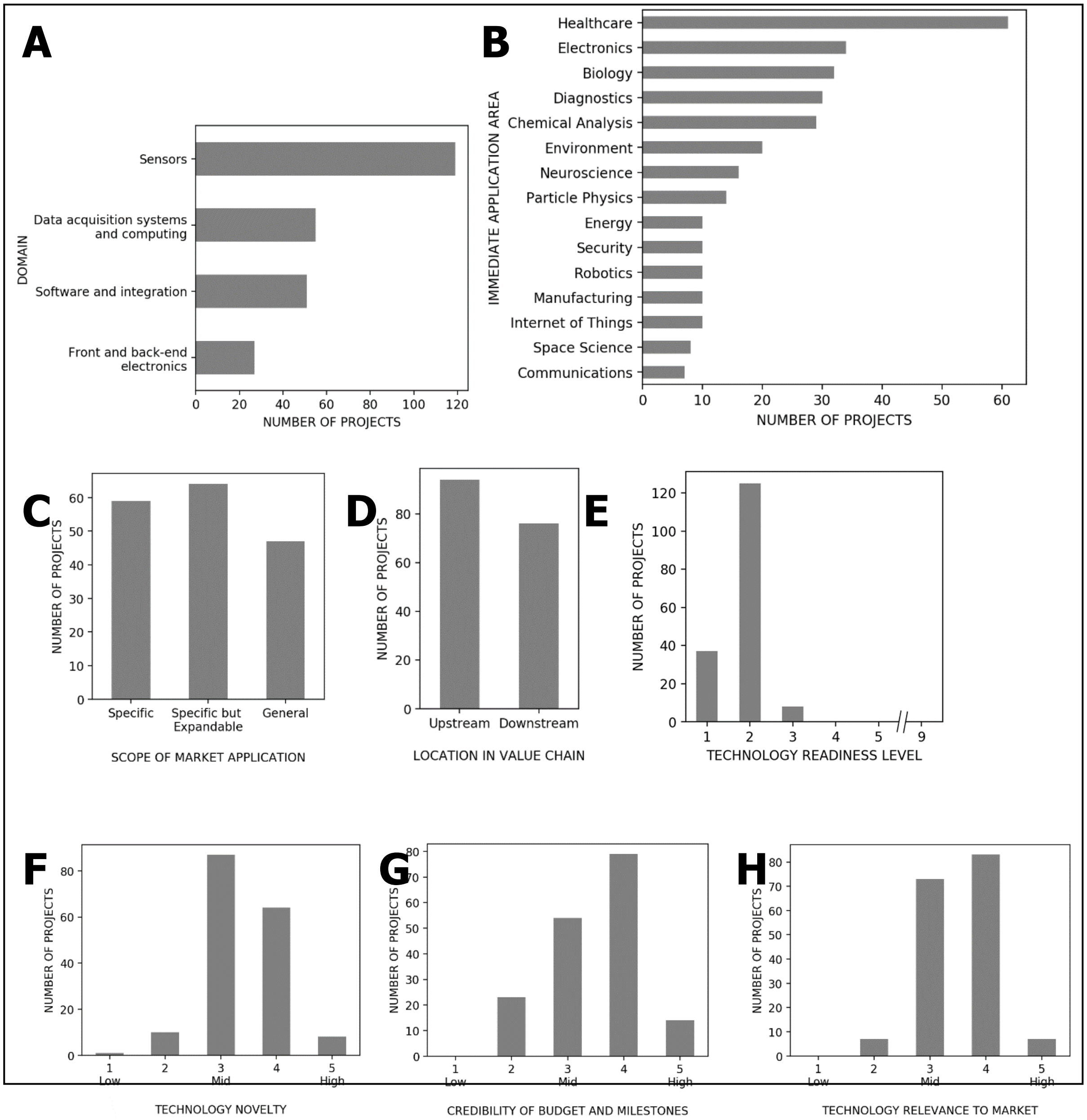
*Figure 2. Visualization of the 170 Funded Projects under ATTRACT.*



Each project is labeled by its acronym. The projects are plotted by processing their textual data (removal of stop words, lemmatization, inclusion of n-grams), performing TF-IDF vectorization and decomposing by PCA into two components. The colors were generated by K-Means clustering. The blue cluster refers to projects in healthcare. The green cluster refers to applications of detectors to various areas. The orange cluster refers to upstream advances in sensor technologies. The code will be available online.

The automated classification was, however, not adequate to fully understand the projects included within ATTRACT. We, thus, conduct further analyses by manually evaluating the textual data of the projects. Figure 3A shows the different technological domains as submitted the participants, which are as follows: sensors (70%), data-acquisition systems and computing (32%), software and integration (30%) and front and back-end electronics (16%). Note that the projects can belong to more than one domain so they do not add up to exactly 100%. As observed, a large percentage of projects are in the domain of sensors. This percentage is not unexpected, as big science infrastructures are generally known for the sophistication of their imaging and detection technologies. The high expertise of these groups in sensor technology, together with the versatility of sensors towards various uses, make them good candidates for exploring alternative commercial applications.

*Figure 3. Summary of the Various Coded Dimensions of the ATTRACT Projects.*



A shows the domains of the projects, as stated in their proposals. B shows the application areas, as coded from analyzing the text. C describes the scope of the market application for each project. D shows whether the application area is upstream or downstream. E describes our rating on the relevance of the proposed technology to the selected market. F describes our evaluation of the technology readiness level of each project. 2G describes our evaluation of the novelty of the technology. 2H shows the evaluation of the credibility of the proposed budget and milestones of each project.

Further analysis was carried out to describe the different features of the funded projects under ATTRACT (Figure 3). Figure 3B shows that ATTRACT caters to a diverse range of application areas, including healthcare (36%), electronics (20%), environment (12%), energy (6%), security (6%) and manufacturing (6%). These projects commit to these areas in varying degrees. Figure 3C shows that the projects are almost equally split in terms of the degree of specificity in the application area. While 35% of the projects are specific to their mentioned application area, there are also a large number of projects offering a general solution to different application areas (28%). An interesting category is the 38% that are specific but expandable projects that have already identified their pilot market but then can easily extend their reach to other areas. Furthermore, Figure 3D shows that there are slightly more projects located upstream in the value chain. These upstream projects (55%) aim to supply companies with knowledge and technologies that can be further processed and integrated towards their offerings. In contrast, downstream projects (45%) cater directly towards solving the problems of its intended market.

Figure 3E shows that the most common technology readiness level was 2, meaning that the projects are only in the stage where the technology and/or application area has been conceptualized. The average TRL across all projects was 1.8. These low TRL values are in line with what was expected from the projects during the proposal call. The low TRLs show that these technologies are still in their early stages, requiring further development towards becoming viable solutions. Their low TRLs have the benefit, however, of giving them the flexibility to find the serendipitous area where their application will have the most impact.

Originating from the leading big science infrastructures, the projects feature some of the most advanced, cutting-edge technologies. Figure 3F shows that the projects are highly novel, with an average rating of 3.4 out of 5. The problem typically with technologies that are too novel is finding areas that would be relevant for their application. However, as seen in Figure 3G, the projects have generally high relevance to the markets they are hoping to serve. Across all projects, the average rating was 3.5 out of 5. This rating implies that a project such as ATTRACT can help activate researchers to find relevant applications for the technologies they are working on. Otherwise, for projects lower in rating, the support provided by ATTRACT enables these projects to refine their technologies to find a better fit with their market of choice or to find a more applicable market to which their solutions can be of value.

To systematically explore the space in the development of their technologies, it is important for the project's team to have a credible plan and list of milestones. Figure 3H shows that the projects were rated highly on this aspect, with an average rating of 3.5 out of 5.

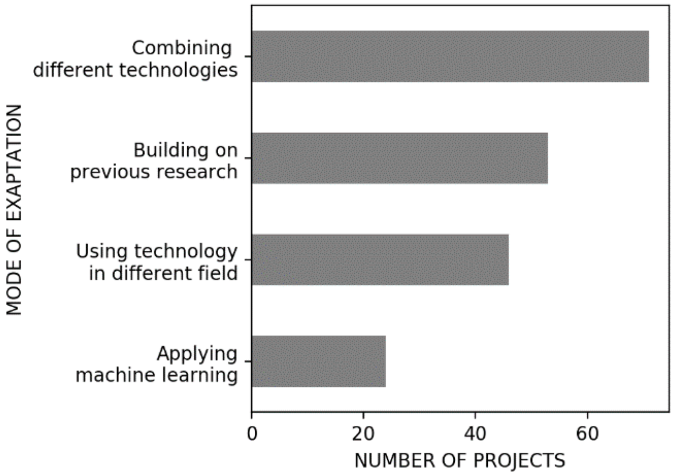
### *5.5.4 Modes towards Serendipity*

In the project text, the researchers typically narrate the mode by which they were able to develop new applications for their scientific research. We identified the recurrent themes by which serendipitous discoveries were actively pursued by project members in our first read through the 170 projects. In the second and third readings, we categorized the projects according to the following criteria:

- Combination of different technologies – technologies or knowledge from different research domains is combined, integrated or assembled together to produce a new application.
- Building on previous research – technologies from previous research work are extended or improved to be more effective or efficient but are still within the same domain or application area.
- Applying technology to another field – technology or knowledge from one domain is used in a new research domain or application area.
- Using machine learning or artificial intelligence – when the computational advances in machine learning or artificial intelligence are used to augment or find new uses for existing technologies.

Note that the projects typically combine these modes to different degrees and so, we coded them according to what is explicitly mentioned in the text. The number of projects in each category is summarized in Figure 4.

*Figure 4. Modes towards Serendipity in the ATTRACT projects.*



### *5.5.4.1 Combination of different technologies*

The most represented mode was the combination of different technologies (41%). Under this category, technologies could come from adjacent or distant domains. Moreover, these technologies could be combined with varying degrees of integration. On one extreme, we identify a subset of projects (16%) where existing, readily available technologies are assembled to develop a new application. For instance, a project called PHIL, which aims to use a photonic system for liquid biopsy, mentions the following:

*"we will design and build the system using mainly commercial solutions for the different system aspects".*

Otherwise, many projects combine the latest advances from distant research areas to create novel solutions. A notable example is the SCENT project, which aims to create new gas sensors. The project mentions that it is:

*"based on merging two up-to-now disjointed macro-disciplines: high-pressure technology and gas-sensing; whose scientific communities are still far one another: the former focusing mainly on synthesis of materials, the latter unaware of HP-potentialities."*

### *5.5.4.2 Building on previous research*

The second mode we identified is extending and building on previous research (31%). Typically, this mode proceeds from re-examining previous research so that new features that have not been previously identified or explored can surface. Pursuing this re-examination typically requires a meticulous re-examination of previously acquired knowledge and finding new perspectives in the existing data. A notable example is the project Random Power, which is a random bit generator for cryptography. According to their proposal:

*"The genesis of the project is an example of ingenuity and serendipity and can be tracked to the effort of understanding random events affecting the response of state-of-the-art detectors of light with single-photon sensitivity."*

Another way that previous research is reinterpreted is by exaggerating features or taking things to the extreme. For instance, there are many projects that examine what possibilities would be opened if current detectors could be applied at extremely cold temperatures or in environments with very high radiation. Similarly, there are projects that develop new application areas through imagining what opportunities can be created if a technology becomes a magnitude more efficient or powerful.

The previous research can also be extended by projecting from the current state of their research a laudable target. By setting a difficult goal, the researchers then leave it to their abilities and to successful development of the project so that they can bridge the gap between this goal and their current state.

#### *5.5.4.3 Applying technology to another field*

Another set of projects (27%) applied a technology from one field to another field. This category coincides best with the previous notions of serendipity – finding new uses from existing things. By exposing a technology to a field that it has not been previously used for, new use cases for the technology potentially emerge. Especially for the big science institutes in ATTRACT, their technologies might be narrowly used within their scientific domain. These new technologies are also able to provide a fresh perspective to the field, proposing new ways to deal with the problems that the existing technologies currently employed within the field may not adequately address.

A notable example of a project is SIMS, which involves designing a seismic imaging and monitoring system. They mention that they will develop a:

*"next-generation MEMS sensor that utilizes patented technology inspired by the search for gravitational waves."*

#### *5.5.4.4 Using artificial intelligence or machine learning*

The final mode we identified involved the application of machine learning for a specific application, accounting for 14% of the projects. This category can be considered a subset of the previous category since machine learning is a breakthrough originating from the computational sciences that is finding new uses in different domains. By being able to find patterns that humans cannot easily identify, it can be said that applying AI or machine learning increases the efficacy of various sensors in what can be obtained from the data it is able to collect.

Many of the projects in this category are in the field of healthcare. The usage of machine learning allows data collected from the various imaging technologies to be brought together and processed to reveal new insights on certain diseases. For instance, the project MAGres plans to integrate various magnetic resonance techniques to obtain a better understanding of the brain tumor glioblastoma. They mention the following:

*"ML [machine learning] methods are the key to unlock the predictive power from the complex and high-dimensional data to be acquired"*

## 5.6 Discussion

We identify four categories of how big science research can be used in previously unexplored ways towards commercial applications. These four modes towards serendipity are (1) a combination of different technologies, (2) building on previous research, (3) applying technology to another field and (4) using AI or machine learning. Compared to the previous studies of serendipity, the categories we describe do not completely coincide with any one proposed typology of serendipity, as summarized in Table 2.

*Table 2. Contributions to the previous literature on serendipity*

| **Cultivating Serendipity** | **Serendipity viewed from its Outcome** | | | **Other Literature on Serendipity** |
|---|---|---|---|---|
| **Categories from ATTRACT**<br>**(This Article)** | Structure of Serendipity (de Rond, 2014) | Typology of Serendipity (Yaqub 2018) | Exaptation of Science-based Innovation<br>(Garud 2018) | |
| **Applying a technology to another field** | Serendipity by way of random variation | Walpolian Targeted search solves an unexpected problem | Franklin's character was previously shaped for some use but is now coopted for a different role (ex. coin as screwdriver) | |
| | Serendipity as the unintended consequence of design | | | |
| **Building on previous research** | Pseudo-serendipity by way of random variation | Mertonian Targeted search solves problem via an unexpected route | | |
| | Pseudo-serendipity as the unintended consequence of design | | | |
| **Combining together different technologies** | | | | Crossovers between components (Fink et al., 2017) |
| **Applying AI/Machine learning** | | | | Computer-aided serendipity (Arvo, 1999) |
| **Untargeted search (during research before ATTRACT)** | | Bushian Untargeted search solves an immediate problem | | |
| | | Stephanian Untargeted search solves a problem later | Milton's<br>character was not shaped for some use but has the potential to be coopted for another use (ex. spandrels) | |

The category of applying technology from one field to another coincides highly with the previous notions of Walpolian serendipity (Yaqub, 2018) and the idea of exaptation (Garud et al., 2018). These two formulations, on a fundamental level, refer to the unanticipated usage of a certain item. A nuanced difference, however, between these previous notions on serendipity is that our categorization stems from a different view of serendipity, i.e., exploring the modes towards its realization. Instead of characterizing it ex-ante, our category describes the actions that researchers are actually taking in the hopes of finding serendipitous applications for their scientific research.

On the surface, extending the previous research does not seem to be related to serendipity. The implied incremental nature of the progress that comes from building on previous research makes it seem that it is not a viable way to cultivate serendipity. However, as we find in the different projects, extending the previous research can be productive, especially if it allows the accumulated wealth of knowledge and experience of various actors to be activated and re-examined. This productivity coincides with how Cunha et al., (2010) sees serendipity, i.e., as the process of metaphorical association – seeing things in a new way. Such activation facilitates researchers to pursue a laudable target that they have not considered doing before.

Compared to the previous typologies of serendipity, we find two new categories. The first one is the combination of different technologies. This conceptualization of the phenomenon is consistent with that of Fink et al. (2017), which relates serendipity to the surprise from the crossover of interdependent components. On a fundamental level, the innovation research has greatly emphasized the role of combining knowledge from diverse domains to generate breakthrough innovation (e.g., Guan and Yan, 2016; Schoenmakers and Duysters, 2010). Nonetheless, it has not been explicitly linked to serendipity due to the lack of empirical studies on its realization.

Finally, in the ATTRACT projects on AI and machine learning were used to process and make sense of the huge quantities of data generated by the various sensors. These technologies are valuable, as they are able to see subtle patterns that are invisible to the human eye. AI and machine learning improve the performance of certain technologies by being able to process large amounts of data and integrate different sources of information to obtain new insights. However, it is important to make a distinction that AI and machine learning were mainly used to integrate the data resulting from the detectors instead of for discovering new applications. Machine learning was not used on a meta-level to discover new serendipitous applications of the technologies, for instance, from mining text from publications and patents. However, with the ongoing progress in these technologies (as in recommender systems), it would be interesting to see how AI and machine learning can directly be used to generate leads for serendipitous connections between various topics (e.g., Arvo, 1999; Giles and Walkowicz, 2019).

### *5.6.1 Implications for theory*

The research on serendipity has evolved beyond the simple conceptualization as an accident or happenstance. Recent developments have allowed serendipity to be scientifically examined by having reformulated it as a capacity, requiring the focus of attention (de Rond, 2014). This paper validates the previously proposed typologies on serendipity through the unique dataset of ATTRACT. While the previous research on serendipity mainly relied on anecdotal stories in the history of science, ours is grounded on the data from the 170 funded projects under ATTRACT. With these projects spanning different domains and varying in their technological features, this gives us access to a large dataset that we can probe to study how serendipity is actively pursued.

Unlike the previous studies of serendipity, which view the phenomenon after it has already occurred, we provide another perspective by looking at the modes towards its realization. This process-oriented data-driven approach allowed us to find two previously unidentified modes wherein serendipity can be cultivated, as follows: combining technologies and using machine learning. More systematic analyses with other novel datasets are needed to corroborate our findings and identify other means that serendipity can be realized.

### *5.6.2 Implications for policy and practice*

Our paper shows how policy can enable researchers to find alternate serendipitous uses for their technologies. The ATTRACT project is consistent with calls by Mazzucato (2013, 2016, 2017), who argues that the government can go beyond its role as a regulator or fixer of markets towards an entrepreneurial role, absorbing the risks in strategic sectors until technologies have reached a sufficiently mature state to be attractive to private and venture capital. This assumes that market mechanisms and private capital alone may not be the most efficient routes to realizing innovation via basic to applied research (Martin, 2016). Specific industrial policies and stimulus instruments are needed to absorb the risks in basic research settings when working with low TRL technologies. This is particularly relevant to ATTRACT in light of the empirical research suggesting that the more the research infrastructure is involved in basic research as part of its mission, the less likely that the organization will be involved in technology transfer activities (Boisot et al., 2011; Rahm et al., 1988); this is certainly the case for several ATTRACT partners.

ATTRACT also resonates with the ‘cooperative technology’ model of technology transfer described by Bozeman (2000), which assumes that government laboratories and research infrastructures can play an important role in technology innovation and economic growth.

With some variation, authors such as Mazzucato and Bozeman echo the original doctrine of Vannevar Bush, i.e., that basic research has a substantial and positive impact on socio-economic innovation via direct and indirect mechanisms. Interestingly, however, the recent literature has argued that while it is commonly believed that Bush maintained an unquestioning faith in an integrated and linear model of innovation, his notion was more sophisticated and involved symbiotic cross-fertilization (Leyden and Menter, 2018). In this view, the authors argue that while Bush saw that basic research and applied research benefit each other, they also succeed by working as separate systems, or stacks. Consequently, scientific and economic policy mechanisms should seek to coordinate the two systems, allowing each to operate through its own logic and success criteria, yet simultaneously cultivating specific points where they can nurture each other (Cunningham et al., 2013; European Commission, 2016b; Leyden and Menter, 2018). ATTRACT does not presume to be the definitive word on how to accomplish this coordination task. Indeed, faithful to its genesis in scientific institutions, ATTRACT should be seen as an experiment in innovation policy (Bakhshi et al., 2011). With its focus on the revelation of information and cross-fertilization of technology and entrepreneurial options, it is experimental at an operational level. With its novel constellation of actors, resources, design, and governance, ATTRACT is very much an experiment in innovation policy.

## 5.7 Conclusions

We have described the ATTRACT project, which is a novel innovation policy instrument to find new applications for the breakthrough imaging, detection, and computational technologies of Europe´s leading scientific research infrastructures.

We have described the philosophy behind the project, discussing the history of big science and the issues with regard to assessing its socioeconomic impact. Where ATTRACT is still in-process, the large data set from the proposals allows us to view serendipity in a unique, unprecedented manner. Specifically, the 170 projects allow us to probe serendipity in a quasi-experimental setting with some controls. We identify several novel modalities of serendipity that emerge from the data.

There are many interesting avenues for future research. First, it is a widely accepted wisdom that increasing the collisions between different actors promotes the chances of serendipity. As such, it is valuable to understand how the various partners working in the projects were able to find each other and create new applications for their previous technologies. Incorporating insights from the alliance and network literature would create new insights in the serendipity literature.

Faithful to its genesis in scientific institutions, ATTRACT is best viewed as a policy experiment. Where a complete evaluation of it will require more time, the initial evidence suggests that policymakers can play a purposeful and effective role in fostering derivative benefits from public investments in big scientific research infrastructures.

## References

Almirall, E., Wareham, J., 2011. Living Labs: Arbiters of midand ground-level innovation. Technol. Anal. Strateg. Manag. https://doi.org/10.1080/09537325.2011.537110

Arvo, J., 1999. Computer aided serendipity: the role of autonomous assistants in problem solving. Proc. 1999 Conf. Graph. interface '99.

Aschhoff, B., Sofka, W., 2009. Innovation on demand-Can public procurement drive market success of innovations? Res. Policy. https://doi.org/10.1016/j.respol.2009.06.011

Auerswald, P.E., Branscomb, L.M., 2003. Valleys of death and Darwinian seas: Financing the invention to innovation transition in the United States, in: Journal of Technology Transfer. https://doi.org/10.1023/A:1024980525678

Autio, E., Hameri, A.-P., Bianchi-Streit, M., 2003. Technology transfer and technological learning through CERN's procurement activity. CERN.

Autio, E., Hameri, A.P., Nordberg, M., 1996. A framework of motivations for industry - Big science collaboration: A case study. J. Eng. Technol. Manag. - JET-M. https://doi.org/10.1016/S0923-4748(96)01011-9

Autio, E., Hameri, A.P., Vuola, O., 2004. A framework of industrial knowledge spillovers in big-science centers. Res. Policy. https://doi.org/10.1016/S0048-7333(03)00105-7

Bahar, S., 2018. Spandrels, Exaptations, and Raw Material. https://doi.org/10.1007/978-94-024-1054-9_15

Bakhshi, H., Freeman, A., Potts, J., 2011. State of Uncertainty: Innovation policy through experimentation. Nesta.

Banke, J., 2010. Technology readiness levels demystified [WWW Document]. Natl. Aeronaut. Sp. Adm. URL https://www.nasa.gov/topics/aeronautics/features/trl_demystified.html

Beale, R., 2007. Supporting serendipity: Using ambient intelligence to augment user exploration for data mining and web browsing. Int. J. Hum. Comput. Stud. https://doi.org/10.1016/j.ijhcs.2006.11.012

Beard, T.R., Ford, G.S., Koutsky, T.M., Spiwak, L.J., 2009. A Valley of Death in the innovation sequence: an economic investigation. Res. Eval. 18, 343–356. https://doi.org/10.3152/095820209X481057

Bianco, W., Gerhart, D., Nicolson-Crotty, S., 2017. Waypoints for Evaluating Big Science*. Soc. Sci. Q. https://doi.org/10.1111/ssqu.12467

Boisot, M., Nordberg, M., Yami, S., Nicquevert, B., 2011. Collisions and Collaboration: The Organization of Learning in the ATLAS Experiment at the LHC. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199567928.001.0001

Borrás, S., Edquist, C., 2013. The choice of innovation policy instruments. Technol. Forecast. Soc. Change. https://doi.org/10.1016/j.techfore.2013.03.002

Borrás, S., Laatsit, M., 2019. Towards system oriented innovation policy evaluation? Evidence from EU28 member states. Res. Policy. https://doi.org/10.1016/j.respol.2018.08.020

Bush, V., 1945. Science, the Endless Frontier: A Report to the President by Vanevar Bush, Director of the Office of Scientific Research and Development. July 1945. US Gov. Print. Off. Washingt. 1945.

Butler, D., 2008. Translational research: Crossing the valley of death. Nature. https://doi.org/10.1038/453840a

Castelnovo, P., Florio, M., Forte, S., Rossi, L., Sirtori, E., 2018. The economic impact of technological procurement for large-scale research infrastructures: Evidence from the Large Hadron Collider at CERN. Res. Policy. https://doi.org/10.1016/j.respol.2018.06.018

Chesbrough, H.W., 2003. Open Innovation: The New Imperative for Creating and Profiting from Technology., Harvard Business School Press, Boston. https://doi.org/10.1111/j.1467-8691.2008.00502.x

Corneli, J., Jordanous, A., Guckelsberger, C., Pease, A., Colton, S., 2014. Modelling serendipity in a computational context. arXiv Prepr. arXiv1411.0440.

Crease, R.P., Martin, J.D., Pesic, P., 2016. Megascience. Phys. Perspect. 18, 355–356. https://doi.org/10.1007/s00016-016-0193-0

Cunha, M.P. e., Clegg, S.R., Mendonça, S., 2010. On serendipity and organizing. Eur. Manag. J. 28, 319–330. https://doi.org/10.1016/j.emj.2010.07.001

Cunningham, P., Edler, J., Flanagan, K., Laredo, P., 2013. Innovation policy mix and instrument interaction: a review. Manchester Univ. Manchester.

de Rond, M., 2014. The structure of serendipity. Cult. Organ. 20, 342–358. https://doi.org/10.1080/14759551.2014.967451

Dew, N., 2009. Serendipity in Entrepreneurship. Organ. Stud. 30, 735–753. https://doi.org/10.1177/0170840609104815

Doing, P., 2018. Velvet Revolution at the Synchrotron, Velvet Revolution at the Synchrotron. https://doi.org/10.7551/mitpress/7537.001.0001

Edler, J., Fagerberg, J., 2017. Innovation policy: What, why, and how. Oxford Rev. Econ. Policy. https://doi.org/10.1093/oxrep/grx001

Edler, J., Georghiou, L., 2007. Public procurement and innovation-Resurrecting the demand side. Res. Policy. https://doi.org/10.1016/j.respol.2007.03.003

European Commission, 2016a. Open innovation, open science, open to the world - a vision for Europe. https://doi.org/10.2777/552370

European Commission, 2016b. Supply and Demand Side Innovation Policies.

European Commission, 2012. Commission Recommendation of 17.7.2012 on access to and preservation of scientific information.

Fink, T.M.A., Reeves, M., Palma, R., Farr, R.S., 2017. Serendipity and strategy in rapid innovation. Nat. Commun. 8, 1–9. https://doi.org/10.1038/s41467-017-02042-w

Florio, M., Forte, S., Sirtori, E., 2016. Forecasting the socio-economic impact of the Large Hadron Collider: A cost–benefit analysis to 2025 and beyond. Technol. Forecast. Soc. Change. https://doi.org/10.1016/j.techfore.2016.03.007

Florio, M., Sirtori, E., 2016. Social benefits and costs of large scale research infrastructures. Technol. Forecast. Soc. Change. https://doi.org/10.1016/j.techfore.2015.11.024

Frank, C., Sink, C., Mynatt, L., Rogers, R., Rappazzo, A., 1996. Surviving the valley of death: A comparative analysis. J. Technol. Transf. https://doi.org/10.1007/BF02220308

Frost & Sullivan, 2015. 2015 Top Technologies in Sensors & Control (Technical Insights).

Galison, P., 2016. Meanings of scientific unity: The law, the orchestra, the pyramid, the quilt and the ring, in: Pursuing the Unity of Science. Routledge, pp. 12–29.

Galison, P., Hevly, B., 1992. Big Science: The Growth of Large-Scale Research. Stanford University Press.

Garud, R., Gehman, J., Giuliani, A.P., 2018. Serendipity arrangements for exapting science-based innovations. Acad. Manag. Perspect. https://doi.org/10.5465/amp.2016.0138

Giles, D., Walkowicz, L., 2019. Systematic serendipity: A test of unsupervised machine learning as a method for anomaly detection. Mon. Not. R. Astron. Soc. https://doi.org/10.1093/mnras/sty3461

Giudice, G.F., 2012. Big Science and the Large Hadron Collider. Phys. Perspect. https://doi.org/10.1007/s00016-011-0078-1

Guan, J.C., Yan, Y., 2016. Technological proximity and recombinative innovation in the alternative energy field. Res. Policy. https://doi.org/10.1016/j.respol.2016.05.002

Hallonsten, O., 2014. How expensive is Big Science? Consequences of using simple publication counts in performance assessment of large scientific facilities. Scientometrics. https://doi.org/10.1007/s11192-014-1249-z

Hammett, F.S., 1941. The "meaning" of science. Science (80-. ). https://doi.org/10.1126/science.93.2425.595-a

Héder, M., 2017. From NASA to EU: The evolution of the TRL scale in Public Sector Innovation. Innov. J.

Heidler, R., Hallonsten, O., 2015. Qualifying the performance evaluation of Big Science beyond productivity, impact and costs. Scientometrics. https://doi.org/10.1007/s11192-015-1577-7

Heinze, T., Hallonsten, O., 2017. The reinvention of the SLAC National Accelerator Laboratory, 1992–2012. Hist. Technol. 33, 300–332. https://doi.org/10.1080/07341512.2018.1449711

Hellström, T., Jacob, M., 2012. Revisiting "Weinberg's Choice": Classic Tensions in the Concept of Scientific Merit. Minerva. https://doi.org/10.1007/s11024-012-9203-9

Hiltzik, M., 2016. Big Science: Ernest Lawrence and the Invention that Launched the Military-Industrial Complex. Simon & Schuste.

Hudson, J., Khazragui, H.F., 2013. Into the valley of death: Research to innovation. Drug Discov. Today. https://doi.org/10.1016/j.drudis.2013.01.012

Islam, N., 2017. Crossing the Valley of Death-An Integrated Framework and a Value Chain for Emerging Technologies. IEEE Trans. Eng. Manag. https://doi.org/10.1109/TEM.2017.2685138

Kotkov, D., Wang, S., Veijalainen, J., 2016. A survey of serendipity in recommender systems. Knowledge-Based Syst. https://doi.org/10.1016/j.knosys.2016.08.014

Leyden, D.P., Menter, M., 2018. The legacy and promise of Vannevar Bush: rethinking the model of innovation and the role of public policy. Econ. Innov. New Technol. https://doi.org/10.1080/10438599.2017.1329189

Lundvall, B.Å., 2010. National systems of Innovation: Toward a theory of Innovation and Interactive Learning, National Systems of Innovation: Toward a Theory of Innovation and Interactive Learning. https://doi.org/10.7135/UPO9781843318903

Lundvall, B.A., Borrás, S., 2009. Science, Technology, and Innovation Policy, in: The Oxford Handbook of Innovation. https://doi.org/10.1093/oxfordhb/9780199286805.003.0022

Mak, K.K., Pichika, M.R., 2019. Artificial intelligence in drug development: present status and future prospects. Drug Discov. Today. https://doi.org/10.1016/j.drudis.2018.11.014

Makri, S., Blandford, A., Woods, M., Sharples, S., Maxwell, D., 2014. "Making my own luck": Serendipity strategies and how to support them in digital information environments. J. Assoc. Inf. Sci. Technol. 65, 2179–2194. https://doi.org/10.1002/asi.23200

Markham, S.K., Ward, S.J., Aiman-Smith, L., Kingon, A.I., 2010. The valley of death as context for role theory in product innovation. J. Prod. Innov. Manag. https://doi.org/10.1111/j.1540-5885.2010.00724.x

Martin, B.R., 2016. Twenty challenges for innovation studies. Sci. Public Policy. https://doi.org/10.1093/scipol/scv077

Mazzucato, M., 2016. From market fixing to market-creating: a new framework for innovation policy. Ind. Innov. https://doi.org/10.1080/13662716.2016.1146124

Mazzucato, M., Semieniuk, G., 2017. Public financing of innovation: New questions. Oxford Rev. Econ. Policy. https://doi.org/10.1093/oxrep/grw036

McCay-Peet, L., Toms, E.G., 2015. Investigating serendipity: How it unfolds and what may influence it. J. Assoc. Inf. Sci. Technol. 66, 1463–1476. https://doi.org/10.1002/asi.23273

Merriam-Webster, 2020. Serendipity [WWW Document]. Merriam-Webster.com Dict. URL https://www.merriam-webster.com/dictionary/serendipity (accessed 2.25.20).

Merton, R.K., Barber, E., 2004. The travels and adventures of serendipity: A study in sociological semantics and the sociology of science. Princeton Univrsity Press.

Mowery, D.C., 2012. Defense-related R&D as a model for "grand Challenges" technology policies. Res. Policy. https://doi.org/10.1016/j.respol.2012.03.027

Organisation for Economic Co-operation and Development (OECD), 2014. The Impacts of Large Research Infrastructures on Economic Innovation and on Society: Case Studies at CERN.

Organisation for Economic Co-operation and Development (OECD), 2003. Turning Science into Business: Patenting and Licensing at Public Research Organisations.

Oxford University Press, 2019. Serendipity [WWW Document]. Lexico.com. URL https://www.lexico.com/definition/serendipity (accessed 2.25.20).

Rahm, D., Bozeman, B., Crow, M., 1988. Domestic Technology Transfer and Competitiveness: An Empirical Assessment of Roles of University and Governmental R&D Laboratories. Public Adm. Rev. https://doi.org/10.2307/976993

Rolfstam, M., 2009. Public procurement as an innovation policy tool: The role of institutions. Sci. Public Policy. https://doi.org/10.3152/030234209X442025

Rosenman, M.F., 2001. Serendipity and Scientific Discovery, in: Creativity and Leadership in the 21st Century. pp. 187–193.

Scherer, F.M., 1982. Demand-Pull and Technological Invention: Schmookler Revisted. J. Ind. Econ. https://doi.org/10.2307/2098216

Schmookler, J., 1962. Economic Sources of Inventive Activity. J. Econ. Hist. https://doi.org/10.1017/S0022050700102311

Schoenmakers, W., Duysters, G., 2010. The technological origins of radical inventions. Res. Policy. https://doi.org/10.1016/j.respol.2010.05.013

Schopper, H., 2016. Some remarks concerning the cost/benefit analysis applied to LHC at CERN. Technol. Forecast. Soc. Change 112, 54–64. https://doi.org/10.1016/j.techfore.2016.02.007

Scudellari, M., 2017. Big science has a buzzword problem. Nature. https://doi.org/10.1038/541450a

Weinberg, A.M., 1964. Criteria for scientific choice II: The two cultures. Minerva. https://doi.org/10.1007/BF01630147

Weinberg, A.M., 1963. Criteria for scientific choice. Minerva. https://doi.org/10.1007/BF01096248

Weinberg, A.M., 1961. Impact of large-scale science on the United States. Science (80-. ). https://doi.org/10.1126/science.134.3473.161

Wilson, D.A., 1991. The Vannevar Bush Legacy. Science (80-. ). 251, 210–210. https://doi.org/10.1126/science.251.4990.210

Wolfe, A.K., Bjornstad, D.J., Shumpert, B.L., Wang, S.A., Lenhardt, W.C., Campa, M.F., 2014. Insiders' views of the valley of death: Behavioral and institutional perspectives. Bioscience. https://doi.org/10.1093/biosci/bit015

World Intellectual Property Organization, 2010. Managing IP at CERN. WIPO Mag.

Yaqub, O., 2018. Serendipity: Towards a taxonomy and a theory. Res. Policy. https://doi.org/10.1016/j.respol.2017.10.007

# 6

# 6. From Bits to Atoms: White Rabbit at CERN

The article that constitutes this chapter consists in a micro-study of a single case that gives us insight into the different *mechanisms* that help reconcile the main tensions between the two exogenous influence presented (i.e open science and technology transfer).The study empirically investigates White Rabbit, an *open-source hardware* initiated at CERN and transferred to multiple industrial settings

## 6.1 Abstract

The success of Open Source Software (OSS) has inspired others to adopt the 'open source way' of development to the field of electronic hardware design: Open Source Hardware (OSH). While there are many expectations that an open-source ethos will influence commercial hardware development to the same degree that it has influenced software development, yet little is known about how the transposition of open-source development to a form of technology 'object' that has material components (hybrid objects) might be unsuccessful if the conditions salient in OSS development are not equally applicable in OSH. We study the development of White Rabbit (WR), an OSH initiated at CERN and deployed as a powerful precision and synchronization technology in many industrial settings where time accuracy is critical. Through the investigation of WR, our study contributes to recent conceptualizations of digital objects by uncovering the differences from hybrids to purely non-material digital objects and elucidates what happens when we transpose the OS model of development to a hybrid object. As a lens to understand how different attributes of objects require different development models, we adopt relevant constructs from Transaction Costs Economics (TCE) and examine its utility as a predictive theory of OSH development.

**Keywords:** Open source hardware, hybrid objects, development, transaction costs economics.

## 6.2 Introduction

"Oh dear! Oh dear! I shall be too late!" Alice follows the time-obsessed hare down the rabbit hole into Wonderland. The White Rabbit is the first Wonderland character that Alice encounters in *Alice's Adventures in Wonderland,* a fantasy novel by English mathematician Charles Lutwidge Dodgson published in 1865. White Rabbit (WR) is also the name of an open source hardware (OSH) that consists of a fully deterministic Ethernet-based network for data transfer and synchronization. The technology was developed in 2008 by the European Organization for Nuclear Research (centre européen pour la recherche nucléaire; CERN) to provide a sequencing and synchronization solution for CERN's geographically distributed accelerator network. WR was developed as an OSH through a sustained collaboration among traditional vendors, peripheral research organizations, and a heterogeneous community of voluntary contributors. WR was born as the evolution of CERN's General Machine Timing (GMT) program and is currently the clock and event distribution system of their accelerators where time accuracy at the nanosecond[15] level is required. After its implementation at

[15] A nanosecond (ns) is an SI unit of time equal to one billionth of a second, that is, 1/1,000,000,000 of a second, or 10−9 seconds.

CERN, WR was adopted by other scientific research infrastructures and subsequently implemented in a variety of industrial settings where time accuracy is critical, including high-frequency trading (HFT) matching engines in financial services, telecommunications networks, automated vehicles, modern central navigation systems for air traffic control, and smart grids.

The success of Open Source Software (OSS) has inspired others to adopt the 'open source way'[16] of development in the field of electronic hardware design: OSH. OSH is a term for hardware or tangible artifacts – machines, devices, or other physical things – for which the design is made publicly available in a way that anyone can study, modify, distribute, make, and sell either the design or hardware based on this design (OSH Association). OSH typically comprises both material and non-material layers; that is, it can be viewed as a stack of technologies with a physical form combined with embedded operating, middleware, or application-level software. Although OSH is most common in scientific research infrastructures (Balka 2011; Boisseau et al. 2018; Mellis and Buechley 2012; Pearce 2012), it has now attracted the attention of a wide range of industrial organizations that need to develop never-seen-before technologies not easily acquired through commercial vendors. The domain of OSH includes a diverse range of projects and products such as computer systems and components, scientific machines and tools, robotics, home automation, and medical and biotech instruments (Pearce 2012). Some compelling examples of OSH[17] include Arduino[18], RepRap[19], and the Open Compute Project[20]. Projects range from small-scale, do-it-yourself hardware projects for electronics hobbyists to complex projects that require highly sophisticated expertise, long development cycles, and industrial manufacturing capabilities, which render them cost-prohibitive to hobbyists or small research laboratories (Balka et al. 2009; Boisseau et al. 2018; Oberloier and Pearce 2017). For example, RISC-V[21] is gaining momentum as an OSH instruction set architecture (ISA) in both research and commercial organizations that seek to avoid the non-recurring engineering costs of specialized integrated circuits.

With this recent evolution in OSH, some scholars have predicted that an open-source ethos will influence commercial hardware development to the same degree that it has influenced software development (Balka et al. 2010; Powell, 2016). However, the transposition of open-source development to a form of technology ‘object’ that has material components might be unsuccessful if the conditions salient in OSS development are not equally applicable in OSH.

[16] We borrow this expression from Howison and Crowston (2014).

[17] See a comprehensive list of examples OSH at ohwr.org.

[18] arduino.cc

[19] reprap.org

[20] opencompute.org

[21] riscv.org

Following Faulkner and Runde (2019), we define *objects* as entities that endure (i.e., “something that exists through time and is fully present at each and every point in time over the period of its existence” p. 5) and entities that are structured, that is, entities composed of a number of distinct parts (henceforth *components*). We employ the term ‘object’ in the same spirit as Faulkner and Runde (2009, 2013, 2019) and Kallinikos et al. (2013) to designate purposefully engineered objects rather than any object that occurs naturally. The universe has all types of objects; we refer here only to a subset of them that we term digital objects. Digital objects have not only a *function* that “members of some community impose on that object in pursuit of their practical interests” (Faulkner and Runde 2013 p. 807) but also a *form*, which means that they possess the characteristics and capabilities so that the object's function can be performed. Both function and form give the object its technical identity. For instance, using Faulkner and Runde's (2013) example, an application for network monitoring derives its technical identity by facilitating the monitoring of devices connected to the local network. Digital objects have components or constituent parts that do not have a technical identity themselves to the degree that they do not fulfill a function given by a community, but they possess different *attributes*; that is, digital objects have defining properties according to how the components work, how they are arranged, and how they interact with one another.

A principal *attribute* to distinguish is the difference between material and non-material components, which is their embodiment. The notion of embodiment (Yoo 2010) owes its legacy to the philosophy of phenomenology (Boland 1986; Heidegger 1962) and refers to “the property of being manifest in and of the everyday world” (Dourish 2001 p.18). Material components have spatial attributes (i.e., shape, volume, mass, and location). That is, material components have a “physical mode of being” (Faulkner and Runde 2013 p. 806), which makes them “rigid, stable and tangible” (Yoo 2010 p. 222) as opposed to non-material components, which exist “in a logic state, which makes them malleable and fungible” (von Briel et al. 2018 p. 281). Digital objects with material and non-material components fall into the group of ***hybrid objects,*** or ***hybrids*** (Faulkner and Runde 2019). Hybrid objects include any hardware with middleware or software (Yoo 2010) and encompass many of the objects being developed in OSH projects.

To our knowledge, little scholarly work has systematically investigated how the premises of open-source development differ when applied to hybrids. As other scholars have emphasized, the open-source model may not easily be transposed to hybrid development (Balka 2011; Balka et al. 2010; Boisseau et al. 2018; Oberloier and Pearce 2017; West and Kuk 2016) or any object different than software (Lerner and Tirole 2003). Therefore, further research is warranted given the expected impact of OSH. Moreover, an analysis of the open-source development of hybrids offers the variance needed to theorize, beyond OSS, the relationship between the attributes of object components (i.e., *what* is being developed) and the organizational conditions of their development model (i.e., *how* it is being developed). Accordingly, the research question that this study asks is

> *How do* the *attributes of a hybrid object and its components affect the open-source model of development?*

To answer our research question, we first engage in a review of recent conceptualizations of digital objects to delineate the attributes of the digital objects that contain both material and non-material components (hybrids) to understand 'what' agents act on when they develop them (Faulkner and Runde 2019; Kallinikos et al. 2013; Yoo et al. 2010). We adopt the definition of development as the "social process of designing, developing, and implementing the technical artifact, usually in a specific organizational context and over time" (Akhlaghpour et al. 2013 p.152). In a second step, we review what we know about hybrid development to understand how hardware that contains middleware and software has traditionally been developed. Third, we review the IS literature on the 'open source way' of developing software. Our goal is to extract from this literature the common characteristics of how work is organized in OS development and the conditions that underlie it (i.e., the prerequisites for the occurrence of OS development) (Benkler 2002; Dahlander and Magnusson 2008; Feller and Fitzgerald 2002; Fitzgerald 2006; Fitzgerald and Feller 2002; Howison and Crowston 2014; O'Mahony and Ferraro 2007).

As a lens to understand how hybrid component attributes and their interaction require different development models, we briefly review transaction costs economics (TCE) (Williamson 1975, 1985, 1996) as a high-level theoretical frame. We find the logic of TCE to be useful because it works through the *strategic alignment hypothesis*, namely, that transactions with different attributes will align with governance structures that vary in their relative ability to economize on particular attributes of transaction costs (Williamson 1996). We believe that this underlying mechanism is similar to and useful for understanding the relationship between the attributes of a hybrid object's components and how they align with an appropriate model of development. We appropriate relevant concepts from TCE and explore both their adequacy and limitations in explaining hybrid object development.

The remainder of the paper is organized as follows. After discussing the theoretical underpinnings of our research study and analysis, we then describe our research context around WR and its development, the research design and analytic methods. We also present the findings and discuss the theoretical and practical implications of the study, its limitations, and future research.

## 6.3 Theoretical underpinnings

### *6.3.1 The attributes of hybrids*

Beginning with Orlikowski and Iacono's (2001) empirical study about the treatment of digital technology in IS research, scholars have portrayed the specific features of digital technology in diverse ways (e.g., Ekbia 2009, Kallinikos et al., 2011, 2013, Faulkner and Runde 2009, 2019; Yoo et al., 2010), with the greatest emphasis on the non-material aspects of digital objects. Attributes such as non-rivalry, infinite expansibility, reproducibility (Faulkner and Runde 2009, 2013) and largely unstable, unbounded (Ekbia 2009), interactive, fluid, editable or distributed attributes (Kallinikos et al. 2010, 2013 p.360; Manovich 2001) are supported by examples such as blogs, wikis, personal profiles in social media, booking systems, digital libraries, files, images, films or videos, and open-source software. This literature stream is less attentive to the physical nature of components, for instance, their capability to be reproduced, distributed, or their stability over time. Thus, there is an opportunity to better understand the attributes of hybrid digital objects where material characteristics are present.

From our review of the literature, five salient attributes are relevant for our analysis (e.g., Ekbia 2009; Faulkner and Runde 2009, 2013, 2019; Kallinikos et al. 2010, 2013; Yoo 2010; Nambisan et al. 2017). These five attributes are 1) *embodiment*, 2) *modularity*, 3) *granularity*, 4) *editability*, and 5) *reproducibility*. Our study applies these five attributes to assess how these traits vary when a digital object contains material components. We provide the construct definitions and our conceptual departure from the related notions in Appendix A.

Embodiment refers to the component's material or non-material state as described previously (Faulkner and Runde 2009, 2011; Yoo et al. 2010). Modularity describes components as either 1) loosely coupled – where functionalities are dependent yet distinct from one another or 2) tightly coupled – where components are more closely integrated and responsive to, but less distinct, from one another (von Briel et al. 2018; Kallinikos et al. 2010, 2013; Manovich 2001; Yoo 2010). Granularity is also determinative, as it affects the degree to which a development task can be decomposed into smaller units to be completed by smaller teams or individuals (Kallinikos et al. 2010, 2013; Kallinikos and Mariátegui 2011). Taken together, modularity and granularity describe a great deal about the composition of the object and components, that is, their relative sizes, how they are arranged and relate to one another, and their degree of interdependence (von Briel et al. 2018; Kallinikos et al. 2010). We are also concerned with the degree to which object components are modifiable, specifically, editability, as it affects the degree to which multiple implementations, customizations, or forking are possible at specific points in the technology stack (von Briel et al. 2018; Ekbia

2009; Faulkner and Runde 2019; Kallinikos et al. 2010). Reproducibility is associated with embodiment, as it describes the pragmatic or economic cost of producing and distributing multiple units of the object or component (Kallinikos et al. 2010). In this sense, although digital objects with non-material components (e.g., a web browser) can be downloaded unlimited times once the code is written, each copy of a hybrid object or component requires physical production and distribution (von Briel et al. 2018; Faulkner and Runde 2009, 2013, 2019; Kallinikos et al. 2013).

### *6.3.2 The development of hybrids*

The development of hybrids has traditionally been characterized as requiring a number of discrete and sequential steps that are not as easily decomposed, distributed or completed in parallel processes (DeMicheli and Sami 2013). Although there is no single approach to hybrid development, our purpose is to identify the common features across the literature.

The first step in the development of hybrids is the logical design, which is represented in the schematic diagram. The schematic diagram provides no information on the physical arrangement or interconnection of the parts; it is only a logical depiction of the object. Where one could argue that hybrids and pure software design are similar up to this point (logical design), they diverge from here. Hybrid objects require a *translational action* to go from the digital representation of the object (the logical design) to the object itself. Translational action refers to "practices associated with movement from one layer of the bearer to another" (Faulkner and Runde 2019 p.10). For hybrid objects, the material attributes of the components (e.g., size, heat, etc.) and the interconnection of the parts must be considered. Moving from the schematic to the actual physical layout is somewhat of an art form, as the physical nature of the components must be considered (Ackermann 2009). Electronic design automation (EDA) can aid the developer to generate a netlist that describes each set of electrical connections by grouping them into a 'Net', a group of components that are electrically tied together. This netlist also describes the electrical value and physical attributes of the components from component libraries. However, despite the benefits of EDA software, substantial human expertise is required to evaluate the challenges of size constraints, heat, radio interference, external connections, component cost and other operational and environmental factors; two equally qualified designers could easily produce two different circuit boards of varying quality based on the same schematic (Ackermann 2009).

Sophisticated hybrid objects with multiple components can have subassemblies. When the object is more complex, it is more difficult to complete the detailed development of any one part until the entire subassembly is developed, which reinforces the sequential nature of the process. For such objects, no matter how much care is put into the object architecture and design, unexpected side effects arise when each prototype is assembled and tested for the first time; these are side effects that will not appear when the components are tested in isolation (Pan et al. 2018). For this reason, where commercial software testing is often conducted by independent software specialists who test individual components and their integration according to testing scripts (Wareham and Sonne 2008), for such hybrid objects, this testing is often performed by the engineers who design the object given their more tightly coupled and integrated nature (Drechsler and Breiter 2007). Moreover, the integration of different subassemblies requires ensuring the compatibility of the different components with market standards (DeMicheli and Sami 2013; Gajski and Vahid 1995).

Several observations about the traditional development of hybrids are worth noting. When the phases of development are independent, (i.e., logical design, schematic capture, physical design, prototyping, and testing), they are highly interdependent and sequential. Changes in the fundamental design are more difficult and expensive to modify later in the development cycle, as a change of one component “is likely to require extensive compensating changes in the designs of many interrelated components” (Sanchez and Mahoney 1996 p.65 ). This generates a certain inflexibility to engineering modifications later in the process, which imposes some stability on the core structure and principal components (DeMicheli and Sami 2013). In addition, the use of non-standard components requires more time between development iterations for “procuring materials, creating tooling, trial runs, product assembly, [and] quality control” (von Briel et al. 2018 p.283; Marion et al. 2012). Non-standard designs also have an acute effect on testing costs, which limit the frequency of design iterations and further constrain the possibility of modifications by different developers (Gajski and Vahid 1995; Mellis and Buechley 2012).

Various software tools can only partially alleviate these challenges. Although EDA offers substantial automation benefits, as mentioned, human expertise plays a substantial role in the actual physical design of the object (Ackermann 2009). Other software tools exist for tracking and integrating concurrent modifications introduced by different developers (Mellis and Buechley 2012). However, the limited maturity of these tools requires far more centralized direction-giving; this effect is exacerbated by the addition of multiple software layers on top of the hardware (Pan et al. 2018; Drechsler and Breiter 2007). Even if employing virtual prototypes (Bogers and Horst 2014) or advanced manufacturing techniques (e.g., 3D printing) (Bogers et al. 2016), the development of hybrids “involves more activities such as transferring premature prototypes into designs that can actually be manufactured” (von Briel et al. 2018 p.283; Yu et al. 2018), which incur time, especially when compared to the modifications to software based on writing lines of code (Mellis and Buechley 2012).

Finally, the high interdependence of physical components can aggravate the business and financial aspects of the development process to control costs, which encourages linear process planning to define who contributes at various points of the process (von Briel et al. 2018; Yu et al. 2018). Overall, such centralized direction-giving in the development process has typically been exercised in organizational hierarchies, that is, following formal rules within an organization to preserve "agency over component supply and functions" (von Briel et al. 2018 p.283) or contractual mechanisms that specify development processes and outcomes that preserve input control over the development.

### *6.3.3 Conditions for OS development*

The basic characteristics of the open-source model have been articulated in a large number of publications by its advocates (e.g. Raymond, 1999; Cook, 2001; Linux Documentation Project, 2001; Masum, 2001) through diverse case studies of OS development projects (Mockus et al., 2002; Scacchi, 2001). Essentially, the open-source model has been described as an alternative organizational model for development, which is neither market nor hierarchy (Shah 2006). Diverse and partially overlapping approaches have described it as commons-based peer production (Benkler 2006), a community-based model (Shah 2005, 2006), open sourcing (Ågerfalk and Fitzgerald 2008), collective invention (Allen 1983), private-collective innovation (von Hippel and von Krogh 2003) or distributed innovation (Lakhani and von Hippel 2004).

Early OS research primarily concentrated on delineating the unique characteristics of the '*open source way'* of software development (Crowston and Howison 2006; Feller and Fitzgerald 2002, 2002; Mockus et al. 2002; Raymond 1999), how open source communities coordinate work (Ben-Menahem et al. 2015; Crowston and Howison 2006; Howison and Crowston 2014; Koch and Schneider 2002; Krogh and Hippel 2006) and how they are governed (O'Mahony and Ferraro 2007; Sharma et al. 2002; Tullio and Staples 2013). As the success of OS initiatives progressed and commercial companies increasingly engaged in OS communities, scholars studied the transformation of OS into a more mainstream and commercial form of developing software, which is labeled as OSS 2.0 (Fitzgerald 2006). With a strong commercial orientation, OS went from a phenomenon of "ideologically driven developer communities" (Rolandsson et al. 2011 p.577) to a commercial model of developing software where many companies engage with communities in collaborative developments (Niederman et al. 2006).

Although OS is not a homogeneous approach to software development, the specific *attributes* (i.e., common characteristics of how work is organized) and *conditions* (i.e., prerequisites for the occurrence of OS development) that underlie its model of development are common in this literature. These attributes are (a) the voluntary nature of the collaboration where agents work autonomously and self-select their tasks in (Crowston, 1997; Howison and Crowston, 2014; Lindberg et al., 2016; Maha and Vaast, 2015; Shah 2005, 2006) (b) a loosely centralized collaboration (Cutosksy et al., 1996; Feller and Fitzgerald 2000, 2002) (c) where geographically distributed teams or individuals (Cook 2001; Feller et al. 2008; Feller and Fitzgerald 2002, 2002; Markus 2007) (d) work in parallel development in an asynchronous collaboration of tasks supported by (Cook 2001; Feller et al. 2008; Feller and Fitzgerald 2001, 2002; Markus 2007) (e) infrastructural tools such as the internet and concurrent versioning of software (Baldwin and Clark 2006; Egyedi and Joode 2004; Feller et al. 2008; Feller and Fitzgerald 2002). Table 5 in the appendix provides a summary of the main attributes and related sources in the literature.

Underlying such characterization of OS development are some requisite *conditions.* These are a) *modularity* (Benkler 2002; Fitzgerald 2006; Howison and Crowston 2014; Lindberg et al. 2014; MacCormack et al. 2006), *b) granularity* (Benkler 2002, 2006; Howison and Crowston 2014; Lindberg et al. 2014), and *c) low integration costs* (Benkler 2002; Feller and Fitzgerald 2002; Howison and Crowston 2014; Langlois and Garzarelli 2008). As Howison and Crowston (2015) argue, to enable asynchronous collaboration, the "open superposition" of tasks is necessary. This requires that each module creates an "(adequately) finished artifact" (Howison and Crowston 2015 p. 44) that can be completed by an individual programmer in a geographically distributed environment. This further assumes not only that tasks can be broken down into smaller independent problems (*modularity*) but also that such tasks are sufficiently granular for independent and geographically distributed individuals to understand and complete them (*granularity*). In other words, "to pool a relatively large pool of contributors, the modules should be predominantly fine-grained, or small in size. This allows the project to capture contributions from large numbers of contributors whose motivation level will not sustain anything more than quite small efforts towards the project" (Benkler 2006, p. 10). Moreover, with requisite modularity and granularity, OS contributors can practice what Howison and Crowston (2014) call productive deferral, where difficult tasks can be deferred to allow developers to work on easier tasks (which is therefore asynchronous and non-linear development).

Finally, these independent modules must be re-integrated to form a useful system. This requires low instantiation costs, that is, the costs of rebuilding and adding additional layers to existing work and quality controls over the modules and the costs of integrating completed modules and making them interoperable (Howison and Crowston, 2014; Benkler 2006). When these appropriate *integration characteristics* (e.g., efforts and costs) are sufficiently low, non-linear, asynchronous development is increasingly feasible: complex, functionally interdependent work can be broken down and completed without prohibitive decomposition and reintegration efforts. Table 5 in Appendix A provides a summary of the main conditions for OS development and the related sources in the literature.

### *6.3.4 A Transaction Costs Economics Perspective on Hybrid Development*

Transaction cost economics (TCE) has had an established tradition in management studies since the seminal publications of Williamson (1975, 1985). TCE has been applied in information systems research, particularly in an attempt to understand how ICT reduces external and internal coordination costs, thereby affecting firm size and managerial controls (Gurbaxani and Whang 1991; Malone et al. 1987). The literature on TCE is vast, and excellent reviews of it exist (e.g., Macher and Richman 2008). It is important to emphasize that although the core concepts of TCE are described in the works of Williamson (1975, 1985, 1996), TCE discourse has been applied to so many domains in management, law and social science that a comprehensive interpretation of the theory is well beyond the scope of this paper. What is relevant for our analysis is that TCE works through the *discriminating alignment hypothesis*: *that transactions with different attributes will align with governance structures that differ in their relative ability to economize on particular attributes of transaction costs* (Williamson 1996). Given transaction attributes, the resulting transaction will be governed by mechanisms conceptualized as falling at some point on a market-hierarchy continuum. Where the costs of using market-based mechanisms are excessively high, transactions will be internally integrated into a single organization or firm. With progressively lower external coordination costs and asset specificity, transactions can be completed in governance forms that can be considered to be decreasingly complex and less centrally governed. The most commonly cited TCE transaction attributes are defined in Table 5 in Appendix A.

TCE is a predictive theory. By exploring the conceptual similarity between *transaction attributes* and *component attributes*, the logic of TCE can be extended to predict the development models (as outcomes roughly equivalent to TCE governance structures) of hybrid objects according to their component attributes (Niederman et al. 2006). We therefore focus on several key constructs of TCE, primarily product and production attributes, that are most useful to our analysis and show how they relate to the five component attributes discussed earlier. Note that there is some overlap among the concepts, which make a one-to-one mapping difficult. We have attempted to simplify at an appropriate level to understand the most important conceptual relationships and extrapolate predictive statements on how the development of hybrid object components will be governed based on the related TCE logic.

*Table 1. Relating the hybrid component attributes to TCE constructs*

| Component Attribute | Relevant TCE Construct | Rationale | TCE Prediction |
|---|---|---|---|
| Embodiment<br>*Material*<br>*Non-material* | Asset Specificity | TCE says very little about the differences between tangible and intangible assets; specificity could vary widely in both cases. | None |
| Modularity<br>*Tightly coupled*<br>*Loosely coupled* | Interdependence | Tightly coupled components are more interdependent.<br>Higher interdependence requires more centralized coordination. | *The development of tightly coupled components is coordinated through more centralized governance.* |
| | | Loosely coupled components are less interdependent. Lower interdependence requires less-centralized coordination. | *The development of loosely coupled components is coordinated through less centralized governance.* |
| | Product/Process complexity | Complex components and production processes require higher monitoring costs. | *Development processes with higher monitoring costs are coordinated through more centralized governance.* |
| | Monitoring Costs | Simple components and production processes require lower monitoring costs. | *Development processes with lower monitoring costs are coordinated through less centralized governance.* |
| *Integration characteristics* | | Tightly coupled and complex components are difficult to decompose and re-integrate, which increases integration costs. | *Components with high integration costs are coordinated through more centralized governance.* |
| | | Loosely coupled and less-complex components are easier to decompose and re-integrate, which decreases integration costs. | *Components with low integration costs are coordinated through less centralized governance.* |

| | | | |
|---|---|---|---|
| Granularity<br>*High granularity*<br>*Low granularity* | Duration<br>Frequency<br>Monitoring Costs | Highly granular components can be developed with a shorter duration, greater frequency, and lower monitoring cost.<br>Low granularity components are developed with a longer duration, lower frequency, and higher monitoring cost. | *Components with higher granularity are coordinated through less centralized governance.*<br>*Components with low granularity are coordinated through more centralized governance.* |
| Editability<br>*High editability*<br>*Low editability* | Asset Specificity | A component or IT object that is editable is more readily configurable to alternative uses. Asset specificity decreases with greater editability.<br>A component or IT object that has low editability is less readily configurable to alternative uses. Asset specificity increases with lesser editability. | *Components that are less asset-specific are coordinated through less centralized governance.*<br>*Components that are highly asset specific are coordinated through more centralized governance.* |
| Reproducibility | Transaction Risk (financial & legal) | Components with low technical sophistication and low economic costs are easily reproducible.<br>Components with high technical sophistication and high economic costs are difficult to reproduce. | *Easily reproducible components confer lower financial risk and can therefore be developed through less centralized governance.*<br>*Difficult to reproduce components confer greater financial risk and are therefore developed through more centralized governance.* |

What is evident from this exercise is that although there are many useful similarities, some logical extensions are required to equate TCE transaction attributes to hybrid object component attributes. We assess these limitations in the Discussion section of this article.

For our analysis, we do not consider all the transaction governance outcomes identified in the extensive TCE literature (e.g., joint ventures, franchising, complex versus simple contracts, etc.). Rather, we collapse them into two generalized governance modes based on decreasing levels of centralized coordination and direction-giving, namely, 1) *hierarchical control* and 2) *contractual agreements*, with the addition of 3) *voluntary contributions* from the OS literature discussed previously, given that the OS literature has argued that OS communities are a competing form of governance mode along hierarchies and markets (Benkler 2002; Demil and Lecocq 2006; Niederman et al. 2006; Watson et al. 2005). *Hierarchical control* refers to any activity that is completed, coordinated, or controlled by a single organization. *Contractual agreements* refer to the processes and outcomes described and committed to by transacting parties as stipulated in legally binding agreements. *Voluntary contributions* are the contributions of individuals or organizations that contribute to OS development processes without any pecuniary compensation or legal obligation. A key characteristic of voluntary contributions is that they are organized in a decentralized manner, with each contributor self-selecting their tasks and foregoing any managerial process or price established in contractual agreements (Niederman et al. 2006; Watson et al. 2005).

## 6.4 Research context and methods

We engage in an inductive, longitudinal, in-depth case study about WR, an OSH developed at CERN in collaboration with more than 31 additional organizations. WR offers a powerful opportunity for theory generation by being “paradigmatic of some phenomena of interest” (Gerring 2007, p. 101), where “its extreme value on an independent or dependent variable of interest” helps us to theorize an emerging phenomenon. As a highly complex and sophisticated hybrid object, the approach allowed us to explore deeply contextualized patterns in the open-source development of a hybrid. We studied the ecosystem of the organizations that contributed to the development of WR, namely, firms’ network and research organizations and their interaction when developing a hybrid object such as WR.

### *6.4.1 Research Context*

Since the 1970s, particle physicists have used the so-called Standard Model to describe the fundamental structure of matter. CERN has deployed the world’s most powerful particle accelerators and detectors to test the predictions and limits of the Standard Model, and most recently, they corroborated the existence of the Higgs boson. WR is the name of an OSH initiated in 2008 when engineers at CERN were confronted with limited bandwidth and the impossibility of dynamically evaluating the delay induced by the data links that constitute CERN's geographically distributed computing infrastructure. WR was developed with the following unprecedented specifications: a) the transfer of a time reference from a central location to many destinations with an accuracy better than one nanosecond and a precision better than 50 picoseconds[22]; b) the ability to service more than 1,000 nodes; c) the ability to cover distances of the order of 10 km; d) and data transfer from a central controller to many nodes with a guaranteed upper bound in latency.

Prior to WR, the extant synchronization standard for Ethernet networks was the Precise Time Protocol (PTP), which is standardized as IEEE 1588. WR extends PTP in a backwardly compatible way to achieve sub-nanosecond accuracy (Moreira et al. 2009). "The combination of deterministic latencies with a common notion of time to within one nanosecond allows WR to be a suitable technology to solve diverse problems in distributed real-time control and data acquisition" (Lipiński et al. 2011 p.2).

WR started as an OSH in 2008 when CERN decided to collaboratively develop the technology with any voluntary contributor willing to join the endeavor. An open call was placed in CERN's vendor ecosystem, supported by a repository, wiki, developers' mailing list, workshops and other collaborative tools. Most importantly, an open source hardware license was created to govern the rules of sharing, distributing and selling the WR designs. Very early on, GSI Helmholtzzentrum für Schwerionenforschung, a large-scale accelerator facility in Germany, joined the development together with two companies that started contributing to the WR hardware, gateware, and software development. Motivated by the purposive engagement of CERN, a larger group of companies and research organizations joined (31 at present) to progressively shape a diverse and vibrant ecosystem of organizations that contributed to the development of WR components. Since its beginning, the number of contributors that joined the WR community has grown beyond any expectation and has surpassed CERN´s capability of keeping track of the different applications of WR, reuses or adaptations. Although the initial intentions of CERN were to evolve the General Machine Timing (GMT) protocol, by deciding to develop the technology as an OSH, it eventually grew into a "multilaboratory, multicompany and multinational collaboration developing a technology that is commercially available, used worldwide, and incorporated into the original PTP" (Lipinski et al. 2018 p.2)

### *6.4.2 Data Collection and Sources*

Our study relies on a diverse set of primary and secondary data to provide richness and enhance the validity of our findings (Alvesson and Sköldberg 2009; Klein and Myers 1999). We collected data across three years (2017–2019) and conducted more than 35 interviews. In addition to the interviews, direct observations were conducted from two study visits to CERN in 2017 and 2018, including the participation in a WR developer workshop. Interviews were

[22] A picosecond is an SI unit of time equal to 10−12 or 1/1,000,000,000,000 (one trillionth) of a second.

chosen on the initial recommendation of the WR lead team at CERN, with subsequent recommendations from the interviewees. Our objective was to interview a representative cross-section of the WR community. The interview process was concluded when no significant additional insights were obtained from the data and theoretical saturation was achieved. The major themes in our interview protocol are summarized in Appendix B. In our results (section 4), we present interview excerpts from the study, with alphanumeric key identifiers (corresponding to Table 2) that represent quoted interviewees.

Secondary sources included information retrieved from the WR repository and Wiki, which contains general information about the WR project (i.e., newsletters, a list of companies involved in the WR ecosystem, and presentations and reports from workshops), information about WR technology (i.e., synchronization, data delivery and standardization in IEEE1588-2008) and the WR system (i.e., the switch, master and node), a list of users of WR technology, and information about the open hardware license. We also gathered data from the websites of WR users and suppliers, the research project websites that have integrated WR, social media and academic publications. Table 2 presents the details of each of these sources.

### *6.4.3 Data Analysis*

We performed a three-stage inductive analysis by relying on established procedures for inductive research (Miles and Huberman 1994). We iterated between data and theory to shed light on emergent themes and constructs. The first stage was devoted to reading the abundant material available online about WR. We produced brief summaries that moved from technical descriptions to managerial inferences. Second, in-depth interviews were conducted to understand the primary agents involved in the WR community, their contributions to the technology development and how the development was organized since its inception until present. Three rounds of interviews were implemented in this process, as Table 2 describes.

We iteratively analyzed the interview transcripts by coding relevant observations and contrasting them with our analysis of secondary sources. Data was coded by one of he co-authors and it was progressively discussed with the other co-author, especially when the categorization was unclear to reach an agreement. We generated research memos that synthesized the emergent themes identified in the analysis and compared them with prior research. Finally, we confronted the empirical data with theory. Table 4 and Appendix C provide a detailed description of the progression of our empirical analysis towards the theoretical constructs.

To validate our findings, we applied the respondents validation (Miles and Huberman 1994) by sharing our initial findings with the participants of the study and the WR community. The preliminary results were presented at a workshop on 6-7 October 2018[23] to an audience of 56 participants of the WR community to gather feedback about the main results of the study. Additionally, a draft was shared with the interviewees to solicit their feedback and identify gaps in the technical details of WR technology and development history. Finally, we triangulated the results with an independent study performed in parallel in October 2018 based on a text-mining analysis of the WR Repository and WR community exchanges. This parallel study informed the sequential data collection phases by helping to identify new contributors to WR development and by disentangling the separate developer contributions to WR components.

[23] The workshop information is published at https://www.ohwr.org/projects/5/wiki/oct2018meeting.

*Table 2. Details on the data collection and use in the analysis*

| **Source of data** | **Type of data** | **Description** | **Identifiers** | **Use in the analysis** |
|---|---|---|---|---|
| *Interviews* | *First Round*<br>n=18 | Research scientists and engineers in research infrastructures | RSE1 (2 interviews)<br>RSE2 (2 interviews)<br>RSE3 | To gather data and an overall understanding of the process, different phases, agents and actions in WR development. |
| | | Personnel at the technology transfer offices of the research infrastructures | RT1<br>RT2<br>P1 | |
| | | Other staff in research infrastructures involved in WR development | R1<br>R2 | |
| | | Companies developing software | CS1 | |
| | | Companies developing hardware | CH1 (2 interviews)<br>CH2<br>CH3 | |
| | | Companies implementing pilots of WR with different customers | CD1<br>CD2 | |
| | | Customers of WR not involved in WR development | CA1 | |
| | *Second Round*<br>n=13 | Research scientists and engineers at research infrastructures | RSE4<br>RSE 5 | To gather data on how work was organized and coordinated in the WR development process within the different phases identified. |
| | | Other staff in research infrastructures involved in WR development | R3<br>R4<br>R5 | |
| | | Companies developing software | CS2 | |

|  |  |  |  |  |
|---|---|---|---|---|
|  |  | Companies developing hardware | CH4 |  |
|  |  | Companies implementing pilots of WR with different customers | CD3<br>CD4 |  |
|  |  | Customers of WR not involved in WR development | CA2<br>CA3<br>CA4<br>CA5 |  |
|  | *Third Round*<br>n=4 | Research scientists and engineers in research infrastructures (n=3) | RSE1<br>RSE 5<br>RSE 6 | To verify the interpretations and provide increasing detail on WR-specific components and each development model (i.e., hierarchical control, contractual agreements and voluntary contributions) per component used over time. |
|  |  | Companies developing hardware | CH1 |  |
| *Observations* | *First Visit* | April 2017: Visit CERN to see infrastructure, timing systems department, technology transfer office |  | To gain additional understanding about WR contributors, users, the interactions among them and how they organize the work. |
|  | *Second Visit* | October 2018: Workshop on WR with more than 56 participants.  Presentation of preliminary insights about the study to gather feedback from participants. |  |  |
| *Secondary data* | Repository | 5,076 commits,<br>36 developer members |  | To gather data and obtain an overall understanding of all WR technology, its components, interdependences, cycles of development, different component versions, meetings among contributors, and main events in WR development. |
|  | Wiki | Documentation about<br>- WR Technology: WR Switch; Master (Data, Timing); Node (WR PTP Core); WR good practice guide; Calibration (default parameters for WR switches/nodes, procedure); data-delivery; synchronization; Standardization in IEEE1588-2008; and a Frequently Asked Questions section.<br>- WR Users: 30 users of WR and 16 evaluating the technology (documentation about the organizations, descriptions, and presentations)<br>- WR Projects: 13 publicly funded projects using WR |  |  |
|  | Newsletters | 5 newsletters (2013, 2014, 2015, 2018) |  |  |

| | | | |
|---|---|---|---|
| | Meeting Minutes Published | 10 meeting minutes (2008- 2018) | |
| | Workshop | 10 Workshops (2008- 2018)<br>1 Developer meeting (2010)<br>2 Tutorial WR workshops (2017, 2018) | |
| | Blogs/Websites | 43 websites of users and projects | |
| | Publications | Presentations (n=64)<br>Papers (n=53)<br>Master thesis on WR (n=2)<br>Posters (n=2)<br>Demos (n=3 in 2010 and 2013)<br>Training material (n=2 in 2013 and 2016)<br>Test reports (n=18) | |
| *Other data* | Parallel WR Study | Text-mining analysis of an independent study implement by another researcher studying WR | To triangulate facts and observations regarding WR development with the analysis of commits and contributors. |

## 6.5 Findings

### *6.5.1 Isolating the Attributes of a Hybrid*

We decomposed WR into its different components to identify their specific attributes (see Figure 2 for a graphical representation of the components). Table 3 provides a description of each WR component with excerpts from the interviews and data to substantiate their attributes. For each component, we qualify a) *embodiment* – dichotomously as material or non-material – and the four additional attributes of b) *modularity,* c) *granularity,* d) *editability* and e) *reproducibility* as a matter of degree, that is, high, moderate or low.

*Figure 1 WR representation of switch synchronization hierarchy (Moreira et al. 2009)*

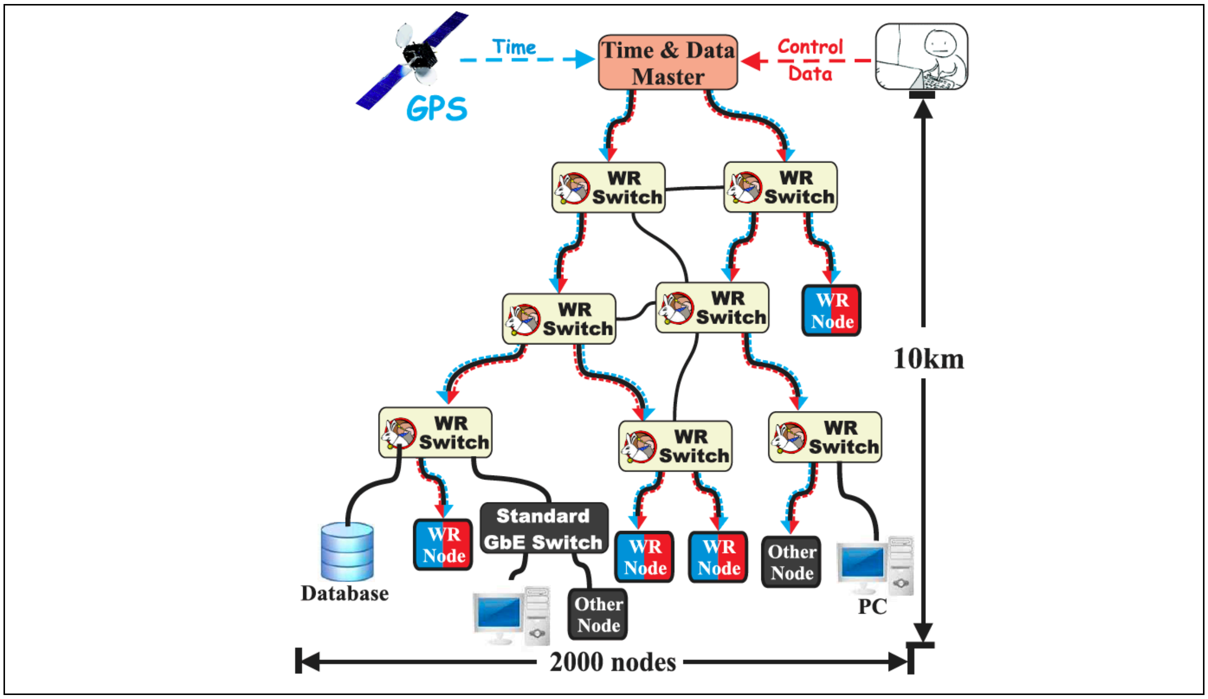


*Table 3. Data analysis and theoretical constructs*

| **WR components** | | | **Illustrative examples of empirical observations and interview excerpts** | **Theoretical observations** | **Theoretical constructs** |
|---|---|---|---|---|---|
| **Software** | **Switch** | Dedicated switch software | “*Switch is like a router with a very precise timing. Nodes is a distinct component but WR is everything, and how you implement both: it has a specific circuit that allows implementing both*” CH1<br>“*It is much faster to develop the software than anything else. So, for instance, they could not test it until we managed to get a prototype of the hardware*” RSE5<br>“*The development cycle was longer for the hardware than the software because if we lose something, we could compile it. It takes us a second, and then you test if it works. The test cycle is very fast*.” RSE 1<br>“*I was working in the protocol and the PTP itself, whereas A was working on some hardware; then, I was also working on some gateware parts of the switch and then B was integrating all together. B was not coordinating; he was integrating and taking different inputs and trying to make them work. For example, the software that interacts with gateware and hardware needs to speak with one and the other. You still need to integrate them” RSE6.*<br>*"It is a multilevel process. For the software, it required to be integrated with the hardware. Tests for each of the components, and as soon as they were integrated, we run other tests. The testing is done by the same developers; we do not have a separate team for testing*” *RSE6.* | - No major differences between the software for the switch and the node<br>- Development of WR software was faster than for the other components<br>- Easy to reproduce with other developers contributing<br>- Loosely coupled, different developers, contributing in parallel to the general and dedicated software for both the switch and nodes<br>- Highly granular enabling two companies, developers at the sponsor organization and other distributed voluntary developers to contribute in parallel<br>- WR software was constantly edited and generated more than five prototype versions | Non-material<br>Modularity (high)<br>Granularity (high)<br>Editability (high)<br>Reproducibility (high) |
| | | at91bootstrap-3.3 | | | |
| | | barebox-2014.04 | | | |
| | | Linux-3.16.38 | | | |
| | | buildroot-2016.02 | | | |
| | | General-purpose software – used both in the switch and node | | | |
| | **Node** | General-purpose software – used both in the switch and node | | | |
| | | Dedicated software for the node | | | |

| | | | | | |
|---|---|---|---|---|---|
| **Gateware** | **Switch** | General-purpose gateware – IP Cores used both in the switch and node | “*Gateware and software it is easier to split it among companies, but hardware does not make sense*” RSE6<br>*"We had different actors working in parallel. I was coordinating the contributions that came from gateware Y that was integrating everything together. In the beginning, we had two companies helping with the software and gateware, the other two for the hardware. X was integrating everything" RSE6.*<br>*“We are now developing gateware for new designs of the nodes, so we are supporting different applications of the nodes because it depends on each application” RSE6.* | - No major differences of the gateware for the switch and the node<br>- Development of the gateware was split among different contributors<br>- Easy to reproduce with other developers contributing<br>- Components more tightly coupled compared to the software of the switch and node; different developers were contributing, but most work is performed by a core group<br>- Gateware was edited but was more stable; three prototype versions | Non-material<br>Modularity (moderate)<br>Granularity (high)<br>Editability (high)<br>Reproducibility (high) |
| | | Dedicated switch gateware | | | |
| | | Module specifications: hdlspec | | | |
| | | Gateware-software interface | | | |
| | **Node** | General-purpose gateware – IP Cores used both in the switch and node | | | |
| | | Dedicated gateware for the node | | | |
| **Hardware** | **Switch** | Electronics & Mechanics | “*If we say a switch, we think about a hardware box” RSE1*<br>“*The switch has 18 ports, and it is a completely different functionality. It has to forward data between ports. The switch looks like 18 ports that are interconnected. It implements more standards and because it is a generic device and needs to allow different configurations. You implement many more protocols than in the nodes. Basically the switch is much more complex than the node because it is 18 times the node; plus, each of the ports needs to interact between themselves; plus, you need to implement more flexibility because it needs to allow different types of configurations; plus, you need to implement more* | - Highly stratified (many layers) but with high interdependencies among layers for efficacy and performance issues (more granularity is associated with less time accuracy)<br>- Low granularity – entity block; more granularity translates into inefficiencies and lowers the switch performance.<br>- Low modularity – components tightly coupled; splitting the switch translates into more standards to be applied to relate one to | Material<br>Modularity (low)<br>Granularity (low)<br>Editability (moderate)<br>Reproducibility (low) |
| | | WR Switch Box: It is a white metal 19” 1U case with two cooling fans in the back | | | |
| | | Main PCB: It contains the main electronics components, ARM processor, Xilinx FPGA chip, | | | |

| | | | | | |
|---|---|---|---|---|---|
| | | oscillators, memories, etc. | *features let's say*" *RSE6*<br>*"It would not work to decentralize the WR hardware design. It needs to be one company that controls the design for one device. For software, you can have different people working on different parts. On gateware, it works decentralized like the software, but for hardware, this is not possible, especially for the switch. I do not see it working" RSE6.*<br>*"The switch is quite a compact device; it needs to work like one unified device, and if you have different companies, you need to define different interfaces between the parts that they are designing, test each part, see that they work together, and you make it much more complex, too much work and much more expensive. For the precision, it is also better that you do not have so many connectors; here, it was no practical" RSE 1.*<br>*"For example, when I was working on one IP core in the switch, I did test it alone, I gave it to X who was integrating everything and then we were debugging" RSE6*<br>*"You could not make it more granular; it would make it many costs and extra work and less efficacy in terms of precision. It would be harder to make it work" CH1* | another, less accuracy as it is lost across modules, more effort to split and tightly coordinate work across teams, more costly<br>- High stability of the switch version – the actual version is very similar to the first version in 2012; very stable as the development cycles were too long and very costly in terms of prototyping and testing<br>- Very difficult to reproduce; requires manufacturing companies with engineering expertise to reproduce each unit (the marginal cost is significantly higher than 0) and distribute them | |
| | | Backplane PCB: It contains electrical connections to 18 SFP cages, debug USB-UART ports, LEDs, etc. | | | |
| | **Node** | WR PTP core | "*If we think of a node, we think about an IP core that you can instantiate in different hardware*" RSE2<br>"*Node is an end device whether it receives or sends staff to one port. You throw or you digest the data. It is like one of the switch ports; plus, you need to implement, like in the switch, WR protocol*" RSE 6<br>"*The development of the node it is easier than the switch*" CH1<br>"(Referring to WR hardware development) *In hardware, it can* | - Less stratified than the switch; it is 1/18 times the switch (fewer layers)<br>- Heterogeneous instantiations of the nodes depending on the context of the WR implementation (e.g., sea, altitude, pressure, etc.)<br>- Moderate editability as the number of versions of the node is associated with all | Material<br>Modularity (low)<br>Granularity (low)<br>Editability (moderate) |

| | | | | | |
|---|---|---|---|---|---|
| | | | *take us weeks or months to do the same for software. This is the same for open or not open stuff. It costs you a lot to do a prototype in terms of money, time; you need to wait for it and test for it. It is much more difficult: more expensive, more time, more difficult. Once you do it, you do not change it very often*" RSE1<br>“*The node is different because the first node was a spec board and designed here, and then one company developed a simplified version. Some people took this design and made different formats, and this was without ourselves doing it, we did not pay for the design, it was because people needed it*” RSE 6 | the diverse WR implementations<br>- Low reproducibility requires manufacturers to source prototypes and distribute them<br>- Low modularity – splitting the node is ineffective; each version of the node had a core group of developers | Reproducibility (low) |

Some general patterns should be noted. For the material components of WR, specifically, the hardware layers of the switch and node, we found a) *low modularity* and b) *low granularity*. Where the switch was more stratified compared to the node, both had high interdependencies among the layers required for timekeeping precision; additional layers reduced accuracy. Likewise, the components could not be split into more granular parts as this would generate interoperability problems across component interfaces and more standards to implement, which would further impede chronological precision. Additionally, for the switch and nodes, we found c) *moderate editability*; multiple developers can act on the design of the switch and the node and modify them but to a lesser degree than the software components. Although the switch was a more stable technology, the node offers greater editability, and as a result, there are five times more versions of the node compared to the switch. In addition, we found d) *low reproducibility*, with an average cost per node in the range of 1,500 dollars and the switch approximately 10,000 dollars. The development process involved physical prototyping that must be manufactured and acquired to test its performance, with reproducibility as an important attribute. As CH1 describes, "*If someone wants to use WR, I need to manufacture it. However, in software, if I need to modify something, I do not incur NRE [referring to non-recurring engineering costs] because I modify [and] compile and users or any developer tries it, but in WR, I need to manufacture another prototype, and these costs are neutral for me; I need to incur costs in electronics. This process of manufacturing has additional costs, and of course, if I need to sell it, I need to certify it to ensure that it is safe, and this has important additional costs*".

Across the software and gateware layers of WR, we found *high levels* of *a) modularity b) granularity c) editability* and *d) reproducibility*, as described in Table 3 through our data, which made the development of such layers faster and the distributed contributions easier to organize. As RSE1 explains, "*The development cycle was longer for the hardware than the software because if we lose something (in the software), we could compile it. It takes us a second, and then, you test if it works. The test cycle is very fast*". As RSE6 further describes, "*Gateware and software it is easier to split it among companies, but hardware does not make sense*".

### 6.5.2 Three Phases of the Evolution of WR Development

The hybrid model for developing WR underwent ***three main phases***, where the hierarchical, contractual and voluntary contributions varied over time (see Table 4).

The first *phase* (from 2008 to 2012) began with the project launch in 2008 and concluded when WR achieved the first version of the switch and the node in 2012. WR was launched as an OSH project, a decision that is consistent with CERN's traditional operational philosophy and raison d'être. As Bij et al. (2013) explained, "OSH also fits CERN's role of transferring the technologies it has developed to industry and to stimulate industry with innovative products such as the WR*"*, p.7. As RSE 5 further describes, "*To develop WR as open source would help us to get specialist knowledge, where we know that small companies play a large role; but on the other hand, we would need to support them and help them to achieve the quality we need. Companies benefit from that process because it helps them improve and produce better hardware*".

However, given that WR would be a sophisticated technology conceived for a very specific purpose and due to the different interdependencies and highly integrated nature of the switch and the node, it quickly became evident that the design of the first versions of both the switch and the node needed to be controlled and directed by the sponsor, CERN. As stated by one hardware developer,*"CERN was our grandfather — not only when we were developing WR together but also in the first moments when we were wondering what was next"* (CH1).

Contractual arrangements with two hardware suppliers and two software suppliers enabled the development of the first prototype, which required tight coordination among tasks and development teams. The development followed the four major sequential and consecutive steps of 1) requirement and specifications, 2) design, 3) implementation, and 4) testing, where the outcomes of each step were highly dependent on the results of the previous step. Although some parallel and asynchronous development was conducted by peripheral research organizations, overall, the development followed planned and sequential phases with *low voluntary contributions*. This phase is characterized by a predominance of *hierarchical control* by the sponsor. Beyond fulfilling the technical requirements of the switch and node design, the strong initial protagonism by CERN can be viewed as part of an initial community-building phase to build interest and confidence in the project. As stated by a CERN engineer, *"After allocating some funding for the first companies to join, we needed to convince others to invest in developing WR* [as voluntary contributors]*, which was not an easy task. I had to reach out proactively to the companies we knew could do it and convince them. We needed first, to select companies that had not only the expertise but also the capacity afterward to provide support to the WR product"* (RSE 5). The initial reliance on hierarchical control supplemented with several key contracts served as a signaling mechanism: *"We knew CERN was serious this time by engaging firms, and this also sent a message to other organizations* [as voluntary contributors] *that could collaborate with us"* (CH2). These initial signals were determinative in convincing voluntary contributors to later join in the development cycle: *"Other organizations and potential users joined and agreed to invest in the development because they saw other companies developing WR, and they know* *that those companies will actually be able to provide the technology once the R&D process is finished"* (CH2). CH1 further explained the implications of such engagement for companies: "*the difficulty of WR, but this is the same for any other OSH, is that it needs to be manufactured along the process. That means that there are additional costs. For example, you have the additional costs of qualification to prove that it works and later on in the process, to certify that the design works, and this is an overhead that you do not have in the software layers of WR or any software (…) and if I need to introduce any modification, then I need to start all over again".*

The *second phase* (from 2012 to 2015) began with the first WR prototype release. At this point, first users began implementing WR and reported bugs, whose fixes were incorporated into further designs. Novel instantiations of the node began to appear based on the unique requirements of the installations of other scientific research infrastructures. As such, this phase has *high voluntary contributions* and minor changes to the design of the switch but many new designs and configurations for the nodes. This phase is characterized by the low direction-giving by the sponsor (the switch was stable) with many new designs of the nodes emerging from a growing WR community.

Some extraordinary examples of WR implementations that lead to new node designs and switch modifications include meteorology research institutes that need to transfer time from atomic clocks over distances up to 1,000 km, the neutrino telescope KM3Net located in the deepest seas of the Mediterranean, and a five-cubic kilometer Cherenkov submarine detector in Toulon (France), Sicily (Italy) and Peloponnese (Greece). At 4,410 meters above sea level, China built the Large High Altitude Air Shower Observatory (LHAASO), the world's largest and most sensitive cosmic-ray observatory for gamma-ray astronomy, which consists of more than 6,300 detectors and 12 telescopes. Four layers of WR switches (583 in total) covered 7,344 nodes of a Square Kilometer Complex Array (KM2A) detector and a Water Cherenkov Detector Array (WCDA) (White Rabbit wiki).

The *third phase* (from 2015 to 2020) began when WR started a standardization process to guarantee the stability of the technology, which raised awareness about the potential of WR across industries. In this phase, WR reuse and implementations emerged in telecommunications, financial services, smart grids, air traffic control, electronics and industry 4.0 applications. As a result, new versions of WR switches and nodes were developed as proprietary applications and not disclosed to the WR community. *Voluntary contributions* by the senior contributors to WR (both companies and peripheral research organizations) were balanced in this phase by proprietary contributions to the switch and node designs. *Hierarchical control* was exercised by the sponsor, not in development, but in the standardization of the core technologies along with the coordination and aggregation of the WR community contributions.

In this phase, we find a growing number of increasingly heterogenous adoptions by industry. Examples include Vodafone, which conducted a successful proof of concept in 2017 to distribute accurate timing through the live Vodafone network where time was measured with a surprisingly small error of less than one nanosecond over a cascade of four sites that spanned a total distance of 320 km. As reported in the WR wiki, "*Needless to say that this result builds strongly on the outstanding work delivered by the WR community over the past years, and we are thankful to all of you who contributed. We are absolutely convinced that with WR, you have created a game changer that will enable marvelous new technologies in the future!*"(JK, The Netherlands, 9/6/17). In financial services, the Frankfurt Stock Exchange implemented WR because it needed a time synchronization technology superior to the current standards of NTP and PTP. As CH1 describes, "*Financial transaction organizations are required by law to prove that the time reference used for stamping transactions is UTC [Coordinated Universal Time] traceable. Thus, the accuracy required is in the millisecond range, but WR allows the nanosecond range with high accuracy, allowing legal timestamping applications.*" Similar implementations at other financial exchanges are appearing in the media, as *"The financial industry has easily become the most obsessed with time" (Markoff 2018 p.1)*. Table 5 provides detailed information on the adopters of WR throughout the three main phases.

*Table 4. A hybrid model of development over time*

| | Phase 1 – Design process (first prototypes) | | | | Phase 2 – Users/developers join for testing and design new versions | | | | Phase 3 – Applications outside the scientific industry, forking, and parallel proprietary developments | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 |
| **Phase Characteristics** | Trigger: WR designs kick off with one supplier supporting WR specifications<br>• Design of the first version of the switch and node controlled by CERN (1 manager, one integrator, two coordinators of components)<br>• Contractual agreements with HW and SW suppliers to allow a first prototype to emerge that required strong coordination among tasks and teams<br>• Low voluntary contributions that include few research infrastructures<br>• High direction provided by the sponsor was given to the design | | | | Trigger: First commercial WR prototype<br>• First users of WR contribute reporting bugs to the switch, which impact further switch designs<br>• High voluntary contributors to design multiple versions of the node, conditional on the different applications of WR<br>• Contractual agreements for WR for *production*<br>• Low direction given by the sponsor as the switch was stable, while there was high generativity as the new designs of the nodes were shared in the repository as the WR community was growing | | | | Trigger: Standardization process of WR raises awareness across industries<br>• First implementations in other industries (e.g., financial services, telecommunications, etc.)<br>• New proprietary versions of WR switch and nodes conditional on the particular setting emerge; designs were developed inside the organizations and not disclosed to the community<br>• Voluntary contributions are balanced by proprietary contributions to the switch and node designs | | | |
| **Events** | WR workshop 1. Project start. | Review of Switch MCH card v1. | Switch MCH card v2 PCB ready. | 4th WR Workshop. PTP working on a WR node. | Contract with supplier for assemblage (production of prototypes)<br>CERN received 4 WR V3 | WR Starting Kit available. | 8th WR Workshop, CERN, Geneva (Switzerland). | The standardization process in IEEE with many users of WR. | Release 5.0 of White Rabbit switch software is out. 45 improvements and extension | The second producer of White Rabbit switch hardware joins. | 2nd WR Tutorial Workshop, Beijing (China). | WR officially standardized. |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | switches. | | | | s made. | | | |
| | White Rabbit workshop 2. | White Rabbit workshop 3. Demonstration of a stabilized link. | Report on measuring propagation delay in FPGAs builds groundwork. | WR Demonstration package documented (SPEC-Switch-SPEC). | 6th WR Workshop. | WR switch V3.3 passes all CE marking tests (power, electrical safety, labels, EMI). | EISCAT, Sweden implements WR. | Release 4.2 of White Rabbit switch software, with many improvements and SNMP control. | EPFL-STI-IEL-DESL-ELL, Distributed Electrical Systems Laboratory implements WR. | More users of WR-Borse Frankfurt; Vodafone. | The first major beam time using WR-based timing system at GSI (134 WR nodes, 32 WR switches). | A new working draft of the open-source hardware license. |
| | Diligent requirements collection through CERN-supplier contract-to gather joint (cross-organizational) specificati | Contract to create a web-based Open Hardware Repository portal CERN-supplier contract.<br>Contained: file repository | WR switch PTP compliance tested in ISPCS 2010 plugfest. | 5th WR Workshop. | The first deployment of a system based on WR synchronization technology was successfully done in May 2012 in Gran | GSI, Germany received the first production batch of V3.3 switches. | Istituto Nazionale di Ricerca Metrologica (I.N.RI.M.), Italy implements WR. | CHIRON-IT, The Netherlands implementation of WR. | Fermilab, USA implements WR. | MIKES, the center for metrology and accreditation of Finland, has connected the Metsähovi Geodetic | 10th WR Workshop, Geneva (Switzerland). | New users of WR. |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ons. | , a wiki, and general project documentation. | | | Sasso and CERN for the CERN Neutrino to Gran Sasso (CNGS) measurements. | | | | | Research Station to the official time of Finland, UTC [MIKE]. | | |
| | Contract with supplier for WR switch hardware design (e.g., cards). | Provision of an open repository to allow community collaboration. | First developer meeting. Basic Ethernet switching demonstrated. | | WR presented at IEEE 802.1 (AVB Gen 2) meeting in view of standardization. | CERN, Switzerland received 10 V3.3 switches. | LHAASO<br>(Tsinghua University,<br>China) implements WR. | Culham Center for Fusion Energy (G.<br>Naylor, S.Hall, B.Huang,), UK<br>Implements of WR. | KM3NET (15<br>countries) implements WR. | 1st WR Tutorial Workshop, Barcelona (Spain). | IMPCAS - The Institute of Modern Physics (IMP) of the Chinese<br>Academy of Sciences, China. | Bolsa de Madrid (Spain) – WR pilot. |
| | Contract with supplier for WR switch software and gateware. | | *Plugfest* to check interoperability with other companies' hardware | | Mini WR network with two SPEC cards documented (SPEC-SPEC). | China Spallation Neutron Source Institute of High Energy Physics | | DLR (German Aerospace Center, Institute<br>for Technical Physics), | LNE-SYRTE: the French National<br>Metrology Institutes | Vodafone proof of concept of WR in the Netherlands. | Baikal Deep Underwater Neutrino Telescope -<br>BDUNT, | KRISS (Korea) implements WR. |

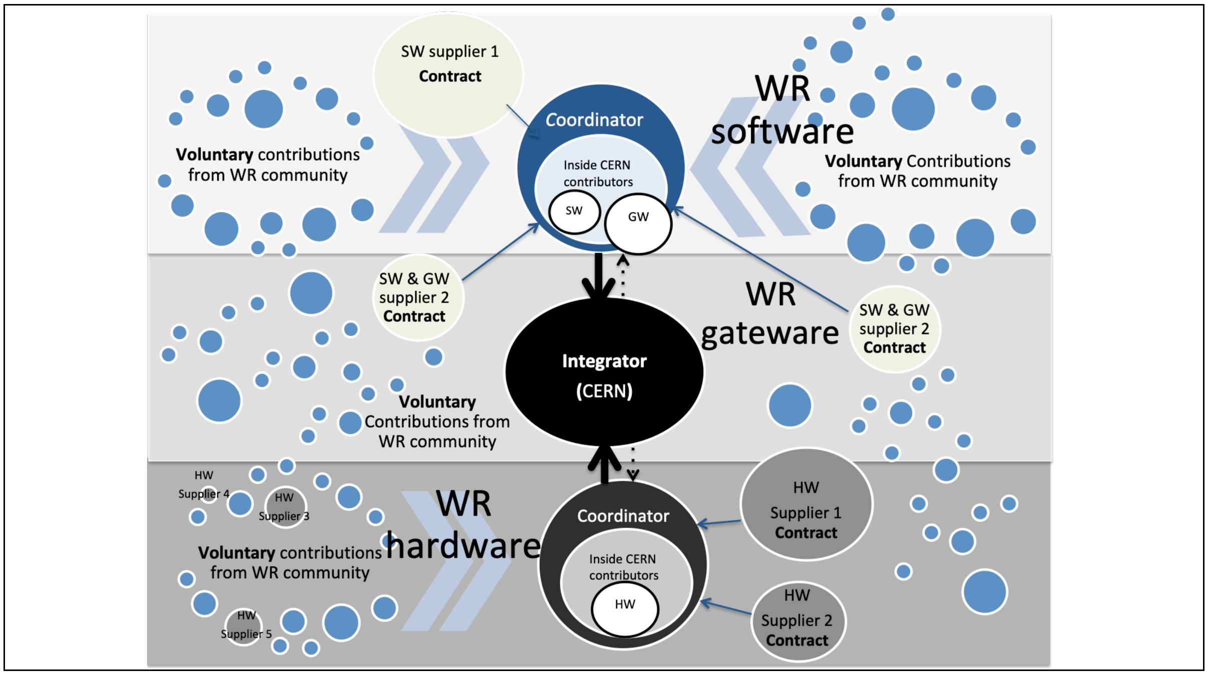
| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | designs and showcase WR to attract developers. | | | Chinese Academy of Sciences –CSNS implements WR. | | Germany implements WR. | (2015-), France. | | Russia implements WR. | |
| | | | | | WR PTP extension proposed to be included in the revision of IEEE PTP standard at ISPCS. | CNGS<br>Timing for neutrino measurements implements WR. | | ELI-ALPS, Hungary implements WR. | Struck Innovative Systeme proprietary WR hardware. | Frankfurt Stock Exchange - proof of concept. | Frankfurt Stock Exchange’s deployment of WR. | D-TACQ Solutions Ltd. proprietary WR. |
| | | | | | 7th WR Workshop. It is showing the final production version of the switch. | DESY, Germany implements WR. | | ELI-BEAMS, Czech Republic implements WR. | N.A.T. proprietary WR hardware. | INFN di Tor Vergata, Italy implements WR. | | Picoquant proprietary WR . |
| | | | | | CTA - Cherenkov Telescope Array, | MIKES (Center for metrology | | ESRF, France implements WR. | Struck Innovative Systeme | Ettus Research proprietary WR . | | SyncTechnology proprietary WR. |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | 32<br>countries implementation of WR. | and<br>accreditation, Finland) implements WR. | | | proprietary hardware. | | | |
| | | | | | ESS Bilbao, Spain implements WR. | Vrije Universiteit Amsterdam, Dept of Physics and<br>Astronomics, The Netherlands. | | JINR, Russia implements WR. | Teledyne SP Devices proprietary WR. | National Instruments proprietary WR. | | |
| | | | | | | | | Paris Observatory<br>CNRS-INSU-IN2P3, France implements WR. | | Sundance Multiprocessor<br>Technology proprietary WR. | | |

### *6.5.3 WR Development: A Hybrid Model for a Hybrid Object*

When CERN decided to develop WR technology as OSH, it soon became evident that the very specific function, sophistication, and interdependent nature required that the core technological design be established before significant voluntary contributions could be incorporated. As a result, traditional hierarchical and contractual mechanisms were employed towards greater directionality and control in the initial development phases. After the core technologies were developed and stable, an increasing amount of voluntary contributions were incorporated at all levels of the WR technology stack. We label this as a *hybrid model of development*. The tightest hierarchical controls were enabled through employment in traditional engineering and development companies with CERN as sponsor and integrator. Moderate control was enabled by formal contracts that enabled the development of specialized components that were beyond the scope of capabilities for the core WR development organizations. Voluntary contributions originated in the WR community, with some centralized coordination and integration by CERN. This hybrid model is depicted in Figure 2.

*Figure 2 Overview of WR development organization*



#### Hierarchical control

The most central function for CERN as the sponsor and principal user of WR was to provide sponsorship, legitimacy and centralized control to the project. This means the coordination of the initial specifications of WR across both contractual partners and voluntary contributors, the internal management of the more complex and interdependent WR components (the hardware), and the aggregation and integration of the voluntary contributions. Given the heterogeneity of these diverse contributions, CERN put a team in place to orchestrate WR development. As RSE explains, "*A was coordinating the contributions that came from gateware, B from the software and C of the hardware of WR switch and node. D was integrating everything together. In the beginning, we had two companies helping with the software and gateware, the other two for the hardware. Each of us was coordinating the contributions of the company and the ones coming from other organizations. D was integrating everything* […] *and E was in charge to coordinate everything as part of the department role more from a management perspective*". As CH1 claims, "*WR worked essentially because of the leadership of CERN*".

The need for a differentiated approach based on component attributes is evident in the following quote where RSE 6 explains how the hybrid model worked: "*There was internal work at CERN, different work at companies and then other contributions by other organizations that voluntarily joined and contributed, and all this work was coordinated and integrated at CERN*". RSE 2 further justifies that "*Whenever we want to have something done, we put a contract. Volunteer contributions are nice, contributions for free we accept them as developed packages, it shows that it works, and then we integrated them in the switch. However, when you need something specific, and if you do not know if it works or if it will not work, we need to control it. For software, you can be a one-man company, whereas for a company that develops hardware, you need licenses, expensive equipment, and before you get paid for what you do, you need to send it for production, which costs money, and for prototyping, which costs money again, and this does not translate into software*".

**Contractual agreements**

The contractual agreements employed were clustered around the following four main activities: 1) contracts awarded to companies to gather and manage WR *specifications* across the WR development community; 2) contracts to develop the repository and main hub for WR collaboration; 3) further contractual arrangements to contribute to the first switch and node prototypes across the software, gateware and hardware components; and 4) contracts for prototyping, where manufacturers were asked to produce a few units of WR switch and nodes and distribute them across the community for testing. All of these contractual arrangements specify that all documentation that results from the development must be shared in the repository and are governed under an open-source license.

An interesting facet of the contractual agreements was that many vendors included voluntary contributions as part of their deliverable. In these instances, complete documentation also included contracting partners' efforts in supporting other voluntary contributions to WR development. That is, if their component included volunteer contributions from the WR community, they were equally responsible for this. As RSE4 explains, "*You have to be ready to document and publish everything. Support may take more than you want*".

### Voluntary contributions

As WR deployments increased beyond the original scope of scientific research infrastructures, a more heterogeneous community of WR users engaged in developing the software, gateware and hardware to customize it to the specific operational requirements of their diverse applications. A portfolio of tools common to OS initiatives was used in WR development that facilitated the customized applications in addition to the standard WR layers. These tools included documentation wikis, issue tracking, dedicated mailing lists, peer review over email, regular face-to-face meetings, dedicated workshops, and proprietary tools to allow distributed hardware development such as electronic design automation and field-programmable gate array (FPGA) development tools.

The voluntary contributions were diverse. Some were focused on the core technologies, whereas others emphasized the more peripheral aspects of the nodes and software. Some volunteers contributed to testing, where other volunteers participated in WR OSH communities with the explicit purpose of cultivating skills that could be monetized as they worked with their own clients. As RSE 6 shares about one organization that voluntarily contributed, "*X´s contribution to WR was a measurement of one of the key things that WR used to reach the nanosecond. They invented this.*" Another example of a voluntary contribution is "*Y was contributing from the very beginning although it is hard to point to one thing. They were contributing to some modules in the switch to some extent*" (RSE 3). In some cases, the voluntary contributions were related to testing the first WR prototype in 2012: "*When organizations started using WR, they started to find bugs and were doing bug reports but not a development of some kind. All of these bug reports resulted in new releases of the switches, and people [and] other research infrastructures helped developing new releases"*. Other companies contributed to WR communities to learn: *"Other organizations contributed to discussions. Minimum effort contributors but their business idea was to contribute to the discussions so that they could be the first to use WR in case they had the first client to make sure that they could use it"* (RSE 5).

In addition, a set of regulatory devices, such as the creation of a new open-source hardware license, were also put in place to agree on codified norms across organizations and ensure the stability of the core technology. This element is particularly important given that whether via contracts or via voluntary engagement, all contributors in WR development are organizations. As RSE 6 explains, "*We always find companies in open hardware ...it* *depends on the type of hardware. If it is simple hardware, then you will find individuals with tools that allow simple designs, but for designs that are complex such as WR, only companies and organizations [participate] because the tools cost many money.*" Previous research in technology ecosystems and platforms has emphasized the importance of standards and disciplined versioning of core technologies as insurance of a fair economic return on investments by implementors, re-sellers, and complementors (Wareham et al. 2014).

## 6.6 Discussion

### *6.6.1 Theoretical Implications*

First, our study contributes to recent conceptualizations of digital objects by uncovering the differences from hybrids to purely non-material digital objects. The study of WR identifies that the physical nature of the components of hybrid objects deviates from the attributes commonly associated with digital objects in both essence and degree. WR, and by extension, many sophisticated hybrid objects that contain material components, exhibit less *editability* and less *reproducibility* and are less *modular* and less *granular.* As a consequence, the efforts for their decomposition and reintegration are *higher* compared to pure non-material digital objects. As such, the nominal and relative attributes of the hybrid object components of a) embodiment, b) modularity, c) granularity, d) editability, and e) reproducibility have strong implications for how their development is organized (Akhlaghpour, Wu, Lapointe, Pinsonneault, 2013). As our analysis shows, at certain levels, these attributes can inhibit the possibility for development to be completed in conditions normally ascribed to OS. In the case study of WR, this resulted in the coexistence of voluntary contributions combined with traditional hierarchical and contractual models of development.

This rationale can be specified as follows. The development of *tightly coupled* components requires highly sequential processes with intensive coordination and control over the activities. A change of one component may require extensive compensating modifications in the designs of many other interrelated components. Consequently, the development cycles of hybrids are longer, and modifications and one point may require more time and resources given the component interdependencies. Relatedly, the *granularity* of the components is important if the nature of the object does not permit a reduction into numerous independent elements. It follows that the *editability,* or the ability to modify it continuously and systematically, can be lower in hybrids due to tighter integration with the hardware. Where this is not always the case (personal computing devices are an obvious example), many OSH projects are designed to address specific needs that have yet to be fulfilled by mature commercial HW/SW solutions and consequently, will likely have a higher level of coupling between these layers. The material aspect of hybrids can reduce their *reproducibility*, which implies non-negligible production and distribution costs and dissuades voluntary contributions, as volunteers normally need to incur significant expenses related to prototyping and testing. As evidenced in this case, this is very much correlated with the sophistication and economic costs of the technology.

We have re-visited some fundamental ideas of TCE to explain the modalities of developing hybrid objects. TCE certainly does not explain all aspects of our OSH phenomenon, and we clearly do not want to oversubscribe it as a theoretical lens (Fischer 1977). Through the discriminating alignment hypothesis, TCE predicts_governance structures according to the transaction attributes (Williamson, 1996). In exploring the conceptual equivalence between transaction attributes and component attributes, we demonstrate that some TCE logic and constructs are particularly well-suited to make predictive statements about when OSH development will be 1) hierarchically or 2) contractually governed or 3) built on voluntary contributions in a more traditional OS manner. We simplified our use of TCE by collapsing the development models to two major outcomes (hierarchical control and contracts, supplemented by voluntary contributions), but given the novelty of OSH, we believe that this predictive capacity should not be underestimated. CERN chose to develop WR by leveraging the expertise of a significant number of heterogeneous voluntary organizations willing to develop WR as an OSH; however, following a similar logic of the strategic alignment hypothesis, WR component attributes determined the need for a mixed model of development to emerge that combined voluntary contributions with commercial contracts and hierarchical control.

Most likely, the most potent TCE construct in our analysis is *interdependence*, as it envelopes *modularity* as a predictor of development governance. Object components with low modularity are more tightly coupled. This applies to both their physical and logical attributes of all technology layers and, therefore, also to their development process that requires greater linear coordination and centralized control. This logic applies equally to the TCE constructs of *product complexity* and *monitoring costs*, which although they are also strong predictors, they can be more ambiguous as observable attributes of objects or components. In addition, the TCE constructs of *duration* and *frequency* are also useful in that they indirectly relate to the *granularity* of the component. TCE posits that transactions with a shorter duration and frequency are governed by less centralized, simpler market-based governance forms. Similarly, more granular components can be developed in a less-centralized or voluntary mode. The TCE construct of *asset specificity* is somewhat synonymous with editability to the degree that the component can be re-configured for alternative uses (low asset specificity = high editability). As TCE predicts that low asset-specific transactions are governed through less centralized governance, it follows that we found that the more highly editable layers of WR were developed through less centralized contractual or voluntary processes. *Transaction risk* is more of an omnibus TCE construct that refers to the potential economic loss, IP infringement, enforcement costs, or any general legal or financial loss. It follows from TCE that higher transaction risks will be governed through more complex and centralized governance forms. Where these elements were present in our analysis of WR, they did not emerge as a focal concern from our respondents to the same degree. This may be a consequence of the unique culture and social norms that are common in scientific research organizations. Interestingly, TCE has very little to say about the material versus non-material embodiment of object components. This may be because where there is some correlation between the material attributes of the component and its modularity, granularity and integration characteristics, this correlation can vary considerably based on the design and complexity of the object.

### 6.6.2 Practical Implications

Big-science research infrastructures develop some of the most sophisticated technologies in existence. Researchers are currently experimenting with OSH to develop new complex hybrid objects that will find multiple unintended applications in different industries (Wareham and Pujol 2019). In parallel, commercial interest in OSH is growing, particularly for organizations that want to minimize the non-recurring engineering costs of nonexistent technologies or solutions. Although open source is a powerful model that can serve as a low-cost source of frontier technologies, it might need to be supplemented with more traditional commercial development processes at specific points based on the component attributes. A need for hierarchical control and contractual agreements is likely greater for OSH projects that are highly sophisticated, which require unique expertise and greater financial investments. The WR case illustrates how the transposition of the open-source model to WR was possible with the combination of these traditional managerial mechanisms that allow direction-giving and control at specific, necessary phases of its development. Combined with the generative nature of the OS community, they made it possible for WR to be deployed as a powerful precision and synchronization technology in many industrial settings. It will be compelling to follow the emergence of OSH movements in other realms of high-end commercial computing such as the RISC-V movement in integrated circuits.

### 6.6.3 Limitations and Future Research

Our findings are subject to limitations that warrant further investigation. First, we study an extreme case of OSH developed with the sponsorship of CERN. In this regard, WR is non-representative, but it is studied with the goal of understanding something that is likely to become more predominant in the future. Furthermore, for theory generation, it is beneficial to study cases with high values on variables of critical interest. Obviously, we should be prudent in extrapolating our findings to contexts that do not have the same level of technical sophistication, economic resources, and political stature as CERN, as these factors are clearly influential in the case of WR.

As a technology, there are two aspects of WR that are also exceptional. First, as time measurement in the extreme is very sensitive to both the physical and logical architecture of the technology, WR is very tightly coupled at certain points, and this high interdependence between layers clearly influenced its development model. Other OSH projects may not have the same technical sensitivities and may therefore be amenable to a wider range of development modes. Second, WR is not a general use technology such as an operating system or scripting software; it was commissioned with a very specific purpose and is therefore intolerant to significant variance in its performance. Clearly, OSH projects that are more general purpose and not constrained by such rigid outcome requirements might be tolerant of greater scope drift or more organic development processes. It follows, then, that additional research is needed in OSH to investigate different types of hybrids in a wider variety of contexts to further substantiate the relationships between the attributes of hybrid components and multiple forms of development.

There are some parallels between the WR OS community and the literature on platform complementors (Constantinides et al. 2018; Tiwana et al. 2010; Wareham et al. 2014). Specifically, the WR switch is similar to a stable platform core. The OS community behaves like platform complementors that develop more customized implementations at the node, gateware and software layers for specific contexts and thus attract a large number of heterogeneous contributors that pursue their own innovation strategies and commercial goals. As research on technology ecosystems and platforms is currently more extensive than OSH, any identifiable similarities or differences could offer valuable insights.

## 6.7 Conclusion

A nanosecond is roughly the time that it takes light to travel one foot and has long been considered a critical metric in computing (Markoff 2018), even in the era of single-box/single-location computers. Currently, the industrial internet is pushing the adoption of massive sensor data and real-time communications; software, hardware, and data are now scattered over heterogeneous grid, mesh and cloud computing installations. For these geographically dispersed applications, the accurate measurement of time is commensurately difficult – yet critical – in industries that are time-sensitive or even “obsessed with time” (Markoff 2018 p.1).

WR was developed as an OSH to address the distortions created by time latency in CERN’s geographically distributed network. Born as a natively open-source endeavor, WR development was governed through a variety of high, moderate or de-centralized governance, that is, hierarchical, contractual, or voluntary contributions, respectively. The attributes of the object components were clearly determinative in the choice of the development model. Our analysis identified and described these causal relationships and showed how different developmental modalities can co-exist and complement one another towards the development of hybrid objects with diverse component attributes. We further demonstrated how, after the initial sponsorship by CERN, the subsequent WR implementation and adaptation by other scientific infrastructures and industry was possible due to a vibrant open source community capable of customizing and thereby further evolving the more modifiable layers of the WR technology stack.

## References

Adler, P. S. 2001. “Market, Hierarchy, and Trust: The Knowledge Economy and the Future of Capitalism,” *Organization Science* (12:2), p. 20.

Ågerfalk, Pär J., and Brian Fitzgerald. 2008. "Outsourcing to an unknown workforce: Exploring opensourcing as a global sourcing strategy." *MIS quarterly* (32:2), pp. 385-409.

Akhlaghpour, S., Wu, J., Lapointe, L., and Pinsonneault, A. 2013. “The Ongoing Quest for the It Artifact: Looking Back, Moving Forward,” *Journal of Information Technology* (28:2), pp. 150–166. (https://doi.org/10.1057/jit.2013.10).

Allen, R. C. 1983. "Collective invention". *Journal of economic behavior & organization,* (4:1), pp. 1-24.

Alvesson, M., and Sköldberg, K. 2009. *Reflexive Methodology: New Vistas for Qualitative Research*, (2nd ed.), Los Angeles ; SAGE.

Baldwin, C. Y., and Clark, K. B. 2006. “The Architecture of Participation: Does Code Architecture Mitigate Free Riding in the Open Source Development Model?,” *Management Science* (52:7), pp. 1116–1127. (https://doi.org/10.1287/mnsc.1060.0546).

Balka, K. 2011. *Open Source Product Development: The Meaning and Relevance of Openness*, Springer Science & Business Media.

Balka, K., Raasch, C., and Herstatt, C. 2009. “Open Source Enters the World of Atoms: A Statistical Analysis of Open Design,” *First Monday* (14:11). (https://doi.org/10.5210/fm.v14i11.2670).

Balka, K., Raasch, C., and Herstatt, C. 2010. “How Open Is Open Source? - Software and Beyond: HOW OPEN IS OPEN SOURCE?,” *Creativity and Innovation Management* (19:3), pp. 248–256. (https://doi.org/10.1111/j.1467-8691.2010.00569.x).

Benkler, Y. 2002. “Coase’s Penguin, or, Linux and ‘The Nature of the Firm,’” *The Yale Law Journal* (112:3), p. 369. (https://doi.org/10.2307/1562247).

Benkler, Y. 2006. *The Wealth of Networks: How Social Production Transforms Markets and Freedom*.

Ben-Menahem, S. M., von Krogh, G., Erden, Z., and Schneider, A. 2015. “Coordinating Knowledge Creation in Multidisciplinary Teams: Evidence from Early Stage Drug Discovery,” *Academy of Management Journal* (59:4), pp. 1308–1338. (https://doi.org/10.5465/amj.2013.1214).

Bogers, M., Hadar, R., and Bilberg, A. 2016. “Additive Manufacturing for Consumer-Centric

Business Models: Implications for Supply Chains in Consumer Goods Manufacturing," *Technological Forecasting and Social Change* (102), pp. 225–239. (https://doi.org/10.1016/j.techfore.2015.07.024).

Bogers, M., and Horst, W. 2014. "Collaborative Prototyping: Cross-Fertilization of Knowledge in Prototype-Driven Problem Solving," *Journal of Product Innovation Management* (31:4), pp. 744–764. (https://doi.org/10.1111/jpim.12121).

Boisseau, É., Omhover, J.-F., and Bouchard, C. 2018. "Open-Design: A State of the Art Review," *Design Science* (4), p. e3. (https://doi.org/10.1017/dsj.2017.25).

Boland, R. J. 1986. "Phenomenology: A Preferred Approach to Research on Information Systems," in *Trends in Information Systems*, NLD: North-Holland Publishing Co., pp. 341–349.

Bonaccorsi, A., and Rossi, C. 2003. "Why Open Source Software Can Succeed," *Research Policy* (32:7), Open Source Software Development, pp. 1243–1258. (https://doi.org/10.1016/S0048-7333(03)00051-9).

von Briel, F., Recker, J., and Davidsson, P. 2018. "Not All Digital Venture Ideas Are Created Equal: Implications for Venture Creation Processes," *The Journal of Strategic Information Systems* (27:4), pp. 278–295. (https://doi.org/10.1016/j.jsis.2018.06.002).

Constantinides, P., Henfridsson, O., and Parker, G. G. 2018. "Introduction—Platforms and Infrastructures in the Digital Age," *Information Systems Research* (29:2), pp. 381–400. (https://doi.org/10.1287/isre.2018.0794).

Cook, J. 2001. *Open Source Development: An Arthurian Legend*.

Crowston, K., and Howison, J. 2006. "Hierarchy and Centralization in Free and Open Source Software Team Communications," *Knowledge, Technology & Policy* (18:4), pp. 65–85. (https://doi.org/10.1007/s12130-006-1004-8).

Cutosksy, M. R., Tenenbaum, J. M., and Glicksman, J. 1996) Madefast: Collaborative engineering over the internet. *Communications of the ACM*, 39(9), 78–87.

Dahlander, L., and Magnusson, M. 2008. "How Do Firms Make Use of Open Source Communities?," *Long Range Planning* (41:6), pp. 629–649. (https://doi.org/10.1016/j.lrp.2008.09.003).

DeMicheli, G., and Sami, M. G. 2013. *Hardware/Software Co-Design*, Springer Science & Business Media.

Demil, B., and Lecocq, X. 2006. "Neither Market nor Hierarchy nor Network: The Emergence of Bazaar Governance," *Organization Studies* (27:10), pp. 1447–1466.

(https://doi.org/10.1177/0170840606067250).

Dourish, P. 2001. *Where the Action Is: The Foundations of Embodied Interaction*, Cambridge, Mass: MIT Press.

Drechsler, R., and Breiter, A. 2007. "Hardware Project Management-What We Can Learn from the Software Development Process for Hardware Design?.," in *ICSOFT Proceedings*, pp. 409–416.

Dyer, J. H. 1997. "Effective Interim Collaboration: How Firms Minimize Transaction Costs and Maximise Transaction Value," *Strategic Management Journal* (18:7), pp. 535–556. (https://doi.org/10.1002/(SICI)1097-0266(199708)18:7<535::AID-SMJ885>3.0.CO;2-Z).

Egyedi, T. M., and Joode, R. van W. de. 2004. "Standardization and Other Coordination Mechanisms in Open Source Software," *International Journal of IT Standards and Standardization Research (IJITSR)* (2:2), pp. 1–17. (https://doi.org/10.4018/jitsr.2004070101).

Ekbia, H. R. 2009. "Digital Artifacts as Quasi-Objects: Qualification, Mediation, and Materiality," *Journal of the American Society for Information Science and Technology* (60:12), pp. 2554–2566. (https://doi.org/10.1002/asi.21189).

Faulkner, P., and Runde, J. 2009. "On the Identity of Technological Objects and User Innovations in Function," *Academy of Management Review* (34:3), pp. 442–462. (https://doi.org/10.5465/amr.2009.40632318).

Faulkner, P., and Runde, J. 2013. "Technological Objects, Social Positions, and the Transformational Model of Social Activity," *MIS Quarterly* (37:3), pp. 803–818.

Faulkner, P., and Runde, J. 2019. "Theorizing the Digital Object," *MIS Quarterly*, (43:4) pp. 1-24.

Feller, J., Finnegan, P., Fitzgerald, B., and Hayes, J. 2008. "From Peer Production to Productization: A Study of Socially Enabled Business Exchanges in Open Source Service Networks," *Information Systems Research* (19:4), pp. 475–493. (https://doi.org/10.1287/isre.1080.0207).

Feller, J., and Fitzgerald, B. 2002. *Understanding Open Source Software Development*, Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Fischer, S. 1977. "'Long-Term Contracting, Sticky Prices, and Monetary Policy': A Comment," *Journal of Monetary Economics* (3:3), pp. 317–323. (https://doi.org/10.1016/0304-3932(77)90025-3).

Fitzgerald. 2006. "The Transformation of Open Source Software," *MIS Quarterly* (30:3), p.

587. (https://doi.org/10.2307/25148740).

Fitzgerald, B., and Feller, J. 2002. “A Further Investigation of Open Source Software: Community, Co-ordination, Code Quality and Security Issues,” *Information Systems Journal* (12:1), pp. 3–5. (https://doi.org/10.1046/j.1365-2575.2002.00125.x).

Gajski, D. D., and Vahid, F. 1995. “Specification and Design of Embedded Hardware-Software Systems,” *IEEE Design Test of Computers* (12:1), pp. 53–67. (https://doi.org/10.1109/54.350695).

Gurbaxani, V., and Whang, S. 1991. “The Impact of Information Systems on Organizations and Markets,” *Communications of the ACM* (34:1), pp. 59–73. (https://doi.org/10.1145/99977.99990).

Heidegger, M. 1962. *Being and time*, Oxford: Blackwell.

Hippel, E. V., and Krogh, G. V. 2003. "Open source software and the “private-collective” innovation model: Issues for organization science". *Organization science*, 14(2), pp. 209-223.

Howison, J., and Crowston, K. 2014. “Collaboration Through Open Superposition: A Theory of the Open Source Way,” *MIS Quarterly* (38:1), pp. 29-50.

Kallinikos, J., Aaltonen, A., and Marton, A. 2010. “A Theory of Digital Objects,” *First Monday* (15:6). (https://doi.org/10.5210/fm.v15i6.3033).

Kallinikos, J., Aaltonen, A., and Marton, A. 2013. “The Ambivalent Ontology of Digital Artifacts,” *MIS Quarterly* (37:2), pp. 357–370. (https://doi.org/10.25300/MISQ/2013/37.2.02).

Kallinikos, J., and Mariátegui, J.-C. 2011. “Video as Digital Object: Production and Distribution of Video Content in the internet Media Ecosystem,” *The Information Society* (27:5), pp. 281–294. (https://doi.org/10.1080/01972243.2011.607025).

Klein, H. K., and Myers, M. D. 1999. “A Set of Principles for Conducting and Evaluating Interpretive Field Studies in Information Systems,” *MIS Quarterly* (23:1), pp. 67–93. (https://doi.org/10.2307/249410).

Koch, S., and Schneider, G. 2002. “Effort, Co-Operation and Co-Ordination in an Open Source Software Project: GNOME,” *Information Systems Journal* (12:1), pp. 27–42. (https://doi.org/10.1046/j.1365-2575.2002.00110.x).

Krogh, G. von, and Hippel, E. von. 2006. “The Promise of Research on Open Source Software,” *Management Science* (52:7,), pp. 975–983.

Lakhani, Karim R., and Eric Von Hippel. 2004. "How open source software works:“free” user-to-user assistance." *Produktentwicklung mit virtuellen Communities*. Gabler Verlag. pp. 303-339.

Langlois, R. N., and Garzarelli, G. 2008. “Of Hackers and Hairdressers: Modularity and the Organizational Economics of Open-source Collaboration,” *Industry and Innovation* (15:2), pp. 125–143. (https://doi.org/10.1080/13662710801954559).

Lerner, J., and Tirole, J. 2003. “Some Simple Economics of Open Source,” *The Journal of Industrial Economics* (50:2), pp. 197–234. (https://doi.org/10.1111/1467-6451.00174).

Lindberg, A., Gaskin, J., Berente, N., and Lyytinen, K. 2014. *Exploring Configurations of Affordances: The Case of Software Development*, In *Proceedings of the 20th Americas Conference on Information Systems*. Retrieved from https://aisel.aisnet.org/amcis2014/SocioTechnicalIssues/GeneralPresentations/12/.

Linux Documentation Project (2001) *Linux Documentation Project.* https://www.tldp.org/

last accessed Feb. 28, 2020.

Lipinski, M., van der Bij, E., Serrano, J., Wlostowski, T., Daniluk, G., Wujek, A., Rizzi, M., and Lampridis, D. 2018. “White Rabbit Applications and Enhancements,” in *2018 IEEE International Symposium on Precision Clock Synchronization for Measurement, Control, and Communication (ISPCS)*, Geneva: IEEE, September, pp. 1–7. (https://doi.org/10.1109/ISPCS.2018.8543072).

Lipiński, M., Włostowski, T., Serrano, J., and Alvarez, P. 2011. “White Rabbit: A PTP Application for Robust Sub-Nanosecond Synchronization,” in *Control and Communication 2011 IEEE International Symposium on Precision Clock Synchronization for Measurement*,, September, pp. 25–30. (https://doi.org/10.1109/ISPCS.2011.6070148).

MacCormack, A., Rusnak, J., and Baldwin, C. Y. 2006. “Exploring the Structure of Complex Software Designs: An Empirical Study of Open Source and Proprietary Code,” *Management Science* (52:7), pp. 1015–1030. (https://doi.org/10.1287/mnsc.1060.0552).

Macher, J. T., and Richman, B. D. 2008. “Transaction Cost Economics: An Assessment of Empirical Research in the Social Sciences,” *Business and politics*, *10*(1), 1-63.

Malone, T. W., Yates, J., and Benjamin, R. I. 1987. “Electronic Markets and Electronic Hierarchies,” *Communications of the ACM* (30:6), pp. 484–497. (https://doi.org/10.1145/214762.214766).

Manovich, L. 2001. *The Language of New Media*, MIT Press.

Marion, T. J., Friar, J. H., and Simpson, T. W. 2012. “New Product Development Practices and Early Stage Firms: Two In-Depth Case Studies,” *Journal of Product Innovation*

*Management* (29:4), pp. 639–654. (https://doi.org/10.1111/j.1540-5885.2012.00930.x).

Markoff, J. 2018. “Time Split to the Nanosecond Is Precisely What Wall Street Wants,” *The New York Times*. (https://www.nytimes.com/2018/06/29/technology/computer-networks-speed-nasdaq.html).

Markus, M. L. 2007. “The Governance of Free/Open Source Software Projects: Monolithic, Multidimensional, or Configurational?,” *Journal of Management & Governance* (11:2), pp. 151–163. (https://doi.org/10.1007/s10997-007-9021-x).

Masum, H. 2001 "Reputation layers for open source development". In:Making Sense of the Bazaar: *Pro-ceedings of the 1st Workshop on Open Source Software Engineering*. Feller, J., Fitzgerald, B. & van der Hoek, A.(eds). http://opensource.ucc.ie/icse2001/papers.htm.

Maxwell, J. A. 2013. *Qualitative Research Design: An Interactive Approach*, (3rd ed.), Applied Social Research Methods ; 41, Thousand Oaks, Calif: SAGE Publications.

Mellis, D., and Buechley, L. 2012. “Collaboration in Open-Source Hardware: Third-Party Variations on the Arduino Duemilanove,” in *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW ’12, Seattle, Washington, USA: Association for Computing Machinery, February 11, pp. 1175–1178. (https://doi.org/10.1145/2145204.2145377).

Miles, M. B., and Huberman, A. M. 1994. *Qualitative Data Analysis: An Expanded Sourcebook*, (2nd ed.), Thousand Oaks: Sage Publications.

Mockus, A., Fielding, R. T., and Herbsleb, J. D. 2002. “Two Case Studies of Open Source Software Development: Apache and Mozilla,” *ACM Transactions on Software Engineering and Methodology* (11:3), pp. 309–346. (https://doi.org/10.1145/567793.567795).

Moreira, P., Serrano, J., Wlostowski, T., Loschmidt, P., and Gaderer, G. 2009. “White Rabbit: Sub-Nanosecond Timing Distribution over Ethernet,” in *In Proc. of the International IEEE Symposium on PrecisionClock Synchronization for Measurement, Control and Communication, ISPCS*, pp. 58–62.

Nambisan, S., Majchrzak, A., Song, M., and Lyytinen, K. 2017. “Digital Innovation Management: Reinventing Innovation Management Research in a Digital World,” *MIS Quarterly* (41:1), pp. 223–238. (https://doi.org/10.25300/MISQ/2017/41:1.03).

Niederman, F., Davis, A., Greiner, M. E., Wynn, D., and York, P. T. 2006. “Research Agenda for Studying Open Source II: View Through the Lens of Referent Discipline Theories,” *Communications of the Association for Information Systems* (18). (https://doi.org/10.17705/1CAIS.01808).

Oberloier, S., and Pearce, J. 2017. “General Design Procedure for Free and Open-Source Hardware for Scientific Equipment,” *Designs* (2:1), p. 2. (https://doi.org/10.3390/designs2010002).

O’Mahony, S., and Ferraro, F. 2007. “The Emergence of Governance in an Open Source Community,” *Academy of Management Journal* (50:5), pp. 1079–1106. (https://doi.org/10.5465/amj.2007.27169153).

Oshri, I., Henfridsson, O., and Kotlarsky, J. 2018. “Re-Representation as Work Design in Outsourcing:A Semiotic View,” *MIS Quarterly: Management Information Systems* (42:1), pp. 1–23.

Pan, W., Li, Z., Zhang, Y., and Weng, C. 2018. “The New Hardware Development Trend and the Challenges in Data Management and Analysis,” *Data Science and Engineering* (3:3), pp. 263–276. (https://doi.org/10.1007/s41019-018-0072-6).

Pearce, J. M. 2012. “Building Research Equipment with Free, Open-Source Hardware,” *Science* (337:6100), pp. 1303–1304. (https://doi.org/10.1126/science.1228183).

Raymond, E. 1999. “The Cathedral and the Bazaar,” *Knowledge, Technology & Policy* (12:3), pp. 23–49. (https://doi.org/10.1007/s12130-999-1026-0).

Rolandsson, B., Bergquist, M., and Ljungberg, J. 2011. “Open Source in the Firm: Opening up Professional Practices of Software Development,” *Research Policy* (40:4), pp. 576–587. (https://doi.org/10.1016/j.respol.2010.11.003).

Sanchez, R., and Mahoney, J. T. 1996. “Modularity, Flexibility, and Knowledge Management in Product and Organization Design,” *Strategic Management Journal* (17:S2), pp. 63–76. (https://doi.org/10.1002/smj.4250171107).

Scacchi, W. 2001. "Software development practices in open software development communities: a comparative case study". In: *Making Sense of the Bazaar: Proceedings of the 1st Workshop on Open Source Software Engineering*. Feller, J., Fitzgerald, B. & van der Hoek, A.(eds). http://opensource.ucc.ie/icse2001/papers.htm.

Shah, SK. 2005. *Open Beyond Software*. C. Dibona, D. Cooper, and M. Stone, eds. Open Sources 2. O’Reilly Media, Sebastopol, CA.

Shah SK 2006. "Motivation, governance, and the viability of hybrid forms in open source software development." *Management Science* (52:7) pp.1000–1014.

Sharma, S., Sugumaran, V., and Rajagopalan, B. 2002. “A Framework for Creating Hybrid-Open Source Software Communities,” *Information Systems Journal* (12:1), pp. 7–25. (https://doi.org/10.1046/j.1365-2575.2002.00116.x).

Tiwana, A., Konsynski, B., and Bush, A. A. 2010. “Research Commentary —Platform Evolution: Coevolution of Platform Architecture, Governance, and Environmental Dynamics,” *Information Systems Research* (21:4), pp. 675–687. (https://doi.org/10.1287/isre.1100.0323).

Tullio, D. D., and Staples, D. S. 2013. “The Governance and Control of Open Source Software Projects,” *Journal of Management Information Systems* (30:3), pp. 49–80. (https://doi.org/10.2753/MIS0742-1222300303).

Wareham, J., Fox, P. B., and Giner, J. L. C. 2014. “Technology Ecosystem Governance,” *Organization Science 25*(4), pp.1195-1215.

(https://doi.org/10.1287/orsc.2014.0895).

Wareham, J., and Pujol, L. 2019. “From Big Science to Big Business,” *Research Europe*. 6 June 2019 p.12

Wareham, J., and Sonne, T. 2008. “Harnessing the Power of Autism Spectrum Disorder,” *Innovations: Technology|Governance|Globalization* (3:1), pp. 11–27.

Watson, R. T., Boudreau, M.-C., Greiner, M., Wynn, D., York, P., and Gul, R. 2005. “Governance and Global Communities,” *Journal of International Management* (11:2), Information Technology and International Business: Theory and Strategy Development, pp. 125–142. (https://doi.org/10.1016/j.intman.2005.03.006).

West, J., and Kuk, G. 2016. “The Complementarity of Openness: How MakerBot Leveraged Thingiverse in 3D Printing,” *Technological Forecasting and Social Change* (102), pp. 169–181. (https://doi.org/10.1016/j.techfore.2015.07.025).

Williamson, O. E. 1975. *Markets and Hierarchies, Analysis and Antitrust Implications: A Study in the Economics of Internal Organization*, New York: Free Press.

Williamson, O. E. 1979. “Transaction-Cost Economics: The Governance of Contractual Relations,” *The Journal of Law and Economics* (22:2), pp. 233–261. (https://doi.org/10.1086/466942).

Williamson, O. E. 1985. *The Economic Institutions of Capitalism: Firms, Markets, Relational Contracting*, (1st Free Press pbk. ed.), London: Collier Macmillan Publishers.

Williamson, O. E. 1996. *The Mechanisms of Governance*, Oxford University Press.

Yoo, Y. 2010. “Computing in Everyday Life: A Call for Research on Experiential Computing,” *MIS Quarterly* (34:2), pp. 213–231. (https://doi.org/10.2307/20721425).

Yoo, Y., Henfridsson, O., and Lyytinen, K. 2010. “Research Commentary—The New Organizing Logic of Digital Innovation: An Agenda for Information Systems Research,” *Information Systems Research* (21:4), pp. 724–735. (https://doi.org/10.1287/isre.1100.0322).

Yu, F., Pasinelli, M., and Brem, A. 2018. "Prototyping in Theory and in Practice: A Study of the Similarities and Differences between Engineers and Designers," *Creativity and Innovation Management* (27:2), pp. 121–132. (https://doi.org/10.1111/caim.12242).

# Appendix A

## Key construct definition

*Table 5 Key construct definition*

| | Construct definition | Source | Related notions in the literature | Key departure |
|---|---|---|---|---|
| **About the Object** | **Hybrid object:** Digital objects with material and non-material components. Hybrid objects include any hardware with middleware or software and encompass many of the objects being developed in OSH projects. | (Faulkner and Runde 2019) | Type of digital objects or digital artifacts or IT artifacts. | We extract from the literature on digital objects the traits ascribed to such objects and any reference to digital objects with some degree of physicality. |
| | **Embodiment:** Material (or perpetual) and non-material (or ephemeral) embodiment. | (Faulkner and Runde 2009, 2011; Yoo et al. 2010) | *Numerical representation* (Manovich 2001)<br>*Largely unstable, unbounded and resisting reification* (Ekbia 2009) | We differentiate between material and non-material components. |
| | **Modularity:** "Modularity represents the technical realization of the simple yet powerful idea that integral, en bloc objects or systems are hard to act upon, control, and manipulate" (Kallinikos et al. 2013, p. 360). It is *an* attribute of object components that refers to their faculty of being responsive to and distinct from one another. When they are responsive to and distinct from one another, they are *loosely coupled*; when the components are responsive but not distinct from one another they are *tightly coupled.* | (Kallinikos et al. 2010, 2013; Kallinikos and Mariategui 2011; Manovich 2001;<br>Yoo et al. 2010) | *Communicability, sensibility, and associability* (Yoo 2010; Yoo et al. 2010) | Modularity is the first condition for OS development. OS can be applied to costly and highly complex software development *if* it allows for *modularity*, that is, breaking down complex problems into "smaller, independent or weakly connected problems" that can be then dealt with by diverse agents. Modularity decreases the need for contributors to coordinate their task interdependencies actively. Modularity increases flexibility |

| | | | | |
|---|---|---|---|---|
| | | | | (Benkler, 2002; Fitzgerald 2006;<br>Lindberg, 2013; MacCormack, Rusnak, & Baldwin, 2006; Howison and Crowston, 2014). |
| | **Granularity:** Granularity refers to the ability of an object to be decomposed into numerous, small-grained components. Modularity refers to the relationship between components, whereas granularity refers to the number of units to which one can decompose the object. Both modularity (tightly versus loosely coupled) and granularity (high or low) should be considered to be continuums, that is, matters of degree, not discrete alternatives. | (Benkler 2006;<br>Kallinikos et al. 2010, 2013; Kallinikos and Mariategui 2011; Manovich 2001) | *Infinite expansibility* (Faulkner and Runde 2009, 2011, 2019) | Granularity is the second condition for OS development. An object can be OS developed *if* the components of the object are sufficiently granular or small-grained. The granularity of the components is crucial and determines the possibility of distributed agents to simultaneously cooperate in concurrent tasks in part of the same development process. 'To pool a relatively large pool of contributors, the modules should be predominantly fine-grained, or small in size. This allows the project to capture contributions from large numbers of contributors whose motivation level will not sustain anything more than quite small efforts towards the project' (Benkler 2006, p. 10).<br>Benkler, 2002, 2006; Lindberg, 2013 |
| | **Editability:** Digital objects are pliable and are susceptible to be modified continuously and systematically. Editability can be achieved by rearranging, adding, modifying or eliminating | (Kallinikos et al. 2013; Kallinikos et al. 2010; Kallinikos and Mariategui 2011; Manovich 2001) | *Accessibility* (Benkler, 2006, Lessig 2006)<br>*Adaptability* (Zittrain, 2008; Benkler, 2006; Lessig 2006) | Editability is the third condition of IS and involves the integration characteristics of the object. An object can be OS developed if the |

| | | | | |
|---|---|---|---|---|
| | elements. | | *Addressability* (Yoo, 2010; Yoo et al. 2010)<br>*Interactivity* (Kallinikos et al. 2013)<br>*Openness* (Kallinikos et al. 2013, Kallinikos et al. 2010, Kallinikos and Mariategui, 2011)<br>*Recombinability* (Faulkner and Runde 2009, 2011)<br>*Reprogrammability* (Yoo 2010; Yoo et al. 2010; Kallinikos et al. 2013; Kallinikos and Mariategui 2011; Manovich 2001; Zittrain 2008<br>*Variability* (Manovich, 2001)<br>*Traceability* (Yoo 2010; Yoo et al. 2010)<br>*Transcoding* (Manovich 2001) | cost of integrating independent modules and making them interoperable or the cost of connecting people to tasks is sufficiently low due to efficient and cheap network communications (Benkler, 2002; Langlois and Garzarelli, 2008; Howison and Crowston, 2014). |
| | **Reproducibility:** Minimal marginal cost. | (Faulkner and Runde 2009, 2011) | *Transferability* (Zittrain, 2008), Benkler, 2006), Lessig, 2006)<br>*Distributedness*, which refers to seldom being contained within a single source or institution (Kallinikos et al. 2013)<br>*Non-rivalry*, which concerns the possibility of an object being used simultaneously by a large number of parties (Faulkner and Runde 2013 p.815, 2009, 2011, 2019) | Reproducibility is associated with embodiment, as it describes the pragmatic or economic cost of producing and distributing multiple units of the object or component (Kallinikos et al. 2010). |
| **About the developme** | **Development:** The “social process of designing, developing, and implementing the technical | (Akhlaghpour, Wu, Lapointe, Pinsonneault 2013) | - | Based on TCE, we identify two generalized development models (in |

| | | | | |
|---|---|---|---|---|
| **nt** | artifact, usually in a specific organizational context and over time". | | | TCE terms, governance structures) based on decreasing levels of centralized coordination and direction-giving: 1) *hierarchical control* and 2) *contractual agreements*, with the addition of 3) *volunteer contributions* from the OS literature. |
| | **Organizational attributes of a development process:** major characteristics that describe the organization of a particular (or type of) development process (cf. 'how' a development process is organized). | Based on (Crowston and Howison 2006; Feller and Fitzgerald 2002; Fitzgerald and Feller 2002; Raymond 1999) | - | - |
| | About **OS development**<br>**Autonomy and the self-selection of tasks**: OS is characterized by a collaborative effort where agents combine effort voluntarily and self-select their tasks, which does *not* mean that they do not receive pecuniary compensation (though that may often be true) but rather that the collaborators choose their tasks in a similar way that arises in the assignment of sellers to products in a classic market (Lindberg, 2013). "Work is not assigned to developers; instead, they choose what to work on" (Sharma et al. 2002 p.10). | (Crowston, 1997; Howison and Crowston, 2014; Lindberg, Berente, Gaskin and Lyytinen, 2016; Maha and Vaast, 2015; Shah 2005, 2006; Di Tullio and Stapies, 2014) | - | - |
| | **Loosely centralized**: OS is characterized by distributed teams that have access to the source code, submit code patches to solve problems and | (Cutosksy et al., 1996; Moon & Sproull, 2000, Feller and Fitzgerald 2000, 2002; Feller et | - | - |

| | | | | |
|---|---|---|---|---|
| | add functionalities to the software. | al. 2002) | | |
| | **Virtual boundaries**: OS is characterized by a geographically distributed community defined by virtual rather than physical boundaries. OS communities do not have well-defined boundaries and remain open to new contributors, which can join at any time and are fluid in allowing any member to leave the community. Users can not only contribute to the source code but also test the software, report bugs, or suggest new features. | (Cook 2001; Feller et al. 2008; Feller and Fitzgerald 2002, 2002; Markus 2007) | - | - |
| | **Asynchronous collaboration and open superposition of tasks**: OS is characterized by massive parallel development, debugging, and asynchronous collaboration supported by the internet as a communication, collaboration and distribution platform and by concurrent versioning software. Complex OS collective work can be completed in a sequence of layers or modules with distinct functionality and payoffs that do not depend on future work for its utility. | (Cook 2001; Feller et al. 2008; Feller and Fitzgerald 2001, 2002; Markus 2007) | - | - |
| | **Infrastructural tools that facilitate parallel development**: The internet and concurrent versioning systems allow the submission and responsive testing of code patches and the frequent releases that characterize OS. | (Baldwin and Clark 2006; Egyedi and Joode 2004; Feller et al. 2008; Feller and Fitzgerald 2002). | - | - |
| **About the common TCE concepts** | **Asset specificity**: The degree to which an asset can be redeployed to alternative uses and by alternative users without any sacrifice of productive value. | (Dyer 1997; Macher and Richman 2008; Williamson 1975, 1985, 1989, 1996) | See Table 1, which address the relation among the TCE concepts with | |

| | | | |
|---|---|---|---|
| **used** | **Duration**: The time during which the transaction will transpire. | | attributes of hybrid object components. |
| | **Frequency**: How often specific transactions occur. | | |
| | **Search costs**: Costs associated with searching markets for supplier/product availability and the determination of price and quality. | | |
| | **Uncertainty**: The uncertainty surrounding the transaction that includes market, geopolitical or institutional uncertainties. | | |
| | **Monitoring and enforcement costs**: Includes the costs associated with ensuring that each party fulfills a predetermined set of obligations and with any legal costs required for enforcement. | | |
| | **Interdependence**: The degree to which a product or process can be decomposed into discrete tasks and completed by individual vendors.<br>(*Related to*)<br>*Product and process complexity*: Relating to the number of components and the extent of the interactions to manage between these components. | | |
| | **Transaction risk**: The potential economic or opportunity cost associated with a failed transaction. | | |

## Appendix B

### Example of interview guide

| **About the organization, roles, and responsibilities** |
|---|
| • What does your organization do? |
| • What is your role at the organization? |
| • When did you got involved in WR? |
| • What was your task? |
| • Has your task changed over time? |
| **Initial engagement** |
| • How did the organization know about and initially get involved in WR? |
| • How did your organization fund the investment for collaborating in WR? Did it change over time?<br>• (in case it was via a contract): What was the reason for the contract? Duration? What happened after the contract? |
| **Motivational aspects for collaborating** |
| • What were the motivational aspects behind the collaboration? |
| • How did these motivations change over time? |
| **About WR technology and the process of development** |
| • What are the components, functions, and applications of WR? |
| • Please describe the development cycle of WR (including versions). |
| • How did the development of WR hardware differ from WR gateware and software? |
| • How did the different development tasks relate to one another for hardware, gateware, software? |
| • What were, in your opinion, the major events in the development of WR? Why? |
| **Coordination** |
| • How did you develop your task? |
| • Did you collaborate with someone? |
| • Did you report to anyone inside and outside your organization? |
| • Which tools did you use to develop and communicate the outcomes of your task? |
| • Did you have meetings? For what purpose? |
| • How did you use the repository, wiki, mailing list? Others? |
| **The role of the license** |
| • What is the OSH license? |
| • Did you participate in the debate on the OSH license? |
| • In your opinion, what are the differences between the license versions? |

# Appendix C

*Table 6. Data analysis and theoretical constructs related to the development models*

| Nº | Events | Time | Development phase | Theoretical observation | Theoretical construct |
|---|---|---|---|---|---|
| (1) | Diligent requirements collection through ***CERN- supplier contract***- to gather joint (cross-organizational) specifications- | 2008 | Specifications | Cosylab was called in to gather requirements by using input from CERN, GSI's Facility for Antiproton and Ion Research (FAIR) project, L'Institut de Physique Nucléaire de Lyon (IPNL), and ITER, the international nuclear fusion project. Requirements were collected through the phone, video conferences and in person. The requirements were organized into layers, starting at the lowest (physical) layer and moving up until the event distribution processor. Then, commonalities were listed, and the potential incompatibilities were identified. | Contractual agreement for specifications |
| (2) | *Publication of all specifications by CERN*: All specifications files are published to benefit from peer review and to enable remote collaboration. | 2008 | Specifications | Open publication of all specifications so that developers could join the collective endeavor | Voluntary contributions |
| (3) | Contract to create a web-based Open Hardware Repository portal ***CERN- supplier contract.***<br>Contained: file repository, a wiki and general project documentation | 2009 | Specifications and Design | CERN decided to outsource the development of a repository for the following reasons, according to testimonials:<br>• Time saved by having fast research results<br>• Increased quality of the requirements by adding a wide field of expertise<br>• Time saved by flexible addition of complementary development services<br>• Tailored solution that fit an open-source community | Contractual agreement to develop an infrastructural tool to allow further open source development |
| (4) | Provision of an open repository to allow community collaboration | 2009-present | All | Open repository contents are considered to be the knowledge hub of all the WR community. | Voluntary contributions |
| (5) | Community manager to coordinate the requests of developers and | 2009- | All | A person designated at CERN orchestrates and directs all | Voluntary contributions |

| | | | | | |
|---|---|---|---|---|---|
| | users | present | | outside communications and requests to enter the community, implementation requirements, etc. | |
| (6) | *Contract with supplier* for WR switch hardware design (e.g., cards) | 2008 | Design | CERN outsourced to an expert engineering company the design of the WR switch (only hardware) to accelerate the design process. | Contractual agreement for hardware design |
| (7) | *Contract with supplier for WR switch software and gateware* | 2008 | Design | CERN outsourced to an expert software company the development of core-specific software and gateware for the switch to accelerate the design process. | Contractual agreement for software design |
| (8) | *Quality and design review of software and hardware by CERN design unit and Beams department* | 2008 | Design and Prototyping | In-house: quality review of the design of both software/hardware core WR | Hierarchical control |
| (9) | *Contract with supplier* for assemblage (production of prototypes) | 2012 - 2013 | Prototyping | CERN supplier for the hardware design, outsourced the assemblage of WR switch prototypes. Although it was outsourced by the supplier, CERN had direct control on the quality (1-year iterations due to gaps from production files to actual production). | Contractual agreement for prototypes |
| (10) | *Contract with second supplier for WR production (GSI- Creotech)* | 2017 | Prototyping | GSI contracted a second supplier to produce WR prototypes to look for redundancy in the system. | Contractual agreements for prototypes |
| (11) | *Contract with second supplier for WR production (CERN-Creotech) to compare quality* | 2017 | Prototyping | CERN purchased from second provider WR switch to test the quality and compare it with the first provider. | Contractual agreements for prototypes |
| (12) | *Plugfest* to check interoperability with other companies' hardware designs and to showcase WR to attract developers | 2010 | Design | | Voluntary contributions |
| (13) | Contract CERN- Supplier to develop a WR Starting Kit- to attract users and developers (peer review). | 2012 | Testing (and Design) | Starting Kit consisted of a couple of Spartan-6-based boards called SPEC, one of which can be configured to be a master and the other as a slave to encourage users to perform early-evaluation experiments. *WR starting kit developed* | Voluntary contributions |
| (14) | *Coordination meetings with CERN personnel* | 2008-2019 | All | Section meetings were led by the director of the section to manage and coordinate actions across WR development. | Hierarchical control |
| (15) | *Coordination meetings with the four direct suppliers* | 2008- | Design | CERN- supplier meetings to direct tasks | Hierarchical control |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
|  |  | 2012 |  |  |  |
| (16) | *Coordination meetings with all WR community led by CERN* | 2008-2019 | All | Workshops with all the WR community to coordinate and direct tasks. Conclusion – agreement after the first workshop – Timing Workshop Summary<br>Summary of the Timing Workshop held 15 February 2008: CERN will be the manager for all the tasks related to WR development. | Hierarchical control (Project management – task assignments across community) |
| (17) | *Documentation control* by CERN (including schematics and PCB documentation by CERN) | 2008-2019 | Design | Upload all documentation of the schematics and PCB. Identify incompleteness with respect to prototyping (flaws), improve the quality of documentation and guarantee the accessibility of such information. | Hierarchical control |
| (18) | Release of an open source hardware license v.1.1 to attract voluntary developers (not via contracts) and govern the distribution of the open hardware designs. | 2011 | Design, Testing, and Implementation | A new open source hardware license will attract contributors from outside the organization.<br>The research group released Version 1.1 of its open hardware license (OHL) three months after the initial license was published. The license borrows concepts from open source software licensing models but governs the use of hardware designs instead of source code. | Voluntary contributions |
| (19) | Release second version of open source hardware license | 2011 | All | Release second version of open source hardware license to attract voluntary developers (not via contracts) and govern the distribution of the open hardware designs. | Voluntary contributions |
| (20) | Release of an open source hardware license v.1.1 | 2013 | All | Release of an open source hardware license v.1.1 to attract voluntary developers (not via contracts) and govern the distribution of the open hardware designs. | Voluntary contributions |
| (21) | Development of an open source tool that allows open hardware design (KiKat) | 2019 | Design | Development of alternative tools to the proprietary ones that exist to facilitate open hardware design | Voluntary contributions |

# 7

## 7. Discussion and conclusion

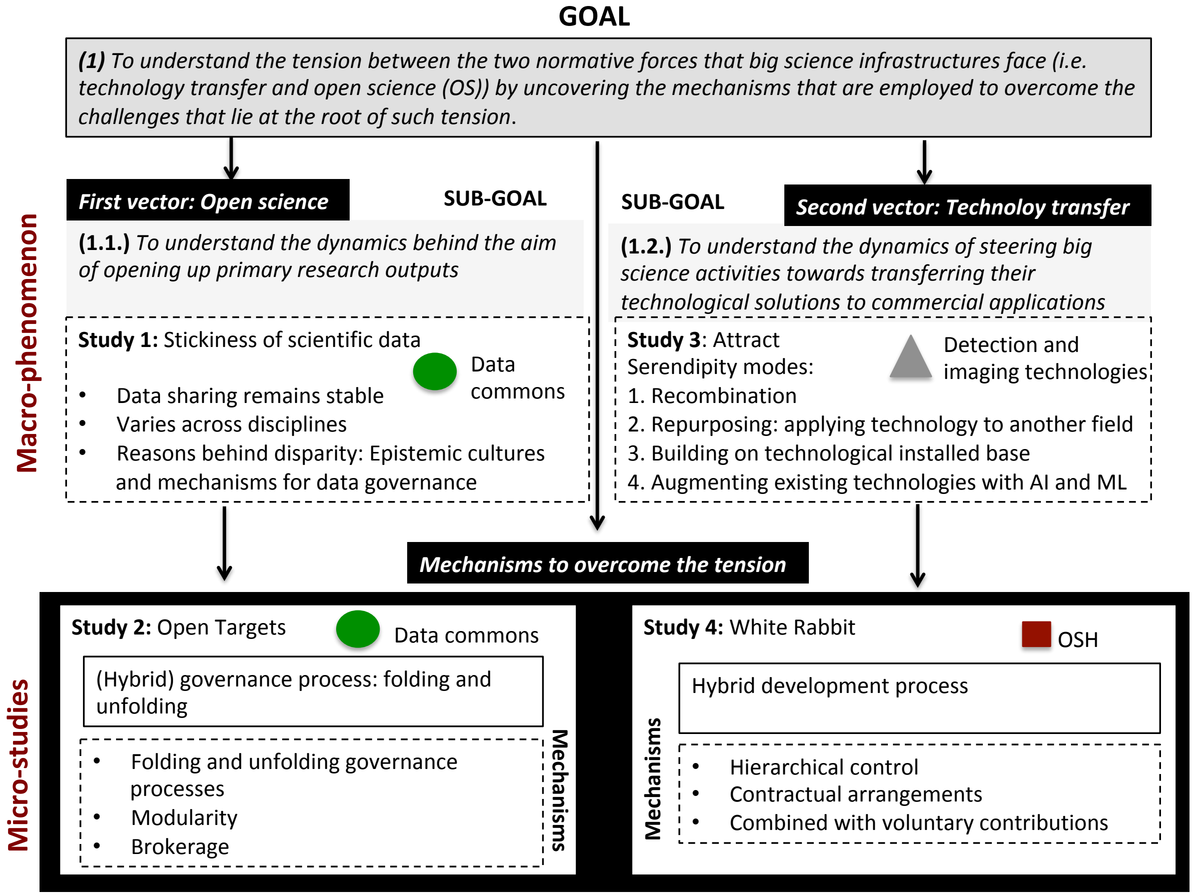
This final chapter integrates the findings of the articles that compose chapters 3,4,5 and 6, as well as discussing the theoretical contributions, managerial and policy implications, limitations of the dissertation, followed by future research opportunities

## 7.1 Theoretical contributions

The main goal of this Ph.D. dissertation is to understand the tension between two competing vectors that are influencing big science infrastructures: 1) calls for more openness in scientific processes and outcomes, and 2) a need for more effective technology transfer.

On one hand, studies #1 and #3 have offered a contextual overview of the dynamics within each of these exogenous forces in isolation and have provided contributions to a) the literature of open science, and b) technology transfer, respectively. A better understanding of each vectorial force has enabled an exploration of the dynamics at the intersection of the two forces. Considering study #2 and study #4 as a joint product, we extract several broader contributions to the study of information systems development at the intersect of the two vectors (figure 1). We conclude with table 1 that summarizes the theoretical contributions and normative implications.

*Figure 1. Overview of the goals of the four empirical studies*



**(1)** *To understand the tension between the two normative forces that big science infrastructures face (i.e. technology transfer and open science (OS)) by uncovering the mechanisms that are employed to overcome the challenges that lie at the root of such tension.*

**(1.1.)** *To understand the dynamics behind the aim of opening up primary research outputs*

**(1.2.)** *To understand the dynamics of steering big science activities towards transferring their technological solutions to commercial applications*

**Study 1:** Stickiness of scientific data

- Data sharing remains stable
- Varies across disciplines
- Reasons behind disparity: Epistemic cultures and mechanisms for data governance

**Study 3**: Attract
Serendipity modes:
1. Recombination
2. Repurposing: applying technology to another field
3. Building on technological installed base
4. Augmenting existing technologies with AI and ML

● Data commons
▲ Detection and imaging technologies
Mechanisms to overcome the tension
Micro-studies

**Study 2:** Open Targets

(Hybrid) governance process: folding and unfolding

- Folding and unfolding governance processes
- Modularity
- Brokerage

**Study 4:** White Rabbit

Hybrid development process

- Hierarchical control
- Contractual arrangements
- Combined with voluntary contributions

### 7.1.1 *Contribution to open science literature*

Open science literature suggests that scientific data sharing confers increased costs to scientists and their institutions without commensurate professional benefits (Borgman 2015; Edwards 2019; Edwards et al. 2011; Tenopir et al. 2015; Wallis et al. 2013). Yet we lack a recent overview of whether, and how, data is shared across scientists.

The inequality of policy attention towards scientific data sharing compared to other trends within open science (i.e. open access) has resulted in a lack of evidence about if scientists share their data, how they share data, and the reasons behind their data sharing behavior. While data about open access publications are readily available, indicators about scientific data sharing have been lacking.

In consequence, generating a comparable global survey data set informs future policy discussions and is one of the primary contributions to this literature stream. Additionally, the analysis of the study #1 data also advances that there is no homogenous explanation of why the number of researchers who have shared their data remains stable from 2016 to 2018. The study confirms the difficulties of scientific data sharing (e.g. Borgman 2015; Edwards et al. 2011; Piwowar et al. 2007), but further contributes by uncovering the delicate system of mechanisms that need to be implemented to align individual and collective incentives in a manner consistent with the specific epistemic cultures of each scientific community and their professional practices. Specifically, in comparing two scientific communities (HEP and MB), the study provides a theoretical explanation of why the slow adoption of data sharing is due to an intertwined web of varied cultures and rational pursuits. Where HEP and MB have significantly different epistemic cultures, research infrastructures, and scientific practices, both communities have established information infrastructures with mechanisms designed to mitigate the domain-specific costs and facilitate data sharing and re-use. In short, the study suggests that modularity and time dilation can be employed as governance mechanisms that reconcile the collective benefits of the scientific community with individual academic career incentives. Both mechanisms can be employed across various epistemic cultures to accommodate divergent practices across scientific communities. Appropriate and transparent governance enacted in the information infrastructures can mitigate the perceived risks preventing scientific data sharing.

### 7.1.2 *Contribution to the literature on technology transfer in big science*

The present dissertation also contributes to the literature on big science by opening the 'black box' of how big science brings new technologies to society in applications previously unanticipated and sheds light on the nature of the serendipity behind such broad and multifaceted process.

While previous literature extensively reflects on the technology contributions of big science infrastructures (Autio 2014; Autio et al. 2003, 2004; Castelnovo et al. 2018; Hallonsten 2014; Heidler and Hallonsten 2015) and has mostly based on anecdotal examples of scientific discoveries and individual scientists' experiences, yet we lacked knowledge on how such infrastructures may purposefully realize such serendipity process in a more systemic level (Autio, 2014).

The third study sits in this literature and explores the dynamics behind the aim of transferring big science solutions to unanticipated market applications through the analysis of the 170 projects funded under the ATTRACT initiative. The analysis uncovers four modes wherein serendipity can be cultivated: Recombination of technologies; repurposing; building and extending technology from previous research; and AI and ML to augment existing technologies.

The study contributes to evolving research on serendipity beyond its simple conceptualization as a natural accident and suggests that big science infrastructure can proactively shape the transfer of their technological solutions to other industrial settings.

### 7.1.3 *Contribution to IS development*

Finally, and probably the major theoretical contribution of the dissertation relies upon the intersection of these two vectors (i.e. open science and technology transfer) and sits at the core of IS development. By assessing two different open science dimensions in Open Targets (OT) and White Rabbit (WR) (i.e. data commons and open-source hardware), the cases help elucidate the specific mechanisms that organizations use to reconcile the tensions caused when for-profit entities contribute to opensource or commons-based resource pools.

In particular, in the case of White Rabbit, we describe the friction of transposing an open-source model of development to digital objects with physical components (hybrids). We isolate how hybrid objects deviate in nature and form, that is, their attributes, in comparison to pure non-material digital objects (Faulkner and Runde 2009, 2013, 2019). Thereafter, the attributes of hybrid objects are analyzed under the conditions of open source development (i.e. high modularity, high granularity, and low integration costs), to understand the applicability of the 'open source way' in OSH development (Benkler 2002; Feller et al. 2002, 2008; Fitzgerald 2006; Howison and Crowston 2014). In the WR case, we found that open-source development was complemented with more traditional commercial development processes at specific points. The case uncovers the need for hierarchical control and contractual agreements that allow direction-giving and coordination in the development of a highly sophisticated OSH such as WR. This was a result of the highly complex and interdependent nature of the technology, which requires centralized control, technical expertise, and more considerable financial investments. Combined with the generative nature of the OS community, WR was successfully developed and deployed as a powerful precision and synchronization technology at many scientific research infrastructures, and sequentially, in numerous industrial settings. This wider diffusion of WR outside of its immediate scientific purview is a result of the open-source community supporting it.

The case of Open Targets moves to a different open science dimension (i.e. data commons). It sheds light on the dynamic governance of an information infrastructure that overcomes the challenge of simultaneously aligning individual and collective interests (Constantanides 2012, Constantinides and Barrett 2015; Hanseth and Monteiro 1997). By integrating ideas from information infrastructure scholarship and collective action theory (Hardin 1968, 1982; Ostrom 1990), we theorize that the openness-attribute of information infrastructure is a manageable with appropriate mechanisms that enable movement from private to open workspaces. This allows contributions to common goods by for-profit companies that need opacity and closure following competitive and market logic. In other words, to overcome the historical, social dilemmas of collective action (i.e., free-riding and overconsumption) and provide the effective incentives for contributors to invest in the commons, two mechanisms are employed that afford the fluid movement between open and closed spaces of work: the principle of modularity, which refers to the technical architecture of the infrastructure, and the role of a broker or a trusted third party, that serves as an arbiter amongst the organizations to orchestrate the exchanges.

Taken together both studies, what we can appreciate is the parallelism between the fluid navigation through open and opaque spaces in Open Targets and the hybrid development process in WR. In OT, such hybridity between open and dark places made 'openness' compatible with the traditional, restricted, and controlled spaces of work where protected R&D processes take place to pursue the competitive race towards a new drug. In WR, the ‘hybrid development process’ made ‘openness’ and the generativity of open-source compatible with hierarchical control and contractual agreements to coordinate and afford direction-giving in the development of a complex OSH such as WR. What the combination of the two studies teaches us is that ***‘openness’ needs some degree of opacity*** to find the proper equilibrium between the two vectorial forces. The studies go one step further and advance *how* organizations can navigate across the shadows; that is, graduated levels of transparency and accessibility.

In both cases, the very prominent ***role of the digital artifact*** is evident. In OT, the technical attributes of a multi-layered infrastructure designed around the principle of modularity afforded navigation between scientific openness and closed market logic, making individual and collective interests compatible. In dissecting the technical attributes of the ‘object’ to understand where and how different agents act, we can decipher the mechanisms that emerged to make for-profit and community-based collaboration in both White Rabbit and Open Targets development. In both studies, by identifying the technical attributes of the artifacts we manage to relate them to the governance approach that stakeholders follow to find the optimum equilibrium between openness and technology commercialization.

Equally important and related to the technical characteristics, both studies feature ***the role of the organizational attributes*** that accompany the development of a data commons infrastructure or open-source hardware development. The *arbitrage* role of an operational team at Open Targets that behaves as a trusted-third party governing the exchanges between the organizations, or the orchestrating role of CERN who grandfathered and directed the development of White Rabbit.

In sum, when contrasting the dynamics of each of the vectorial forces (i.e. open science and technology transfer) with the friction *across* the vectors, we elucidate the intricate complexities that interact when scientific institutions attempt to simultaneously foster openness in research processes while boosting the commercialization of their technologies. Our case studies show how the technical attributes of a digital 'object' or information infrastructure combine with effective arbitration towards effective policy interventions. Table 1 provides an overview of the contributions and normative implications of the different studies in this dissertation.

.

*Table 1. Overview of the contributions of the different studies and normative implications*

| **Study #** | **1: The stickiness of scientific data** | **2: Opaque spaces of the commons: Governing information infrastructures in Life Sciences** | **3: Systematising serendipity for big science infrastructures** | **4: From bits to atoms: White Rabbit at CERN** |
|---|---|---|---|---|
| **Research Question** | Do researchers share their data? How do they share their data? Which mechanisms emerge to enable researchers to share their data? | How do organizations develop commons-based information infrastructures that govern access to collective resources while simultaneously protecting the members' private interests? | Which are the formative conditions of serendipity transforming big science research towards commercial applications? | How do the attributes of a hybrid object and its components affect the open-source model of development? |
| **Theoretical Foundation** | Epistemic cultures<br>Collective action theory | Information Infrastructures<br>Collective Action theory | Serendipity | Digital objects and IT artifacts<br>Open-source<br>Transaction Costs Economics |
| ***Contribution*** | Epistemic cultures (communitarian versus individualistic) coexist with rational cost-benefit estimations<br>The principles of modularity and time dilation are mechanisms that allow fostering data sharing practices by making compatible individual and collective interests. Both mechanisms allow mitigating differences in more communitarian and individualistic scientific epistemic cultures | Two dynamic processes: Folding and unfolding to transition from open to opaque spaces of work<br>The two processes are afforded by the principles of modularity (technical architecture of the infrastructure) and brokerage (organizational attributes of the infrastructure) | Four serendipity models:<br>1. Recombination<br>2. Repurposing: applying technology to another field<br>3. Incremental: build and extend technology from previous research<br>4. AI and ML to augment existing technologies | Hybrid development model<br>The physical nature of the components of hybrid objects inhibits the conditions of open source development and leads to the emergence of a hybrid model that combines hierarchical control, contractual arrangements, and voluntary contributions. |
| ***Normative Implications*** | Sharing scientific data is not a dichotomous decision, but it needs to establish a *degree* towards *what* data do you share (modularity), and *when* do you share it (time dilation - embargos*).*<br>Mechanisms shaping incentives and rewards need to be designed locally to account for differences in epistemic cultures and suggesting that one-size-does-not-fit-all. | The development of data infrastructures based upon commons needs to allow the dynamic transition between open and opaque spaces of work to preserve the private incentives of for-profit to invest in the infrastructure development and to overcome the historical, social dilemmas of collective action (i.e., free-riding and overconsumption).<br>A modular architecture combined with the role of a broker or a trusted third party, who is assigned coordination and arbitration tasks to orchestrate and mediate the flows of data are required to overcome the apparent incompatibility of openness versus commercialization of R&D outputs. | Big science infrastructure can actively shape the transfer of their technological solutions to alternative industrial settings by proactively cultivating four serendipity models. | The material aspect of hybrids objects can reduce object editability, granularity, *reproducibility, and integration characteristics.*<br>While open source is a powerful model that can serve as to leverage crowd knowledge towards developing frontier technologies, it might need to be supplemented with more traditional commercial development processes at specific points based on the component attributes (i.e. editability, granularity, integration, and reproducibility). |

## 7.2Managerial and policy implications

The managerial implications of the Ph.D. thesis are multiple. With the COVID-19 crisis, the policy attention on open science has grown. The response of the scientific community to COVID-19 outbreak has been to embrace the principles of open science unprecedented levels, including: sharing preprints to speed up access to research outputs; global scientific data sharing related to COVID-19 to accelerate discovery (Pells 2020); and the development of open-source hardware prototypes of ventilators (Buytaert et al. 2020). The examples have only been the tip of the iceberg of what some open science ideologists have been trying to pursue for the last decade. We can predict that the open science policy mandate is not only here to stay but will push much further.

Yet, although we know how openness help to leverage the spread of skills and expertise and accelerate discovery, we see also in the exemplar space of COVID-19 that this goal can be at odds with the companies engaged in the commercialization of ventilators (Buytaert et al. 2020); additionally, it also induces fear into pharma concerning the safeguard of data and knowledge flows in the race towards a COVID vaccine or other treatments (Darzi 2020).

The present dissertation places itself at the center of this tension and tries to overcome the 'dualistic' ideological debates where stakeholders position themselves at the extremes of the open-closed continuum in policy discussions. As such, we try to offer a nuanced perspective on how to pursue a 'smart' openness and inform policy interventions by suggesting governance mechanisms that manage not only to safeguard economic interests of for-profits in their R&D and innovation pursuits, but also to align their individual interests with the open science demands.

While the full industrial impact of scientific data infrastructures based upon commons and open-source hardware is yet to come, both the Open Targets and White Rabbit studies offer inspiring formulas of how they frame the tension as a manageable trade-off with appropriate governance mechanisms. In an exercise to describe the policy relevance of the results, we provide a summary of the policy implications of the studies in table 2.

We conclude by elaborating two examples of the major policy implications and resulting recommendations from the dissertation results: First, we consider the present context where the European Commission is currently investing in the development of a major scientific data infrastructure, the European Open Science Cloud (EOSC) (European Commission, 2019), which is foreseen to accelerate scientific data sharing across European countries and beyond. Our results suggest that *modularity* should be a major characteristic of the EOSC architecture to successfully attract the engagement of a broad range of the extended community, including for-profit entities. Developing an infrastructure with different layers and access rights while simultaneously allowing embargo periods over the data (time dilation) can allow organizations and scientists to navigate across their required levels of opacity to align their specific interests with the common good. Allowing some degrees of ‘darkness’ in EOSC design will also contribute to achieving larger engagement of the wide community and, hence, to its sustainability. Paradoxically, if the rational pursuits of commercial organizations or scientists’ professional incentives are ignored in the technical development and governance of the EOSC, it may jeopardize the uptake of such infrastructures and lead to wasteful public expenditures.

A second example can be taken from open source hardware, where public research infrastructures are experimenting with this new formula in the procurement of their scientific experimental tools (Pearce, 2012). Open-source hardware offers these organizations the possibility to avoid vendor lock-in, to merge disperse expertise from their network of suppliers and contributors, and at the same time, align their public mission by fully disseminating the design of their technologies. However, if funding agencies do not permit some degrees of darkness through permissive open-source hardware licenses that allow proprietary (and non-disclosed) developments to emerge around the core technology, this well-intentioned policy could fail to engage core contributions by those seeking subsequent commercial exploitation on the periphery.

In sum, openness in science needs some degree of opacity to find the proper equilibrium between the two the social benefits of science and the commercial interests of some of its most important contributors. In essence, our analysis suggests that policies calling for carte-blanche openness that ignore the incentives of many profit-seeking organizations that make valuable contributions to the larger ecosystems supporting scientific programs may have undesirable consequences. It is important to move away from naïve ideological debates between the pro-Open with pro-IP advocates and employ hybrid governance approaches that allow resolving the divergent interests of its various stakeholders. This dissertation suggests that policy attention needs to be focused on finding an acceptable equilibrium to make these forces compatible.

*Table 2. Overview of the general policy implications of the different studies*

| **Study #** | **1: The stickiness of scientific data** | **2: Opaque spaces of the commons** | **3: Systematising serendipity** | **4: White Rabbit at CERN** |
|---|---|---|---|---|
| ***Findings*** | 66% of researchers declare making their data available. The % remains stable, with no growth shown over the past two years. Data sharing significantly varies across disciplines. Both communitarian and individualistic scientific communities (different epistemic cultures), employ three mechanisms (with some variation) to enable data sharing in both scientific communities: Modularity; Time dilation; and Boundary organization to establish transparent data governance and mediate the identification of the 'bona fide' researcher. | Folding and unfolding are two governance processes that allow organizations to transition from open to opaque spaces of work. These two governance processes are afforded by the principles of modularity and brokerage that are articulated through the technical and organizational attributes of the infrastructure. These two processes allow overcoming the historical social dilemma of collective action (i.e. free riding and overconsumption) | Four serendipity models:<br>1. Recombination<br>2. Repurposing: applying technology to another field<br>3. Incremental: build and extend technology from previous research<br>4. AI and ML to augment existing technologies | The physical nature of the components of hybrid objects inhibits the conditions of open source development and leads to the emergence of a hybrid model that combines hierarchical control, contractual arrangements, and voluntary contributions. |
| ***General Policy Implications*** | Mechanisms shaping incentives and rewards towards scientists to foster scientific data sharing need to be designed locally to account for differences in epistemic cultures and suggesting that one-size-does-not-fit-all. | The development of public data infrastructures (or the federation of existing ones) needs to allow the navigation between open but also restricted spaces of work combined with embargo periods over the data (or time dilation between the creation and disclosure of the data) to preserve private and individual incentives.<br>These infrastructures need to be governed by a trusted third-party of the community (with bilateral Non-disclosure agreements) that behaves as an arbiter and orchestrates the data flow. | The policy mandate of the increasing impact of big science infrastructures can be materialized in a systemic way by the infrastructures through purposively facilitating inside their activities four serendipity paths (corresponding to the four serendipity models).<br>New public funding instruments can experiment with the four models to accelerate the technology transfer of big science technological solutions to alternative industrial settings | Public procurement policies of research infrastructures at large can foster the model of open-source hardware – as they have done for open-source software. By advocating for this model of development in their procurements they will allow large peer review of their technologies, will avoid vendor lock-in situations, and obtain system efficiency gains by avoiding redundant technology developments across infrastructures. Yet, the voluntary contributions of open source need to be complemented by more traditional commercial development processes at specific points based on the technology attributes (i.e. editability, granularity, integration, and reproducibility).<br>Open-source hardware licenses applied to these public procurements need to consider allowing contributors to engage and not disclose proprietary developments around the core open and standardized hardware technology to protect for-profit interests and enable an ecosystem to emerge around the technology. |

## 7.3 Limitations and future research

Notwithstanding its theoretical contributions, this Ph.D. dissertation also has several limitations. While the individual limitations of each study are explained separately in each chapter, this section aims at providing a holistic perspective.

Regarding the study of data sharing attitudes and practices, where the sample sizes were large, the period between the two surveys was only two years. Given the phenomenon studied, this sampling is likely insufficient to detect long-term patterns. Additional surveys with the same instrument can enrich our current data. Additionally, research that purposefully examines the heterogeneity in data sharing practices across disciplines can benefit from in-depth comparisons of high-intensive and low-intensive data sharing scientific communities to explore whether the mechanisms uncovered in our study to mitigate the domain-specific barriers and facilitate data sharing and re-use are applicable in other scientific contexts.

Open Targets is constituted by some of the world's most formidable research organizations together with highly capitalized pharmaceutical companies. As such, the generalizability of the findings to other information infrastructures in different contexts might also be limited. It should be noted in the case of OT is also exceptional because of the extremely competitive nature of the life sciences industry. Other information infrastructures may not have the historical precedence of secrecy, legal protection, and long investment lifecycles.

The challenge of the case method is to generalize the findings. Nevertheless, it is worth mentioning that there is a trade-off between internal and external validity. White Rabbit is also a very sophisticated and expensive technology. It is plausible that OSH with less cost and complexity could be developed in entirely different modalities. Hence, while we acknowledge the difficulties of generalizing the results of OT or WR to the larger populations, it is equally valid that the internal validity of our findings in both studies is the main focus. Our results are deeply grounded in the contexts under study, and by employing established procedures in inductive research (Miles and Huberman 1994), the two cases sought to maximize the internal validity of our results.

Regarding the study of ATTRACT, we exploit a unique dataset of 170 projects funded with €100,000 to develop a proof-of-concept commercial application within one year. This is also a unique policy intervention historically unprecedented in the European Commissions. As such, we encourage more systematic analysis with other novel datasets to understand other approaches by which a serendipity process can be identified, brokered, and cultivated. In the future, the role of generative computing and machine learning will likely be necessary for this respect.

Finally, by studying an extreme case of OSH developed with the sponsorship of CERN, we acknowledge that WR is non-representative, yet it is studied to understand something likely to become more predominant in the future. Extreme cases are particularly useful for theory generation, as they exhibit high values on variables of critical interest (Gerring 2007). Nevertheless, the level of technical complexity, financial resources, and political stature of CERN are likely unique yet essential influences in the case. Hence, we should be prudent in extrapolating our results to a different context that does not display the same local characteristics. We encourage additional research in OSH to investigate the heterogeneity of hybrids in different contexts to substantiate further the relationships between the attributes of hybrid components and multiple forms of development. Recent announcements of OSH ventilators being developed in response to the COVID 19 virus are an obvious opportunity for such research (Pearce 2020).

## References

Autio, E., 2014. *Innovation from Big Science: Enhancing Big Science Impact Agenda*, Department of Business, Innovation & Skills. Imperial College Business School, p. 76.

Autio, E., Hameri, A.-P., and Bianchi-Streit, M. 2003. *Technology Transfer and Technological Learning through CERN's Procurement Activity*, CERN.

Autio, E., Hameri, A.-P., and Vuola, O. 2004. "A Framework of Industrial Knowledge Spillovers in Big-Science Centers," *Research Policy* (33:1), pp. 107–126. (https://doi.org/10.1016/S0048-7333(03)00105-7).

Benkler, Y. 2002. "Coase's Penguin, or, Linux and 'The Nature of the Firm,'" The Yale Law Journal (112:3), p. 369. (https://doi.org/10.2307/1562247).

Borgman, C. L. 2015. *Big Data, Little Data, No Data: Scholarship in the Networked World*, Cambridge, UNITED STATES: MIT Press. (http://ebookcentral.proquest.com/lib/georgetown/detail.action?docID=3339930

Buytaert, J., Abud, A. A., Akiba, K., Bay, A., Bertella, C., Bowcock, T., ... & Dikic, N. (2020). The hev ventilator proposal. arXiv preprint arXiv:2004.00534.

Constantanides P (2012) Perspectives and Implications for the Development of Information Infrastructures (IGI Global).

Constantinides P, Barrett M (2015) Information Infrastructure Development and Governance as Collective Action. Information Systems Research 26(1):40–56.

Castelnovo, P., Florio, M., Forte, S., Rossi, L., and Sirtori, E. 2018. "The Economic Impact of Technological Procurement for Large-Scale Research Infrastructures: Evidence from the Large Hadron Collider at CERN," *Research Policy*. (https://doi.org/10.1016/j.respol.2018.06.018).

Darzi, A. 2020. "The Race to Find a Coronavirus Treatment Has One Major Obstacle: Big Pharma | Ara Darzi," *The Guardian*. (https://www.theguardian.com/commentisfree/2020/apr/02/coronavirus-vaccine-big-pharma-data).

Edwards, P. N. 2019. "Knowledge Infrastructures under Siege : Climate Data as Memory, Truce, and Target," *Data Politics*, Routledge, March 13, pp. 21–42. (https://doi.org/10.4324/9781315167305-2).

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

European Commission. 2019a. “Facts and Figures of Open Research Data,” European Commission - European Commission. (https://ec.europa.eu/info/research-and-innovation/strategy/goals-research-and-innovation-policy/open-science/open-science-monitor/facts-and-figures-open-research-data_en, accessed April 19, 2019).

Faulkner, P., and Runde, J. 2009. “On the Identity of Technological Objects and User Innovations in Function,” Academy of Management Review (34:3), pp. 442–462. (https://doi.org/10.5465/amr.2009.40632318).

Faulkner, P., and Runde, J. 2013. “Technological Objects, Social Positions, and the Transformational Model of Social Activity,” MIS Quarterly (37:3), pp. 803–818.

Faulkner, P., and Runde, J. 2019. “Theorizing the Digital Object,” MIS Quarterly, (43:4) pp. 1-24.

Feller, J., Finnegan, P., Fitzgerald, B., and Hayes, J. 2008. “From Peer Production to Productization: A Study of Socially Enabled Business Exchanges in Open Source Service Networks,” *Information Systems Research* (19:4), pp. 475–493. (https://doi.org/10.1287/isre.1080.0207).

Feller, J., and Fitzgerald, B. 2002. Understanding Open Source Software Development, Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Fitzgerald. 2006. “The Transformation of Open Source Software,” MIS Quarterly (30:3), p. 587. (https://doi.org/10.2307/25148740).

Hallonsten, O. 2014. “How Expensive Is Big Science? Consequences of Using Simple Publication Counts in Performance Assessment of Large Scientific Facilities,” *Scientometrics*. (https://doi.org/10.1007/s11192-014-1249-z).

Hanseth O, Lyytinen K (2010) Design Theory for Dynamic Complexity in Information Infrastructures: The Case of Building Internet. Journal of Information Technology 25(1):1–19.

Hardin G (1968) The Tragedy of the Commons. Science 162(3859):1243–1248.

Hardin R (1982) Collective action (Published for Resources for the Future by the Johns

Heidler, R., and Hallonsten, O. 2015. “Qualifying the Performance Evaluation of Big Science beyond Productivity, Impact and Costs,” Scientometrics. (https://doi.org/10.1007/s11192-015-1577-7). Hopkins University Press, Baltimore).

Howison, J., and Crowston, K. 2014. “Collaboration Through Open Superposition: A Theory of the Open Source Way,” MIS Quarterly (38:1), pp. 29-50.

Miles, M. B., and Huberman, A. M. 1994. Qualitative Data Analysis: An Expanded Sourcebook, (2nd ed.), Thousand Oaks: Sage Publications.

Pearce, J. M. 2012. "Building Research Equipment with Free, Open-Source Hardware," *Science* (337:6100), pp. 1303–1304. (https://doi.org/10.1126/science.1228183).

Pells, R. 2020. "Coronavirus and Ebola: Could Open Access Medical Research Find a Cure?," *The Guardian*. (https://www.theguardian.com/education/2020/jan/22/people-cant-learn-about-treatments-they-need-why-open-access-to-medical-research-matters).

Piwowar, H. A., Day, R. S., and Fridsma, D. B. 2007. "Sharing Detailed Research Data Is Associated with Increased Citation Rate," *PLoS ONE* (3), p. 5.

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., Pollock, D., and Dorsett, K. 2015. "Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide," *PLOS ONE* (10:8), (P. van den Besselaar, ed.), p. e0134826. (https://doi.org/10.1371/journal.pone.0134826).

Wallis, J. C., Rolando, E., and 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), (L. A. Nunes Amaral, ed.), p. e67332. (https://doi.org/10.1371/journal.pone.0067332).