MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of Master's Thesis

"Datafying relations, Relationalising data: Investigating forms of togetherness at the CERN open data portal"

verfasst von / submitted by

Antonia Winkler, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien, 2023 / Vienna, 2023

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet: UA 066 906

Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet: Masterstudium Science-Technology-Society

Betreut von / Supervisor: Univ. Prof. Dr. Ulrike Felt

Acknowledgements

This thesis is part of the METAFORIS (Making Europe through and for its research infrastructures) project, which is jointly funded by the “DACH Project” funding scheme of the Austrian Science Fund (FWF, I3982) and the Deutsche Forschungsgemeinschaft (DFG). The METAFORIS project is a collaboration between the Science Technology Society (STS) department in Vienna (under the leadership of Ulrike Felt) and the Munich Center for Technology in Society (under the leadership of Sebastian Pfotenhauer).

I want to thank all members of the METAFORIS project for the fascinating conversations we had, which greatly enriched this thesis - Noah Münster, Kamiel Mobach, Nina Klimburg-Witjes, Erik Aarden, Kaye Mathies, Zinaida Vasilyeva and Oguz Özkan.

The last two and a half years at the STS department in Vienna have taught me a lot. Not only have I discovered my passion for social science theory and practice, but I also got the chance to become part of a study and work environment which truly inspired me. During my studies here at the STS department, I have gained the confidence to think about a future in academia. I am extremely grateful to all my teachers and fellow students who made this possible through their encouraging words. I particularly want to thank Ulrike Felt, my supervisor for this thesis. Thank you for giving me a chance to work on the METAFORIS project and allowing me to discover the fascinating world of particle physics. Your thoughtful comments and advice on my work were invaluable. From conceptualizing conference presentations over conducting interviews to developing this thesis, I have benefited immensely from your supervision. I also want to thank Noah Münster for his contribution to this thesis. Your continuous feedback was extremely valuable, and your support throughout this process inspired me greatly.

I additionally want to thank all my interviewees, without whom this work would not have been possible. Thank you for meeting with me and being so open and interested in this project. I greatly admire your dedication to open science and your continuous efforts to make academia a more inclusive place.

Finally, I want to thank Jojo, Becci, Zikka, and Christine for all their love and support.

Abstract English

The increasing datafication of research has restructured working procedures in many scientific fields and continues to reorganize the social, temporal, and spatial aspects of research. Over the last decades, research institutions around the globe have started to rearrange their working procedures with respect to the production, processing and sharing of data. The European Organization for Nuclear Research (CERN), in particular, has invested in the development of unique infrastructures for data processing. This rapid implementation of digital technologies ushered in a new era of sociotechnical organization at CERN and crucially restructured the ways in which researchers structure their working procedures. In order to unfold the ways in which research organization is enacted by data practices at CERN, I will follow the movement of data through and beyond the organization. By investigating the ways in which data is made mobile, the structures necessary to facilitate these data flows gain visibility. This thesis will specifically focus on how data moves from the CERN open data portal (ODP) to diverse locations of research. The foundation of the CERN ODP in 2014 marked a crucial step in CERN's open data strategy. It was fueled by the ambition to make data available to a diverse set of communities. As of 2022, the portal offers more than two petabytes of data and enables educational as well as research-oriented endeavors. Investigating the dynamics around the development and use of the CERN ODP will allow insights into how data practices can facilitate, shape, and direct research organization across a multiplicity of spatial, temporal, and sociotechnical boundaries.

Abstract German

Die zunehmende Datafizierung der Forschung strukturiert Arbeitsabläufe in verschiedenen wissenschaftlichen Bereichen neu und organisiert die sozialen, zeitlichen und räumlichen Aspekte der Forschung um. In den letzten Jahrzehnten haben Forschungseinrichtungen begonnen, ihre Arbeitsabläufe im Hinblick auf die Produktion, Verarbeitung und Nutzung von Daten neu zu orientieren. Insbesondere die Europäische Organisation für Kernforschung (CERN) hat in die Entwicklung einzigartiger Infrastrukturen für Datenspeicherung und Verarbeitung investiert. Die Einführung digitaler Technologien am CERN hat eine neue Ära der sozio-technischen Organisation eingeläutet und strukturiert die Arbeitsabläufe von Wissenschaftlern grundlegend um. Um zu erforschen, wie sich die Organisation von Forschung am CERN durch die Einführung digitaler Technologien verändert hat, werde ich der Bewegung von Daten durch die Organisation und darüber hinaus folgen. Durch die Untersuchung von Datenflüssen werden die sozial-epistemischen Strukturen, die die Flüsse ermöglichen, sichtbar gemacht. Dieses Forschungsprojekt konzentriert sich insbesondere darauf, wie Daten vom CERN Open Data Portal (ODP), CERNs Infrastruktur für den öffentlichen Zugang zu Forschungsdaten, zugänglich gemacht und wiederverwendet werden. Die Gründung des CERN ODP im Jahr 2014 war ein entscheidender Schritt in CERNs open data Strategie. Sie

wurde von dem Bestreben angetrieben, Daten für verschiedenste Nutzer*innen Gruppen zur Verfügung zu stellen. Seit 2022 bietet das Portal mehr als zwei Petabyte an Daten an und ermöglicht sowohl bildungs- als auch forschungsorientierte Nutzung. Eine Untersuchung der Dynamiken rund um die Entwicklung und Nutzung des CERN ODP wird Einblicke in die Daten Praktiken und die Forschungsorganisation am CERN und darüber hinaus ermöglichen.

Contents

Acronyms

ALICE: A Large Ion Collider Experiment

ANT: Actor Network Theory

AOD: Analysis Object Data

ARC: Analysis Review Commitee

ATLAS: A Toroidal LHC Apparatus

CERN: European Organization for Nuclear Research

IT: Information Technology

ODP: Open Data Portal

CMS: Compact Muon Solenoid

DOI: Digital Object Identifier

HEP: High Energy Physics

ICT: Information and Communication Technology

JSON: JavaScript Object Notation

LHC: Large Hadron Collider

LHCb: Large Hadron Collider beauty

METAFORIS: Making Europe Through And For its Research Infrastructures

QCD: Quantum Chromo Dynamics

STS: Science Technology Studies

1 Introduction

> “From a historical perspective the question of how experiments end commands our interest because it directs attention to that fascinating moment in the activities of the laboratory when instrumentation, experience, theory, calculation and sociology meet.”(Galison, 1987, p.1)

Research activities at the world’s largest laboratory for particle physics (CERN) revolve around the construction and use of immense accelerator technologies that collide subatomic particles at increasingly higher energies. CERN’s currently biggest particle accelerator is the Large Hadron Collider (LHC), a 26.7 kilometer long underground tunnel located in the border region of Switzerland and France. Inside the LHC, research collaborations investigate particle collisions at four main detectors: Compact Muon Solenoid (CMS), A Toroidal LHC Apparatus (ATLAS), A Large Ion Collider Experiment (ALICE), and Large Hadron Collider beauty (LHCb). Research in big science projects such as CERN is made up of a myriad of endings and beginnings, all entangled in a web of complex decision-making processes.

While always closely intertwined with one another, these endings can be spatial, social, and temporal. The LHC collaborations continuously adapt the sites of experimentation, such as the spaces of the detector and collider. For instance, over the last decades, there has been a strong growth in the size and complexity of detector and accelerator technologies at CERN. Additionally, the actor groups that are part of a particle physics experiment are subject to constant change. For instance, the increasing complexity of accelerator and detector technologies has integrated engineers in the LHC collaborations and thus rearranged the social composition of particle physics groups. Additionally, the temporal ends of the experiment are continuously (re-)negotiated. The LHC collaborations have to decide when the full potential of the collider has been realized and the time has come to move to yet another, more powerful experimental setup. They further need to continuously evaluate the quality of their analysis practices and decide whether their results adhere to the constantly evolving standards of correctness that prevail within the High energy physics (HEP) community.

The datafication of research activities at CERN has had a significant impact on these considerations. New possibilities for data collection and storage have facilitated the growth of collider and detector technologies. Computing infrastructures developed at CERN enable scientists to access, process, and analyze data in real-time and from around the world (CERN, 2022a). These changes have led to a shift in the research objectives of many physicists at CERN. Next to the operation of detector technologies, they now concern the processing and analysis of datasets. While much research at CERN is still focused on adjusting and improving the various components of the detector, datafication has reoriented the collaborations’ goals to the development of data processing techniques and data analysis software to improve the quality of the data and results. Thus, the place of the experiment is no longer exclusively constituted by the collider and the detector but additionally by the computing infrastructures within which parts of the experiment are performed.

This spatial proliferation of the experiment into the digital has had far-reaching effects on the social and epistemic boundaries of research at CERN. Datafication has introduced new actors into the field of particle physics. Scientists specializing in the development of computational techniques have joined the research collaborations at CERN. Consequently, new types of expertise have gained relevance and have started to restructure the epistemic landscape of particle physics. Further, and not unrelated, datafication has transformed what it means to speak about the temporal outcome of the experiment. Researchers not only need to ask themselves when the detector setup has been sufficiently upgraded and used. In addition, they have to evaluate whether the resulting research data has been processed and analyzed adequately. For instance, current developments in artificial intelligence have enabled new ways of processing particle physics data. Through the use of these new computational methods, the results of particular particle physics analyses can transform significantly. This has raised the question of whether these transformed results should represent a new standard for performing certain HEP analyses. This example demonstrates how computational techniques have acquired agency in shaping the outcomes of contemporary HEP research. Data has become an object of experimentation and theorization that is manipulated through increasingly sophisticated computational tools. But when, if ever, is data considered fully processed and analyzed? Does the experiment ever truly end if new computational techniques have the capacity to constantly challenge analysis practices in HEP?

Data-intensive research at CERN is an instance where the intertwinement of technological, social, and epistemic developments in particle physics research becomes visible. Through datafication, new connections between physicists and computational experts are forged. As such, datafication reorganizes the social landscape of HEP. Additionally, what counts as a valid scientific result in particle physics research is continually re-negotiated in the light of new data processing techniques. In this way, the epistemic underpinnings of research are crucially influenced by datafication. Thus, the study of data practices at CERN promises to disclose some of the crucial socio-epistemic dynamics of HEP research. Through the study of data practices, it is not only possible to carve out particular kinds of knowledge that are produced at CERN but additionally to scrutinize their entanglement with the technological and social underpinnings of research. Therefore, in this thesis, I want to investigate how data practices at CERN transform the socio-epistemic structure of research.

Previous research in the field of Science and Technology Studies (STS) has demonstrated how the implementation and use of different types of data follow diverging logics and have multiple effects on specific areas of research (Bates et al., 2016; Leonelli, 2020). Data practices do not exhibit universal, never-changing dynamics across disciplines. They are strongly situated in their production context. This is particularly true for data practices at CERN, where the socio-epistemic organization of research is unique in many ways. Researchers work in collaborations that comprise up to several thousand members. These exceptionally large group numbers are connected to specific organizational practices. For instance, publications released by CERN collaborations undergo extremely extensive and highly unique internal evaluation and review processes before they are made accessible to the public. These review processes involve various layers of revision within which different subgroups of the collaboration evaluate the results. Researchers have to adapt their data analysis practices to the recommendations made during these review processes. In this way, the organizational dynamics of CERN become reflected in the data practices of researchers.

In order to study these data practices, it is crucial to understand research at CERN as relying on a wide variety of infrastructures that store, organize and distribute the data generated by the research collaborations. While these infrastructures are connected in multiple ways, they often express their own functionalities and logics. For instance, they grant access to different user groups and provide varying types of data. Therefore, in this research project, I chose to investigate one particular data infrastructure that promises to encapsulate a situated answer to the questions raised above: The CERN open data portal (ODP).

The CERN ODP was founded in 2014 and marked an attempt at making large datasets produced by the LHC collaborations available to the public. Driven by the ambitious goal of enabling both research-related and educational (re-)use of data, the development of the ODP was realized by an interdisciplinary team consisting of CERN physicists, the CERN IT department, and the CERN scientific information service (Rao et al., 2019). As of 2020, the portal offers more than two petabytes of data (Simko et al., 2020) and has proven to facilitate educational as well as research-related data (re-)use.

The implementation of the CERN ODP needs to be understood in the context of CERN's long-standing commitments to open and accessible research. The CERN convention states that "experimental and theoretical work shall be published or otherwise made generally available." (*Convention for the Establishment of a European Organization for Nuclear Research | CERN Council*, 1953). Inventions such as the world wide web were developed at CERN to facilitate the dissemination of scientific products and have established CERN as a central advocate in the open access movement (Lipton, 2020). Despite this clear commitment towards open research, the establishment of concrete infrastructures for the open release of resources is an issue of continuous debate at CERN. The open data portal constitutes one attempt at translating CERNs ambitions into concrete infrastructuring practices.

In contrast to the LHC collaborations' internal data infrastructures, the CERN ODP addresses data (re-)use communities that are not necessarily affiliated with CERN. As such, the portal is an instance where connections between actors are forged in new ways. In addition, the research results obtained with CERN open data are associated with new forms of evaluation. In contrast to publications that are released with collaboration internal data, which are considered to be fully fledged “real physics” analyses, research that mobilizes open data is considered to result in less rigorous “proof of concept” publications. Thus, the CERN open data portal is an example of how data practice and the socio-epistemic organization of research are inextricably linked. My research interest particularly lies in the data practices that inform the development and maintenance of the CERN open data portal. Additionally, this work will scrutinize the ways in which data is (re-)used from the portal. Consequently, I aim to understand how these data practices enact the socio-epistemic structure of research.

In order to investigate this, I mobilize the notion of “forms of togetherness” initially introduced by Felt (2009). Forms of togetherness denote ways of being together in academic research settings (Felt, 2009). Thus, they aim to scrutinize how actors are connected and disconnected from one another by the particular places in which research is done. “Forms of togetherness” are epistemic and social at the same time. They bring actors together in particular ways while simultaneously shaping and being shaped by specific kinds of knowledge. In the context of this study, I want to extend my understanding of place to include digital places such as the CERN open data portal. In doing so, I suggest that studying the relevance of place in academic research and its impact on the socio-epistemic constitution of a field should include digital places that have become similarly relevant in shaping togetherness and apartness in academic practice.

Expressed through this conceptual frame, the central concern of this study becomes *how data practices concerning the development, maintenance, and use of the CERN open data portal enact different forms of togetherness.* Scholars in the field of HEP have pointed to the complex and context-specific nature of data produced by the LHC collaborations (Rao et al., 2019), making it extremely hard for external researchers to make sense of and work with data produced at CERN. However, diverse user groups have successfully worked with data from the CERN ODP. Therefore, the open data portal constitutes an intriguing project where diverse forms of togetherness are enabled through the use of complex HEP data.

In studying these forms of togetherness, I contribute to a body of work that has interrogated the interrelations of technological and socio-epistemic structures in the HEP community (e.g., Karaca, 2020; Merz & Sorgner, 2022) Additionally, this study feeds into discussions on the increasing trend towards open data in contemporary research settings and its translation into concrete infrastructuring practices (e.g., Kitchin, 2014; Wessels et al., 2017). In the following, I give a brief overview of the main contents of this thesis.

1.1 Overview

In chapter 2, I explore STS literature which is crucial in addressing my research question. The first subchapter mainly focuses on the notion of the “data journey,” which has been developed by Leonelli and Tempini (2020) to describe how datasets move through different socio-epistemic research settings. By following data journeys through concrete research environments, the disruptive and transformative qualities of data move into the foreground. Additionally, the continuous work that guarantees the transportation of data to different locations of research gains visibility. Thus, data journeys are an instrumental concept for pinpointing data practices in the context of the CERN ODP, where datasets are (re-)used in a multiplicity of socio-epistemic settings.

The second subchapter explores STS literature that focuses on the social and epistemic structures of academic research settings. In this section, I outline different vocabularies for conceptualizing the relationship between the epistemic and the social and ask which aspects of my research question these vocabularies allow to address. I engage in concepts such as "collaboration" (Katz & Martin, 1997), "boundary work" (Star & Griesemer, 1989), and "trading zones" (Galison, 1997) and ultimately arrive at the notion "forms of togetherness" (Felt, 2009) to subsume the different types of socio-epistemic ordering that I aim to observe. In both subchapters, I place a particular focus on the ways in which the literature mentioned above has been applied to the field of high energy physics.

After presenting my research question in section 3, chapter 4 focuses on STS research concerning infrastructures. The development and maintenance practices of the CERN open data team enact the CERN open data portal as a digital infrastructure, within which particular actions become possible while others are inhibited. STS research on infrastructures provides sensitivities for analyzing the actions that are enabled and disabled in particular infrastructural settings. The chapter begins with an outline of the central characteristics of infrastructures (Star & Ruhleder, 1996) and moves on to explore how STS research has conceptualized the role of digital infrastructures in particular. In order to understand how the CERN ODP relates to its larger infrastructural embedding, the subsequent section will focus on the ways in which large-scale research infrastructures such as CERN have been conceptualized in STS research.

In chapter 5, I introduce the methodological approach of this work. I chose to conduct a multi-method study with a focus on semi-structured interviews. Initially, I performed a short bibliometric analysis to gain an overview of the actors enrolled in the production and use of CERN open data. Subsequently, I conducted seven online interviews with CERN ODP developers and users and seven interviews with members of the ATLAS collaboration as part of the METAFORIS team during an on-site visit at CERN. Through analyzing source code and conversations on the development of the portal on GitHub (CERN open data, 2023b), and GITTER (CERN open data, 2023a), I substantiated the findings generated from the interview material. In addition to outlining my methodological considerations, Chapter 5 introduces the actors identified through the bibliometric analysis.

Chapter 6 outlines the results of my empirical investigation. I introduce my findings by outlining an exemplary data journey that will underscore how datasets transform in multiple ways as they travel to the CERN open data portal. The subsequent analysis consists of four main parts. The first subchapter investigates portal development and maintenance practices. It focuses on one particular data practice that proved relevant for all members of the open data development team: the application and classification of metadata elements. In the first part of this section, I investigate how metadata practices in the development phase of the portal enacted particular forms of togetherness between the members of the open data team and the data providers from the LHC collaborations. Subsequently, I outline how these forms of togetherness have transformed since the establishment of the portal. The analysis will show that interpersonal collaboration in the development phase of the portal transformed into semi-standardized types of interaction through digital forms such as a metadata schema. The second and third subchapters investigate two distinct dynamics of open data (re-)use. The second subchapter particularly focuses on data (re-)use of a group that was largely unanticipated by the developers of the portal: members of an LHC collaboration. In the analysis of this (re-)use case, I resume the question raised at the beginning of this introduction and ask how open data (re-)use enacts a particular type of research result. I do so by contrasting the "communitarian" (Knorr-Cetina, 1999) data processing and validation practices that prevail inside the CMS collaboration with the open data (re-)use practices of LHC researchers. The analysis will show that open data publications acquire the status of "proof of concept" works, while publications that have undergone CMS internal peer review structures are considered to be "real physics" analyses. As I will outline, this differentiation produces a particular form of apartness between researchers who work with open data and their respective collaborations. The third subchapter focuses on open data (re-)use by theoretical physicists who are not affiliated with CERN. My analysis will show that (re-)use by theorists crucially revolves around a particular type of data that is produced by the LHC collaborations: Simulated Monte Carlo samples. Monte Carlo datasets are computer-generated simulations of collisions that take place inside the collider. They use theoretical assumptions and detailed knowledge of the detector functionalities to reconstruct particle decays. Researchers test their theoretical assumptions by comparing these simulated datasets to "real world" data generated by the detector. For theorists, these datasets proved to be crucial in performing analyses with open data, particularly because they supplied information on the intricate workings of the detector technology. As such, they brought the theorists closer to the material setup of the experiment and aligned some of their data practices with the experimentalists. Consequently, they enacted a new form of togetherness between those two groups.

In the fourth subchapter of the empirical part, I want to take a step back and ask how physicists at CERN, who are not involved in the construction and use of the CERN ODP, perceive current approaches toward open data. These researchers assume a critical position towards current open data developments. National and supranational funding agencies are increasingly adding open data to their list of requirements. This often exerts pressure on research institutions such as CERN to make data open. In light of this development, I ask how CERN scientists evaluate these funding requirements and what they understand as important steps for eliciting successful open data (re-)use in the future.

In chapter 7, I summarize the findings and draw three central conclusions from my work. To conclude, I outline areas for future investigation that reach beyond the scope of this thesis.

2 State of the art

Contemporary knowledge production at CERN crucially relies on the generation and analysis of data. Data-intensive research methods enable scientists to analyze particle collisions simultaneously and from research institutions around the world. In HEP practice, datasets act as portable representations of phenomena recorded by the detector. As such, they have allowed researchers to reorganize their analysis practices substantially. The following literature review will show that datasets undergo extensive transformations as they move from the sites of their production to various locations of physics analysis. Research in the area of data studies has focused on the ways in which data changes on its journeys and asked how these transformations impact the outcomes of research. The first subchapter of this review assembles such studies and asks how they can be mobilized for the analysis of the CERN open data portal.

2.1 Data journeys

Scholars in the field of data studies have placed a focus on the malleable character of data (Leonelli & Tempini, 2020) and investigated how data moves through different socio-epistemic spaces (Bates et al., 2016). In order to describe the bumpy and intertwined movements of data through time and space, Leonelli (2020) has introduced the notion of the "data journey". Data journeys are characterized by disruption and disunity; they display data as always changing relations in various social worlds. Through this understanding, data is "situate[d] [...] across interconnected sites of practice" (Bates et al., 2016, p.2) and acquires a processual character that strips it of any inherent qualities. Thus, for Leonelli (2020) and Bates et al. (2016), data takes on different types of meaning in different socio-epistemic settings. By studying how different communities of practice use data, research can make those meanings visible and engender an understanding of the radically different purposes of data.

Importantly, Leonelli (2020) points out that following data journeys not only means attending to the transformations of data but additionally focusing on the infrastructures that sustain its travel. By interrogating how data changes during the course of its journey, the effects of infrastructures that shape those journeys gain visibility. Additionally, an investigation of data journeys has the capacity to uncover theoretical assumptions that undergird the development of infrastructures for data travel (Leonelli, 2020).

But how can we best study the movements of data through different socio-epistemic settings? Though data journeys often seem invisible at first sight, they become particularly noticeable when points of friction emerge, that is, when a "cost in time, energy, and human attention" (Edwards et al., 2011, p.669) is needed to facilitate temporal and spatial movement. In this context, the research of Edwards et al. (2011) has pointed to the relevance of metadata ("data about data") as an important cause of data friction. Metadata is contextualizing material supplied with data which often crucially influences the ways in which data can be repurposed. By investigating data journeys in the environmental sciences, Edwards et al. (2011) show how discussions on the quality and interoperability of metadata are crucial in resolving data frictions and facilitating flows. As the authors demonstrate, data frictions are not only resolved by practices of metadata standardization but additionally by personal exchanges between researchers where “incomplete, poorly structured [and] mutable descriptions” (Edwards et al., 2011, p.673) of data are passed on. Edwards et al. ’s (2011) analysis suggests that data journeys can be successfully studied by attending to points of friction, where human attention towards data becomes necessary. Metadata practices are particularly relevant in enabling dataflows and often stand at the center of these data frictions. Edwards et al. ’s (2011) account additionally opens up for questioning the ways in which informal and mutable descriptions facilitate the travel of data.

Aula (2019) has interrogated the notion of data friction on the institutional level. By analyzing the role of governing institutions during reforms in the Finish secondary health data infrastructure, Aula (2019) shows how data frictions can arise between different institutions. Next to infrastructural reforms, the development of new legislative and regulatory measures crystallized as an integral approach to resolving these frictions. The account demonstrates the centrality of regulatory aspects in the movement of data across institutional boundaries (Aula, 2019). In the case of the CERN open data portal, the requirements of funding agencies, which increasingly include open data obligations, are relevant in shaping data journeys. Though they do not constitute a focus of this work, I will briefly discuss their influence on open data implementation in subchapter four of the findings section.

Edwards et al. ’s (2011), as well as Aula ’s (2019) account, demonstrate the relevance of human attention toward data in analyzing data journeys. Nadim (2016) has introduced the term “data labor” to describe the much under-appreciated work of facilitating and maintaining the flow of data through digital infrastructures. In her account on the genetic databases “GenBank and EMBL-Bank”, she demonstrates that sustained human labor is necessary to guarantee the mobility and usability of data (Nadim, 2016). Nadim (2016) conceptualizes databases as “spaces of convergence” (García-Sancho, 2011), that is, spaces that sit at the intersection of different “communities of practice” (Lave & Wenger, 1991). In order to allow data to travel through different life worlds, data laborers need to develop creative and versatile practices; they need to “articulate” data in various ways. Nadim (2016) defines four characteristic types of “articulation work” (Karasti & Baker, 2004).

In reference to Leonelli (2009), the first kind of articulation work is termed “packaging” (Nadim, 2016). Packaging refers to practices through which data is stripped of its original context and re-contextualized according to another use setting (Nadim, 2016). Packaging allows data to converge in ways that fit particular uses and practices. In this relation, one could, for instance, consider metadata application practices. By attaching new types of information to data, its meaning and potential use context can transform in important ways.

Further, Nadim (2016) points to articulation work as a form of maintenance and care. Practices of maintenance are designed to guarantee the "workings of standards and other technologies aimed at the reproduction of sameness" (Suchman, 2007, p.269). Standards, in this sense, are not automatically enforced but require continuous articulation.

Further, data laborers might attend to the provenance of data samples and negotiate their relevance in spaces of convergence (Nadim, 2016). Through foregrounding the ancestry of data, an "infrastructural inversion" (Bowker & Star, 1999) takes place that allows users to retrace and understand how choices, standards, and classifications are implemented in data sharing infrastructures (Nadim, 2016).

Finally, articulation is always accompanied by collective practices of "imagining", "which allow [data laborers] to connect data, phenomena, communities, and technologies in their everyday practices" (Nadim, 2016, p.502). For instance, imaginaries on the destination of data journeys are crucial to many forms of data labor.

By investigating the labors that enable the journey of data through different live worlds, an infrastructure of situated practices becomes visible. In the case of the CERN open data portal, it is essential to investigate how articulation work is connected to imaginaries of openness and seamless access. As an infrastructure for the open publication of data, the CERN ODP is part of the open science movement that has gained momentum in recent decades. By aiming to provide general access to scientific resources, infrastructures such as the ODP are likely to be guided by ideas that often proliferate in the context of open science, such as seamlessness and general connectivity.

Ideas connected to the construction and use of the CERN ODP are likely guided by CERN's longstanding commitment to open research. The CERN convention states that "the results of its [CERNs] experimental and theoretical work shall be published or otherwise made generally available." (*Convention for the Establishment of a European Organization for Nuclear Research | CERN Council*, 1953) More recently, the "CERN open data policy" has updated this commitment toward openness with respect to the accessibility and (re-)use of CERN data (CERN, 2020). The (re-)use policy aims at" [m]aking data available responsibly (applying FAIR standards), at different levels of abstraction and at different points in time", which will "allow the maximum realization of their scientific potential and the fulfillment of the collective moral and fiduciary responsibility to member states and the broader global scientific community" (CERN, 2020, p.1).

In the context of this research project, it is of interest to interrogate how these particular conceptions of openness come to matter in the data practices that undergird the CERN ODP. This means investigating who is anticipated as the users of open data releases. For instance, in the "CERN open data policy", responsibility towards CERN member states as well as "the broader scientific community" is framed as an incentive for publishing open data (CERN, 2020). This invites questions on how membership privileges specific data users and determines CERN's understanding of openness.

As outlined above, data frictions regularly arise throughout data journeys and, if unresolved, restrict the use of data by particular groups. Often, it is not the access to data that confines its use to a specific group but the know-how, software, and equipment needed to process it. Open data initiatives have acknowledged these restrictions and called for the development of strategies that enable broader (re-)use of data. Kitchin (2014) has conceptualized the open data movement as based on three central pillars: openness, participation, and collaboration. Through mobilizing these principles, open data initiatives seek to enlarge possibilities for the exploitation of data and, in the long run, democratize knowledge production practices (Kitchin, 2014). However, concepts regarding the precise implementation of open data strategies remain pluralistic and have found varying expressions within different institutions and initiatives Kitchin (2014). As has been criticized by scholars in the field of data studies, conceptions of data as a pre-existing entity rather than a relational product prevail in open data movements (Leonelli et al., 2017). Data is often understood as a clearly defined entity that can easily move through various social contexts (Leonelli et al., 2017). In contrast, the above-outlined literature has demonstrated the elusive and situated character of data and suggests that data is a relational entity that requires constant work and attention. Therefore, this analysis of the CERN open data portal investigates how conceptions of data as either relational or stable shape practices such as portal development and maintenance.

2.2 Data journeys at CERN

In order to make sense of the data journeys at CERN, it is helpful to consider the particular infrastructures and data types used by the LHC collaborations. CERN operates one of the world's most unique infrastructures for data processing and sharing. The "Worldwide LHC Computing Grid", the LHC collaborations' internal infrastructure for the processing and analysis of research data, connects computing centers from research institutions all over the globe and offers an immense storage capacity of 1 exabyte (CERN, 2021). Even though this computing infrastructure enables the storage and processing of unprecedented amounts of data, the LHC collaborations generate more data than they can keep and analyze. Thus, research practices at CERN revolve around the selection of datasets for storage and analysis. Karaca (2020) has interrogated the data acquisition system used by the ATLAS collaboration at CERN. His account demonstrates that limitations in data storage capacity have resulted in the implementation of various automated selection processes (called "Triggers") that preselect data before it is stored and analyzed by ATLAS scientists (Karaca, 2020). Selection, in this case, constitutes a prerequisite for data travel and processing (Karaca, 2020). The "usability, mobility and mutability" (Karaca, 2020, p.46) of data are tightly related when it comes to the processing of data at CERN.

However, data journeys at CERN are not necessarily confined to the CERN internal computing grid. Besides this extensive internal data infrastructure, the LHC research groups release parts of their data to the public via the CERN open data portal (Lassila-Perini et al., 2021). This research project focuses on data practices accompanying the construction, maintenance, and use of the portal. It will interrogate how data is modified and contextualized before it is uploaded to the ODP. In analogy to Karaca's (2020) account, it will focus on how mobility, usability, and mutability of data intertwine in the context of the portal. The CERN open data portal is a highly heterogeneous site where diverse communities of practice intersect. It aims to provide educational datasets, as well as datasets for analysis on the research level. The portal attracts communities ranging from theoretical particle physicists to machine learning experts. Thus, it promises to encapsulate a great variety of different "articulation works" that sustain the movement of data to these different (re-)use groups. The investigation of data curation work, specifically the analysis of packaging, maintaining, inverting, and imaging practices (Nadim, 2016), will be an entry point for sketching out the relational and transformative qualities of ODP datasets.

It is important to note that data on the CERN ODP takes on a range of different forms in terms of complexity and size. Generally, the ODP developers distinguish between three core types of data for open release: Data for research level (re-)use consisting of "simulated" and "collision" type datasets and "derived" data for educational use. Research-level datasets can reach sizes of more than 100 terabytes (CERN, 2023) and are usually accompanied by extensive metadata information.

"Collision" data is a processed version of data generated by an LHC detector. Collision data which is released on the ODP has already undergone an extensive selection process. Next to Trigger systems which filter data before it is recorded (Karaca, 2020), computational methods are used to generate so-called AOD (Analysis Object Data) datasets from the original raw data files. This processed character of collision data raises the question if ODP development procedures make the different steps of data generation visible to the new data (re-)use contexts in the contextualization of resources.

While collision-type datasets are processed versions of data taken by an LHC detector, simulated data (also called Monte Carlo simulations) are computer-generated datasets that combine theoretical assumptions and information about the detectors' functionalities. Simulations are necessary since HEP physicists consider "all data [...] as mixtures of components—including, besides the signal, the background, the noise, and the underlying event—which are distorted and truncated through detector losses and smearings" (Knorr-Cetina, 1999, p.77). Researchers aim to reconstruct these different aspects of the data through simulated data. They can test particular theoretical assumptions by comparing simulated data with collision data. As Knorr-Cetina (1999) points out, simulated data can not be classified as either solely part of data generation or theory testing. Simulation is part of data generation as it delineates the background from the signal and detector effects from the collision. However, it includes theory testing as it simulates collisions by mobilizing particular theoretical assumptions. Hence, the production of simulated data puts the clear delineation between data production and theory testing into question. In the context of simulated data, it will be particularly interesting to observe (re-)use dynamics. As simulated data includes detailed knowledge of detector functionalities, users unfamiliar with the experimental setup will likely face challenges in working with this type of data.

Educational datasets are “derived” from research-level datasets by “select[ing] the collisions of interest” (CMS open data team) for a particular physics process and summarizing these aspects of the original data in a smaller dataset. Even though the (re-)use of educational data is not the focus of this work, it is interesting to note that data labor concerning the generation of educational samples is connected to considerations about “interesting physics phenomena” (CMS open data team), which can be easily reconstructed. In the context of educational resources, the phenomenon is always already pre-given, and any deviation from the envisioned result would be considered a mistake. As such, educational resources exhibit a dynamic that is distinct from research-level resources, which are, at least partially, considered to offer the potential of a new type of analysis (See 6.2).

2.3 Data Journeys and the socio-epistemic organization of research: Forms of togetherness

In the preceding part of the literature review, I have outlined how STS research has conceptualized the movement of data through various socio-epistemic settings. Additionally, I have argued that highly unique data production and analysis strategies at CERN offer to disclose how a diverse range of practices structure and sustain these data journeys. However, data journeys not only transform datasets themselves but additionally reconfigure the socio-epistemic structures in which they are produced and used. The practices that create and sustain infrastructures, such as the open data portal and provide data through them, bear the potential to forge new ties between previously disconnected communities. In this research project, I not only aim to understand how data changes when it travels to diverse research locations, but I also investigate how the use of this data transforms the socio-epistemic structure of research settings. For the CERN open data portal, as a site that aims to distribute data to a diverse range of communities, it is particularly interesting to investigate how data journeys create connections between previously disparate groups and how the movements of data restructure epistemic practices. In what follows, I will outline different conceptual languages for analyzing the effects of data journeys on the socio-epistemic organization of research and ask how they can help me address my research interest.

The first subchapter will interrogate how the concept of collaboration helps in conceptualizing personal interactions between researchers enacted through data practices at the ODP. The concept of collaboration will be particularly relevant for my analysis of ODP development and maintenance practices, which crucially rely on interpersonal types of exchange. However, it is not only personal interaction that transforms due to changing data practices at CERN. Data practices crucially shape organizational structures that transcend the level of direct collaboration. In high energy physics, where more than 6000 scientists work in one collaboration on shared objectives, research organization is multilayered and complex. It is likely that organizational types other than interpersonal exchange are crucially influenced by data storage, processing, and analysis procedures. In this project, I particularly want to draw on the relevance of nonhuman intermediaries in shaping the socio-epistemic organization of research at CERN.

Data journeys at CERN intersect with various human actors that do not necessarily communicate personally. A scientist overseeing the production of a particular dataset at an LHC collaboration will not necessarily engage in personal exchanges with data users. The interaction of these researchers is mediated through data and the technological infrastructure used to supply it. The increasing datafication of research at CERN establishes datasets (as well as other digital elements) as crucial intermediaries between the practices of researchers. Therefore, it is essential to understand how these nonhumans shape the relations between scientists. In the case of the CERN open data portal, this shift from personal interactions between researchers to communication via digital intermediaries will become explicit. As I will demonstrate, interactions through the technological infrastructure of the ODP between members of the ODP development team will, at times, replace personal types of collaboration. Additionally, this research project investigates how datasets become intermediaries between data production and use communities. As outlined above, datasets at CERN undergo highly selective processes before they are used by physicists internally and released on the open data portal. Therefore, they have particular understandings of the relevance of specific particle collisions already “inscribed” (Johnson, 1988) in them. Together with available metadata, they suggest specific lines of analysis to data users. In this way, the (re-)use of data on the ODP establishes a connection between data producers and users.

While some aspects of the datasets continuously transform as they move through CERN to the ODP, some characteristics remain stable across different (re-)use contexts. For instance, the high complexity of LHC datasets is generally considered a stable feature in the context of research-level use. In the upcoming empirical analysis, I will focus on the ways in which stable attributes of data force different user groups into particular behaviors. Previous research on data as an instrument in research organization has often focused on the volatile aspects of data (Leonelli & Tempini, 2020; Bates et al., 2016). Databases have been conceptualized as places where data is made accessible for radically different purposes through data labors that sustain its travel (Nadim, 2016; Leonelli & Tempini, 2020). In the case of the CERN open data portal, it is not only interesting to observe how data is adapted to fit particular (re-)use cases. Additionally, I will scrutinize the ways in which specific aspects of data remain stable across (re-)use communities. Both aspects of data, its volatile character, and its more solid qualities, shape the epistemic practices of researchers who work with data. Researchers need to adapt to the inscription processes of ODP implementers and, at the same time, adhere to the characteristics of data that prevail across (re-)use communities. The empirical analysis will show that this obligation was the cause of data friction in early use cases of the portal.

It is crucial to note that users might analyze data differently than producers anticipate. They can "describe" (Johnson, 1988) to anticipated kinds of use and develop alternative methodologies for working with data. Thus, connections that arise between data users and producers through the open data portal are not solely defined by the practices of data producers. Data users can work with data in unanticipated ways and therefore shape the outcomes of their research. This is why this project investigates open data (re-)use cases as well as ODP development and maintenance. The extent to which users deviate from expected use practices will be a particular focus.

To subsume the materially mediated and interpersonal types of connection which are established between researchers, I mobilize the term "forms of togetherness" (Felt, 2009). Forms of togetherness are social and epistemic. They underscore the intertwinement of shared epistemic practices and different types of interactions between researchers. The term forms of togetherness was originally introduced by Felt (2009) to address the relevance of place in academic practice. Through this notion, Felt (2009) aims to interrogate when places become "a force with detectable [...] effects on social life" (Gieryn, 2000, p.466). As Gieryn (2000) has argued, "places bring people together in bodily co-presence, which may cover a repertoire ranging from engagement to estrangement" (p. 476). I want to extend this understanding of place in my analysis of the CERN open data portal and move away from conceptualizing place as a real-world, physical space. Rather, I consider the CERN ODP as a new digital "place" that is inhabited by different communities and materialities. Within the place of the ODP, new forms of togetherness arise, for instance, through the exchange of data.

Galison (1997) has underscored the relevance of place in particle physics practice. He outlines how the MIT radiation laboratory of the 1940s was designed so that "engineers and physicists [could] work[] within sight of one another" (p.830). This physical co-presence facilitated the development of a common language and consequently fostered cooperation between the two groups. I suggest that datafication has opened up for cooperation in new, digital places that are no less relevant in structuring togetherness/apartness relations in particle physics research. Like the physical places of research, digital places bring together actors and allow them to experiment and theorize.

As Felt (2009) argues, places are crucially shaped by the power relations that prevail between their occupants. This rings especially true for digital places. At the CERN open data portal, the roles of different occupying groups are clearly delineated. For instance, access and upload rights enact a separation between data producers and data users. These are the preconditions under which connections are established between different researchers along the data journey. They are crucially shaping the forms of togetherness resulting from data practices at the portal. By investigating portal development and maintenance practices, these preconditions will become visible.

To sum up, I mobilize the notion "forms of togetherness" for several reasons:

1. It allows me to conceptualize the CERN open data portal as a place where particular types of experimenting and theorizing become possible.
2. It enables me to explore the social and the epistemic as inextricably linked.
3. It describes both, interpersonal collaboration between researchers and mediated types of interaction.

In this way, forms of togetherness subsume the concepts I introduce in the following subsection.

2.3.1 Collaboration

The concept of research collaboration is a prominent approach to the analysis of organizational developments in scientific fields. It is usually characterized by quantitative assessment strategies and has often been conceptualized as co-authorship in academic publications (Katz & Martin, 1997). However, a solely quantitative evaluation of collaboration has faced increasing criticism by scholars who have argued that co-authorship can serve as an indicator of collaborative activity but only captures a very specific dimension of the research process (Katz & Martin, 1997). Who is listed as an author on a publication and who has contributed to a particular analysis is not always congruent (Katz & Martin, 1997). A striking example of the inability of co-authorship analyses to account for actual collaborative practice comes from the field of High energy physics. Within the last few decades, publications released by the CMS and ATLAS collaboration at CERN showed a strong rise in co-authorship (Kahn, 2018). Authorship lists in CMS and ATLAS publications currently feature over 6000 researchers. With the continuously growing size of the collaborations at CERN, it has become abundantly clear that only a small fraction of the collaboration members are working together in the production of a particular analysis. This invites the question of whether co-authorship, in the context of experimental HEP, has become entirely disconnected from the idea of interpersonal collaboration on a research project.

A move towards practice theory in social science research from the 1970s onwards has elicited a multiplicity of studies that interrogate the daily interactions of researchers (Nicolini, 2012). By focusing on "embodied, materially mediated arrays of human activity" (Schatzki et al., 2001, p.11), practice theory has introduced “a new vista on all things organizational (and social)” (Nicolini, 2012, p.2). In particular, scholars have argued that it is helpful to conceptualize collaborative activities as “multiple and entangled complexes of practices” (Chimirri, 2021, p.363). Analyzing collaboration through the lens of practice theory gives visibility to the situated and fluid character of research organization as being constantly made and remade through practice. Detailed historical and ethnographic studies on the social structures of High energy physics have demonstrated the highly complex and situated character of collaborative practice in HEP institutions (Galison, 1997; Knorr-Cetina, 1999; Traweek, 1988). They specifically demonstrate that co-authorship, in the context of HEP, indicates group membership but not necessarily personal collaboration on a research project. More generally, these accounts suggest that focusing on the research output of HEP collaborations is insufficient in describing which interactions occur during the research process. While changes in co-authorship might indicate transformations in the collaborative practices of HEP groups, the dynamics of these transformations remain unclear. Thus, an increase in the size of research collaborations does not implicate an increase in interpersonal collaboration between researchers.

In line with this qualitative body of work, this research project scrutinizes the day-to-day routines and interactions of researchers and understands research collaboration as the personal interaction of researchers during the research process.

2.3.2 Boundary work and boundary objects

With the notion of boundary work, Star and Griesemer (1989) aim to conceptualize how collaboration is enabled between different communities of practice. Star and Griesemer (1989) argue that boundary work can be understood as a set of practices for creating cooperation without consensus. By analyzing the interactions between scientists and nonscientists in the making of the Berkeley Museum of Vertebrate Zoology, they show how processes of “translation” helped scientists to enroll a diverse range of actors (Star & Griesemer, 1989). Translation, in this context, refers to the re-articulation of the interests formulated by actors in order to align them with the scientist’s goals. According to Star and Griesemer (1989), boundary objects often play a crucial role in such processes of translation. Boundary objects are “both plastic enough to adapt to local needs and the constraints of the several parties employing them, yet robust enough to maintain a common identity across sites” (Star & Griesemer, 1989, p.393). Thus, different actor groups employ varying understandings of a boundary object in local contexts. In order to cooperate with other social groups, a shared, flexible understanding is employed (Star, 2010).

The flexibility of boundary objects can take various forms. Star and Griesemer (1989) outline four distinct types of boundary objects. First, boundary objects can act as “repositories”, that is, “ordered ’piles’ of objects which are indexed in a standardized fashion” (Star & Griesemer, 1989, p.410). Actors can pick specific elements and use them without discussing their choices with other actor groups. A database is a good example of a repository boundary object. Actors with different backgrounds and interests gain access to large amounts of data which they can freely choose from. Second, there are "ideal type" boundary objects. These objects are not specific to a certain social context, but vague and therefore adaptable to various local environments (Star & Griesemer, 1989). For instance, phrases such as "openness" can be an ideal type boundary object as they acquire a multiplicity of meanings in different contexts due to their lack of clear definition. "Coincidental boundaries" are objects "which have the same boundaries but different internal contents" (Star & Griesemer, 1989, p.410). This means that specific aspects of an object are recognized by all parties, while other aspects are adapted to specific social contexts. For instance, Star and Griesemer (1989) describe how maps of California acted as coincidental boundary objects between amateur collectors and biologists. While all maps shared the same outline, they included different contents and thus allowed the different groups to define their meanings independently. Finally, "standardized forms" can serve as boundary objects. They equalize practices along temporal, spatial, and social dimensions and hence enable more seamless forms of collaboration (Star & Griesemer, 1989). Tuertscher et al. (2011) pick up on the relevance of standardized forms in research organization by interrogating dynamics of collaboration in the ATLAS group at CERN. They argue that the ATLAS architecture, as well as the structure of CERN as a whole, can be understood as a 'bottom up' approach to scientific collaboration "without centralized decision making [and] without a centralized control hierarchy" (Knorr-Cetina, 1995, p.125). Rather than a homogenous group structure, the ATLAS architecture is composed of modular substructures (Tuertscher et al., 2011). Modularity in this context refers to the development of partially autonomous subgroups that work together through "standardized interfaces that define the functional, spatial, and other relationships between them" (Tuertscher et al., 2011, p.78). This elaboration resonates with Star and Griesemer (1989) description of "standardized forms" as intermediaries between different communities of practice.

The more recent work of Merz and Sorgner (2022) underscores the centrality of standardization practices as ways of coping with organizational complexity at CERN. Merz and Sorgner (2022) understand institutional complexity as "diverging institutional logics" and argue that bureaucratic governance strategies, the "segmentation of infrastructure", and standardization practices are responsible for navigating rather than releasing this complexity. While the segmentation of infrastructure delineates different organizational logics, bureaucratic governance is crucial in defining priorities and particular courses of action. Standardization facilitates a transition between the different institutional logics.

Merz and Sorgner's (2022) elaboration thus picks up the idea of "standardized forms" as mediators between different organizational logics. The account further underscores the relevance of infrastructuring practices in demarcating particular dynamics. Both of these aspects are essential for my analysis of the CERN open data portal. The CERN ODP is largely disconnected from LHC internal data-sharing infrastructures and thus grants the developers a high degree of freedom in implementing new logics for data sharing and analysis. However, the ODP remains linked to internal storage infrastructures in order to guarantee the transfer of data. Connections between the ODP and internal infrastructures are established through standardization practices, but also through loosely structured forms of collaboration (See 6.1). Star and Griesemer's (1989) elaboration on boundary objects, as well as Tuertscher et al.'s (2011) and Merz and Sorgner's (2022) work on standards acknowledges the relevance of nonhumans in coordinating actor groups. Star and Griesemer's (1989) account underscores the volatility as well as the stability of nonhumans which I pointed to in the introduction of this chapter. However, besides their elaboration on standardized forms, Star and Griesemer (1989) focus particularly on the flexible characteristics of objects. In Star and Griesemer's (1989) conceptualization, boundary objects are just robust enough to maintain a stable identity across different sites. In the context of the CERN ODP, it is helpful to understand infrastructure elements as volatile boundary objects. In the first subsection of the findings chapter, I describe how the open data team implemented digital interfaces for the interaction between different LHC collaborations in the development phase of the portal. As I will show, the flexible design of these interfaces allowed the different collaborations to retain specific data practices without having to negotiate them with the open data team. However, the analysis of portal development practices additionally shows that standardization practices as outlined by Merz and Sorgner (2022) and Tuertscher et al. (2011) are essential in structuring the relationship between the LHC collaborations and the open data team.

Further, this project investigates how data itself becomes a boundary object that mediates between data production and use communities. Here, it is interesting to observe how data can be adapted to fit local use contexts and which aspects of the researchers' analysis practices remain stable across (re-)use contexts. The analysis shows that next to attaining particular types of meaning in specific use contexts, ODP data (on the research level) exhibits some stable qualities that align the analysis practices of data production and (re-)use communities. As such, data not only enables the coordination of various research interests but simultaneously shapes and aligns those interests.

2.3.3 Trading zones

Galison's (1997) research on the historical evolution of particle physics experiments resonates with Star and Griesemer's (1989) conceptualization of boundary work. Similar to Star and Griesemer (1989), Galison (1997) aims to understand how different domains in particle physics, such as theory, experiment, and computation, can cooperate during the course of an experiment without explicit consensus. Galison (1997) describes how the traditional roles of the experimenter as the producer of experimental data and the theorist as the interpreter have become increasingly blurred from the 1980s onward. At the time, the experimental validation of new theories such as Quantum Chromodynamics (QCD) grew increasingly challenging (Galison, 1997). Data collected by particle detectors continuously increased in size and complexity and, as such, became more and more difficult to analyze and interpret. Experimentalists needed to develop tools for delineating "background" and "detector effects" from the signals caused by particle collisions. In order to do so, they developed increasingly sophisticated Monte Carlo simulation techniques. As elaborated in section 2.2, Monte Carlo techniques simulate particle interactions by using a particular theoretical input. Based on principles of random number generation, they model collisions over large numbers of particles. The Monte Carlos generated by experimentalists additionally included detailed knowledge of detector effects. As such, they could be compared to data taken by the detector and allowed the experimentalists to delineate theoretical assumptions from detector effects.

At the same time, theorists used Monte Carlo simulations to model their theoretical assumptions over a large number of collisions. This allowed them to attain stable probability distributions for particular theoretical inputs. In the particle physics community at the time, many debates were held on the role of Monte Carlo simulations. It was unclear whether they should count as part of experimentation or theorization. Galison (1997) argues that Monte Carlo simulations ultimately started to form a "trading zone" between the theorists and experimentalists. Galison (1997) conceptualizes trading zones as "local sites of coordination" between otherwise heterogenous groups with conflicting practices and understandings. In the realm of simulation, experimentalists and theorists shared common terms and concepts to communicate their interests. Outside of this local setting, they grew increasingly estranged from each other's practices.

This is where Galison's (1997) elaboration intersects with Star and Griesemer's (1989) conceptualisation. Rather than assuming a global homogenization of practices, both concepts conceptualize successful cooperation between different communities of practice as locally situated phenomena. However, Galison (1997) moves away from Star and Griesemer's (1989) focus on objects as intermediaries between different cultures and particularly focuses on the development of common languages and concepts. Nonetheless, materiality remains a crucial force in shaping the trading zone and between theorists and experimentalists.

Galison (1997) describes how the field of Monte Carlo simulation grew to a point where scientists "could make a living in the trading zone" (p.770). According to Galison (1997), Monte Carlo specialists were neither experimentalists nor theorists. Rather, they developed a separate, distinct "professional identity". While this is certainly still the case in contemporary particle physics, I want to argue that Galison's (1997) conception of simulation as a trading zone between experiment and theory has undergone some changes. The increasing complexity of detector technologies that Galison (1997) outlines in his account has taken up speed in the last 25 years. This strongly impacted the Monte Carlo simulations that reflect this high complexity by reconstructing the impact of a particle detector on collision events. Working with the simulations of the experimentalists hence requires long periods of engagement and intricate knowledge of the detector technology. On the other hand, theorists have further developed contrasting Monte Carlo simulations that do not consider detector effects. Thus, Monte Carlo production techniques have become increasingly estranged from one another. This raises the question of whether Monte Carlo specialists have to position themselves as either experimentalists or theorists rather than defining themselves at the intersection. While experiment and theory remain dominant categories in particle physics, trading seems to happen rarely via simulation. Theorists might consider published results of experimentalists in their work, but they rarely engage in their data. The open data portal is an example where theorists have used experimental simulated data. As such, it is an infrastructural setting that opens up a local trading zone between experimentalists and theorists. The upcoming analysis will demonstrate that this trading zone exhibits very specific characteristics, in which theorists, in particular, have to undertake extensive efforts to facilitate exchange.

Galison's (1997) elaboration can be understood as a criticism of both, positivism and antipositivism. Rather than understanding the experiment as the driver of theory or theoretical concepts as the drivers of the experiment, Galison (1997) seeks to acknowledge the partly disjointed nature of these two domains. Through such an understanding, he picks up earlier works on particle physics such as those of Pickering (1984) but modifies them in important ways. Pickering (1984) argues that any empirical data allows for a variety of theoretical explanations. The decision on a theory thus depends on the "dynamics of practice": if a theory builds on existing theoretical expertise and if it enables the development and testing of a new experimental apparatus. The choice of a theory for Pickering (1984) is hence a social choice, aiming to sustain the traditional relation between theoretical and experimental practice. Contemporary particle physics practice, however, proves how theories that do not promise experimental validation in the next decades, if ever, have become quite successful. String theory and other theoretical concepts have managed to defend their place in theoretical particle physics for several decades now, even though experimental validation is far out of sight. Contemporary particle physics thus serves as a striking example of the rapidly changing relationship between theory and experiment. Rather than occupying never changing spaces of limited agency, theorists and experimentalists have inhibited historically contingent spaces that follow their own logics.

Galison's (1997) work underscores this point strongly. Solely connected through local trading zones, the different subcultures of physics operate according to their own logics, sometimes aligning and sometimes clashing with one another. In her account on scientific cultures, Knorr-Cetina (1999) asserts the disunity of the sciences by outlining differing epistemic practices in particle physics and biology. Galison (1997) has gone even further by asserting disunity within the field of particle physics itself. Understanding particle physics as segmented and contingent is crucial for my work. In this thesis, I will engage in the narratives of researchers that work in vastly different epistemic traditions (information scientists, IT specialists, theorists, experimentalists, and machine-learning experts). They are at very different stages in their career (pre-doctoral researchers, senior scientists) and work in the context of different institutional settings (CERN, different university settings). It is a presupposition of this work that these different groups will articulate a variety of socio-epistemic standpoints. Consequently, my interest lies in the ways in which these different standpoints are negotiated and coordinated in practice.

2.3.4 Interlaced knowledges

Galison's (1997) account demonstrates how not only material objects, but also shared concepts can create togetherness between different communities of practice. Through developing a shared language, actors with conflicting interests communicate their positions and develop so-called "interlaced knowledge" that allows them to gain a sufficient understanding of each other's interests (Tuertscher et al., 2011). This resonates with Bressan and Boisot's (2011) more recent account on individual and collective learning in the ATLAS collaboration at CERN. Bressan and Boisot (2011) outline how a pool of "shared knowledge, values and norms"(p.203) is necessary to facilitate research in the ATLAS group. At the same time, researchers are expected to bring a variety of specializations to the table to make an individual contribution. Thus, Bressan and Boisot (2011) conceptualize knowledge in the ATLAS collaboration as only being shared at particular points of interest. Bressan and Boisot (2011) additionally point to the relevance of tacit knowledge that circulates within the ATLAS collaboration. Many shared norms, beliefs, and knowledges are not formalized and written down but need to be acquired in personal interactions between researchers.

Bressan and Boisot (2011) demonstrate the relevance of distributed, flexible, and embodied knowledge in the ATLAS collaboration. However, they understand collaboration and learning as mainly constituted by face-to-face interactions between researchers. This project, however, will focus on the relevance of nonhumans in engendering learning processes. It will thus question whether particular kinds of knowledge can only be transferred through personal interaction or if they can also be distributed via intermediaries. Bressan and Boisot (2011) additionally point to the relevance of norms and values in the organization of research at CERN. In the next subsection, I explore literature from the field of valuation studies and ask how these works can inform a study of the forms of togetherness that arise through data practices at the CERN ODP.

2.3.5 The role of norms and values on forms of togetherness

Thus far, I have outlined how nonhumans, such as data, metadata, standards, concepts, and interlaced knowledges, mediate and direct interaction in particle physics research. I have further outlined how these human-nonhuman encounters facilitate new forms of togetherness. In this section, I want to reflect on the ways in which data practices are guided by values and norms that circulate in CERN and the broader scientific community. My empirical analysis suggests that research evaluation strategies inside and outside of HEP are essential in defining how particular research groups use CERN open data.

Boltanski and Thévenot (2006) have investigated how valuation matters in consent processes. They argue that there are a variety of incongruous "modes of justification" that can be mobilized in any valuation situation. There is no objectively right register of justification. However, there are dominant justificatory practices in particular settings. The success of an argument often depends on the ways in which actors manage to ground their narratives in the "orders of worth" that are dominant in a particular situation. Importantly, Boltanski and Thévenot (2006) argue that the acceptability of a particular valuation practice is materially embedded and dependent on the sociohistorical context. In relation to the CERN open data portal, Boltanski and Thévenot's (2006) work raises the question of which justificatory modes become dominant in the preparation and use of data for the portal, how they clash and how they are negotiated. Is the open data portal understood as a distributor of epistemic worth, an accelerator of innovation, or as a way of preserving the scientific heritage of CERN? And how are these values inscribed into the portal's infrastructure and the data supplied through it?

In order to capture the ways in which valuation practices in the context of the ODP relate to CERN as an institution and the scientific community at large, it is useful to consult Fochler et al.'s (2016) work on evaluative principles in the life sciences. Fochler et al. (2016) suggest that "(e)valuative principles" as well as "regimes of valuation" are relevant for the investigation of valuation practices. While "evaluative principles" reflect the particular situations in which valuation processes take place, "regimes of valuation point to the broader discursive, material and institutional background this concrete evaluation draws on" (Fochler et al., 2016, p.180). This distinction allows Fochler et al. (2016) to account for the situated character of valuation as well as its broader context. In this case study, I want to consider how the value systems of researchers who maintain and use the CERN ODP are embedded in larger "regimes of valuation". More specifically, I want to question how CERN internal values relating to the publication of research results and evaluative principles for research outputs that prevail within the broader scientific community are newly negotiated in the context of the CERN open data portal.

A value that is crucial to HEP research practices is collectivity. In the context of this project, I want to explore how circulating enactments of collectivity intersect with data journeys and shape ODP data practices. Knorr-Cetina (1999) has shown how High energy physics research is largely based on a "communal" culture that enacts the research collaboration as the agent of knowledge production. Researchers are thus subordinated to a larger whole; they are subsumed under a collective identity. This collectivity is reflected in many aspects of knowledge production at CERN, e.g., the internal peer review structures (See 6.2). In the context of the ODP, I want to understand how data practices enact particular understandings of collectivity and how understandings of collectivity, in turn, shape data practices. I specifically pay attention to the ways in which individual researchers reemerge as epistemic subjects. Since the ODP is situated at the intersection of CERN and the broader scientific community, the regimes of valuation that prevail outside of HEP will come to matter for researchers who work with data from the portal. In my empirical analysis, I will particularly focus on the ways in which individual, quantitative modes of research evaluation that are dominant in scientific areas other than HEP impact the ways in which researchers work with open data.

Galison (2003) has linked enactments of collectivity in HEP research to the notion of credibility. He argues that a "fragmented 'we'" in HEP would cause distrust in a research result and "erase" individual scientists "as contributing member[s] of the research community" (Galison, 2003, p.336). By drawing on this insight, I want to understand how research produced through the ODP is valued in the HEP community. Given that CERN open data publications largely circumvent collective review processes (See 6.2), I set out to explore if the research results generated from open data remain valid in the eyes of particle physicists.

3 Research questions

This study has departed from an interest in the ways in which data practices at the CERN open data portal enact the socio-epistemic structure of research. The literature I have reviewed has outlined various vocabularies for describing organizational and epistemic transformations resulting from the implementation of a new infrastructure for data sharing and use. In order to understand the more subtle rearrangements in organizational practice that go along with the development of the CERN open data portal, I decided to focus on the term "forms of togetherness" which enables me to describe how actors become engaged and estranged from one another through particular data practices. Additionally, the notion "forms of togetherness" allows me to retain a broad analytical focus and investigate not only how different forms of interpersonal collaboration are enacted by data practices but additionally how mediated types of interaction are performed by the data practices at the CERN open data portal. The main research interest of this work is thus *how data practices at the CERN ODP enact different forms of togetherness.*

The first three subchapters of the empirical analysis focus on particular types of data practice and the specific forms of togetherness that emerge from them. In the first chapter, I discuss the CERN open data team's portal development and maintenance practices. In this context, it is particularly interesting to investigate how interpersonal collaboration in the development phase of the portal turned into materially mediated forms of interaction. The guiding question of this subchapter is thus *how portal development and maintenance practices enact interpersonal as well as materially mediated forms of research collaboration.*

The second chapter focuses on the open data (re-)use practices of researchers who are members of an LHC collaboration. In this context, it is of interest to observe how understandings of collectivity and individuality are reconfigured through the use of the open data portal. The general question for this section is *how data practices enact the relationship between the LHC collaborations and LHC members who use open data.*

In the third subchapter of the empirical part, I focus on *how data practices enact the relationship between theoretical and experimental particle physicists who use and produce CERN open data.* Here, it is particularly intriguing to explore how the common use of data aligns the epistemic understandings of researchers and creates a new form of togetherness between theorists and experimentalists.

4 Theoretical framing

As outlined in chapter 2, “forms of togetherness” are both epistemic and social. They describe how actors relate to one another and simultaneously point to the types of knowledge that emerge from the resulting actor constellations. In scrutinizing how “forms of togetherness” are enacted by data practices at the CERN ODP, I acknowledge the entanglement of technological, social, and epistemic orders. This project thus draws on one of the crucial insights of actor-network theory (ANT) (e.g., Latour & Woolgar, 1979): the tenet of “free association [which] refuses any *a priori* distinctions between what could count as social, natural, or technological”(Michael, 2016, p.34). Rather than presupposing clearly delineated realms, I aim to understand how social, epistemic, and technological domains relate to one another in concrete settings. ANT teaches us that “the ’social’ is not given but a heterogenous product laden with the nonhuman-technologies and natures [which] are as much part of society as humans” (Michael, 2016, p.4). Thus, studying the ’social implications’ of academic research not only means studying the interactions of researchers. Rather, investigating the social means simultaneously investigating the epistemic and the technological.

According to Latour and Woolgar (1979), this is achieved through tracing relations between human and nonhuman actors. As Michael (2016) outlines, following these relations means attending to processes of “circulation - circulation of people, texts, objects and artifacts, bits of natures and cultures” (Michael, 2016, p.5). Since such circulations are contingent and complex, they need to be studied locally and in their particularity. In following data journeys and attending to the ways in which human labor is shaped by and simultaneously sustains such data journeys, this project takes up the suggestion to trace relations. The CERN open data portal is an instance where a variety of human (e.g., CERN external researchers, CMS researchers) and nonhuman actors (e.g., datasets, infrastructure elements) get connected in new ways and as such, it is an intriguing site for the study of circulation processes.

It is crucial to acknowledge that circulation processes at the CERN ODP are strongly shaped by its infrastructural setting. The interaction of human and nonhuman actors at the ODP relies on the functionalities defined by the technological makeup of the portal. The CERN open data portal mediates between the practices of data production and data (re-)use communities. As an infrastructure that sit between those communities, it crucially shapes data (re-)use groups and the ways in which these groups can access and analyze data. STS literature on infrastructures has investigated how infrastructuring processes influence the possibilities and outcomes of research. In the following subsections, I investigate how this literature can help in pinpointing the kinds of action that are enabled or disabled through the CERN open data portal.

Since humans and nonhumans are often strongly entwined in concrete research settings, ANT approaches suggest that we think of them as “hybrids”, rather than clearly delineated entities

(Michael, 2016). In line with this consideration, this project mobilizes the term "infrastructure" as well as the more processual notion "infrastructuring" to indicate that the continuous work that goes into the maintenance of infrastructures and the resulting technological interface are hard to separate from one another. As the following chapter will show, attention to the types of labor that sustain particular infrastructures are ways of rendering them most visible.

The first section of this chapter outlines some crucial characteristics of infrastructures that help make sense of portal development practices and the analysis of data (re-)use cases. The second part of the chapter focuses mainly on digital infrastructures, and the third subchapter explores literature concerning large-scale European research infrastructures to account for the ODP's embeddedness in the larger context of CERN.

4.1 Central characteristics of infrastructures

Star and Ruhleder (1996) have defined infrastructures as embedded, often invisible entities that demonstrate temporal and spatial stability. The invisibility of infrastructures has several interconnected dimensions (Karasti et al., 2016b). First, infrastructures are often used without being recognized. They are structures that enable us to take action without deeper consideration of their meaning. In order to make them visible, Star and Ruhleder (1996) argue that it is helpful to attend to moments of breakdown, in which infrastructures become most apparent. Paying attention to breakdowns not only means focusing on instances where infrastructures cease to function entirely but, more generally, attending to frictions that arise when actors try to use infrastructures.

Second, the work that goes into the construction and maintenance of infrastructures frequently remains unnoticed (Karasti et al., 2016b). Thus, studying infrastructures not only means investigating their material representation but also attending to the practices that create and sustain them. Rather than understanding infrastructures as fixed systems, it is often helpful to think about them as being maintained by constant processes of "infrastructuring". In this research project, I thus focus on data practices that concern the construction and maintenance of the CERN portal. I understand the resulting technological infrastructure, the CERN open data portal, as an additional resource to substantiate my claims rather than my main research objective. While an investigation of development and maintenance practices allowed me to understand how practices become inscribed into material infrastructures, through attending to practices of open data (re-)use, I could, in turn, study how the resulting infrastructure produces organizational forms and determines practices.

As Star and Ruhleder (1996) argue, infrastructures are always connected to larger infrastructural settings. Even though we often do not recognize them as such, infrastructures are everywhere (Star & Ruhleder, 1996). They form interdependent structures and should not be analyzed as standalone entities but within the context of their environment (Star & Ruhleder, 1996). Research activities often involve “multiple, coexisting, nonconforming infrastructures which actors engage [in] at the same time” (Vertesi, 2014, p.264). This non–conformativity does not infrequently demand creative solutions from researchers who seek to craft connections between different infrastructures (Vertesi, 2014). The omnipresence, intertwinedness, and non-conformativity of infrastructures imply that delineating a particular infrastructure for analysis is always somewhat arbitrary. Thus, the ways in which researchers choose to frame their analysis is not self-evident and has a performative effect on the research result (Karasti et al., 2016b). As Vertesi (2014) suggests, it can be particularly interesting to study where infrastructures do not align with one another and efforts are undertaken to create connections. This resonates with (Bowker & Star, 1999) suggestions to study the “resistances” (p.39) that arise when infrastructures are used. Which desired ways of action are inhibited by an infrastructure? When do friction and rupture arise? My choice of the CERN open data portal as a research site reflects this interest in infrastructural resistances. The ODP forms a “bridge” between the different internal data infrastructures of the LHC collaborations and a variety of open data user groups with vastly differing approaches to data analysis. Hence, rupture and resistance are likely to arise. They will not only allow me to gain insights into the infrastructure of the portal but additionally shed light on its connection to internal storage structures.

Star and Ruhleder (1996) argue that infrastructures are, to some degree, stable across time and space. The spatially distributed quality of infrastructures has been studied extensively in STS research (e.g., Larkin, 2013). Infrastructures connect places and disconnect others. They create new places that often reimagine space along the lines of dominant power hierarchies. The temporal dynamics of infrastructures have received less scholarly attention. However, more recent STS studies have shown that infrastructures “mediate[] time as much as [they] mediate[] space”(Appel et al., 2020, p.15). For instance, CERN internal infrastructures mediate between the microscopic, relativistic time of particle collisions and the non-relativistic, macroscopic analysis timescales of HEP research (Traweek, 1988). One of the major goals of the CERN open data portal is the long-term preservation of LHC datasets. Therefore, next to enabling public access to CERN resources, ODP development efforts are geared toward extending dataset lifetimes and, as such, transforming the temporal dynamics of HEP data (re-)use practices. As (Appel et al., 2020) argue, temporal and spatial dynamics of infrastructures are always interlinked. The spatial extension of an infrastructure always follows particular temporalities that matter for the resulting use dynamics of an infrastructure. For instance, the upload of datasets to the open data portal is governed by an embargo period within which datasets are exclusively available to the collaboration. This informs the upscaling of the ODP in terms of storage capacity and is crucial in defining open data (re-)use communities.

Focusing on the temporal dynamics of infrastructures not only includes investigating how the material construction of infrastructures takes place over time or how infrastructures transform the temporalities of the objects and subjects that are connected to them. Additionally, it is crucial to consider how imaginations of the future are inscribed into infrastructural settings. Aula (2019) has argued that processes of infrastructuring are always processes of materializing specific futures, and as such, they entail complex decision-making processes. As a consequence of infrastructuring, some futures are enabled, and others are circumvented. Infrastructuring is, therefore, a deeply political process that shapes our very perception of the world. Similarly, Gupta (2020) has ascertained that infrastructures "are often shaped by and simultaneously shaping understandings of modernity, progress and desirable ways of life" (Gupta, 2020). In the development of infrastructures, anticipating certain futures entails the establishment of particular values. In the interviews I conducted for this project, some members of the ODP development team envisioned the future of scientific research as one where academic resources circulate freely through different societal domains. As Karasti et al. (2016a) points out, "openness" is a value frequently embedded in contemporary knowledge infrastructures. However, "translating the values of openness into the design of infrastructures and the practices of infrastructuring is a complex and contingent process." (Karasti et al., 2016a, p.6). Values and envisioned futures intersect with anticipations of the actors and objects involved in a concrete research setting, and together, they inform situated implementation processes. Therefore, it is not only crucial to investigate which kinds of futures are imagined but additionally how they are inscribed into a particular infrastructure. In the context of this research project, it was interesting to observe how values such as openness needed to be negotiated in relation to the particular kinds of data and (re-)use communities of the CERN open data portal.

Gupta (2020) has argued that the temporalities of an imagined future and the temporalities of concrete infrastructuring processes sometimes fail to align with one another. As a consequence, "ruins" arise. This is an interesting consideration in the context of the CERN open data portal. An imagined open data future entails an understanding of data as a long-living entity that can be exploited without the knowledge of the original data producer. However, extending the lifetimes of LHC resources has proven to be quite challenging. Changing software formats render older data versions extremely hard to work with, and systems for translating different software versions into one another need to be developed. If such creative processes of translations fail to take place, the open data team risks the production of "data ruins": data that can not be used due to software incompatibility issues.

4.2 Digital infrastructures

Digital infrastructures have taken on a particular role in contemporary research environments. Hine (2014) has argued that information and communication technologies (ICTs) have become essential elements of “the complex political landscape and funding climate” that researchers have started to inhabit. Often, ICTs are understood as markers of “efficiency and progress” (Hine, 2014). Additionally, ICTs are connected to ideas of “seamless access” (Hine, 2014) and limitless connectivity. However, these idealized anticipations fail to account for the impact of particular research settings on shaping such digital infrastructures. As Hine (2014) argues, we should not consider digital infrastructures as isolated agents of transformation but rather as technologies that are mobilized to advance situated research agendas. An investigation of ICTs should crucially focus on how they are embedded in the daily work practice of researchers and how they are used as tools to establish new types of socio-epistemic practice. Additionally, it is essential to account for the ways in which circulating imaginaries of efficiency, progress, and seamless access influence the concrete implementation of ICTs.

Since ICTs acquire a multiplicity of forms in different research environments, they can serve various purposes. For instance, Hine (2014) argues that ICTs can be used as tools for manifesting institutional and disciplinary boundaries. High energy physics is regularly portrayed as being at the forefront of computational progress. The development of the world wide web at CERN is an often-told story that underscores the innovative potential of particle physics. Thus, it gives HEP a distinct identity as an innovator in computational fields.

Additionally, digital infrastructures such as databases define particular access, upload, and (re-)use rights that craft interdependent communities of shared digital practice. Membership in an LHC collaboration is partly constituted by access to the digital infrastructures that store research data. As Hine (2014) argues, ICTs are used to “explore a broad range of potential audiences and to articulate [...] relationship[s] with those audiences” (p.244). The CERN open data portal, in particular, reimagines the (re-)use audiences of LHC data. Data, which was previously restricted to the LHC collaborations, is made available to external communities via the ODP. However, data (re-)use through the ODP should not be understood as a way of eradicating the boundary between the LHC collaborations and non-LHC ODP user groups. To characterize ODP (re-)use dynamics, it is crucial to attend to the specific rights and obligations inscribed in the portal. For instance, there is a clear distinction between data producers and users. The LHC collaborations provide the resources for the portal, and CERN as a laboratory provides the computational infrastructure. Users can download and repurpose data but are unable to modify resources. When data from the ODP is repurposed, users are asked to acknowledge its origins by referring to the LHC collaborations. Thus, a clear delineation between data producers and users remains intact. The timescales of data release are similarly important for understanding the hierarchies inscribed into the open data portal. During an embargo period, data can solely be exploited within the LHC collaborations. Therefore, the self-understanding of the LHC collaborations as data-producing communities, data owners, and first analyzers of LHC data remains intact.

Additionally, the CERN open data portal puts the clear delineation between research process and product into question. As Hine (2014) argues, ICTs are often used to “blur the distinction between ongoing work and finished product, adapting the traditional time scale of taxonomic work to an environment that expects demonstrable results in the short term” (p. 244). The CERN open data portal enacts research data, as opposed to LHC publications, as a publishable product in its own right. This is well illustrated by the data citation practices that are recommended on the portal. Each dataset has a unique “digital object identifier” (DOI) that allows researchers to cite data in a uniform way. Datasets thus become publishable and citable items. However, the ability of datasets to act as independent scientific representations is not without contestation in the CERN community. Several of my interviewees have argued that it is challenging to detach HEP data from its local research environment and successfully mobilize it in other contexts (See subchapter four of findings). Digital infrastructures such as the open data portal are hence also a site where understandings of what constitutes a generalizable and objective scientific product are negotiated.

The arguments presented above underscore how digital infrastructures serve multiple and situated purposes. They can diffuse or reinstate institutional and disciplinary boundaries, they can serve as a way of connecting individuals and groups, and they can redefine what counts as the product and the process of knowledge production. Miller and Slater (2000) argue that “it is important to understand the Internet as a symbolic totality as well as a practical multiplicity” (p.16). This holds true for digital infrastructures. While they are often understood in general terms as ways of enabling “efficiency and progress” and “seamless access” (Hine, 2014), they need to be scrutinized in the context of their environments in order to be fully understood.

4.3 (European) research infrastructures

The general aim of this subsection is to think about the ways in which large scale research infrastructures such as CERN have been conceptualized theoretically. While this thesis focuses on one particular part of CERNs digital infrastructure, the open data portal, it is important to understand how the general infrastructural setting affects processes on the ODP. Generally, this section is informed by the tenet that all “knowledge infrastructures are performative of the knowledge being produced” (Karasti et al., 2016a, p.5). For instance, data infrastructures supply particular types of data and metadata and thus, enable specific types of analysis while hindering others. Hence, the question arises what types of research are enabled through infrastructural practice. Additionally, knowledge infrastructures are themselves shaped by social, political, technological and epistemic aspects. Particularly in the context of large scale research, political decision making processes and goals are crucially influencing the development of scientific agendas.

4.3.1 European integration and CERN

CERN is a striking example of how technoscientific and sociopolitical orders depend on one another in the context of large-scale research. Founded in the aftermath of the second world war, CERN was, from the onset, motivated by the ambition to bring unity to the post-war European nations. As Krige (2003) argues, particularly because science was understood as a neutral, unpolitical activity, it was well suited as a first move to the political integration of Europe. The importance of CERN in re-uniting post-war Europe resonates with the work of scholars who have argued that the integration of Europe is a technological as well as a political process (Hallonsten, 2012, 2020). Misa and Schot (2005) assert that European integration needs to be understood as "an emergent outcome of a process of linking and delinking of infrastructures" (Misa & Schot, 2005, p.1). Rather than conceptualizing Europe as a predefined entity, the authors understand Europe as being constantly made and remade through practice. Misa and Schot (2005) particularly argue that integration is achieved through "the circulation and appropriation of knowledge and artifacts" (p.8). Thus, Misa and Schot 's (2005) account demonstrates how political and technoscientific orders interact with one another. Political integration only becomes possible by establishing technoscientific infrastructures that circulate particular knowledges and objects.

This resonates with Barry's (2006) work which introduces the concept of "technological zones". According to Barry (2006), European integration is made possible through the development of zones within which "differences between technical practices, procedures and forms have been reduced, or common standards have been established" (Barry, 2006, p.249). Technological zones can take on the form of shared ways of measurement, the setting of common standards, and the development of practices in accordance with those standards. These shared forms of practice are always already political since they advance processes of integration. They are additionally crucial in directing developments in science and technology. Within big science projects such as CERN, common standards are developed and deployed in daily practice. They are crucial in allowing researchers to step outside their disciplinary and organizational contexts and thus enable communication and coordination. The development of common standards and a shared, interconnected infrastructure are crucial objectives of the CERN open data project. However, due to the marginal position of open data endeavors in CERN's research agenda, they are unlikely to contribute to grand sociopolitical developments such as European integration.

Mobach and Felt (2022) have outlined how the co-productive relationship between Europe and CERN has transformed throughout CERN's existence. Originally envisioned as a project "in which European unity could be accomplished through science" (Mobach & Felt, 2022, p.382), the focus shifted in the 1990s when CERN began to showcase itself as a 'laboratory for the world' (Mobach & Felt, 2022, p.382) in which "Europeanness" became a resource in arguing for the advancement of particle physics research. Hence, whereas particle physics was originally seen as a way of advancing European integration, later on, European values were understood as vehicles for promoting global particle physics practice.

The development of the CERN open data portal was informed by the idea of making CERN resources globally accessible. In the context of the ODP, CERN is hence strongly imagined as a “laboratory for the world” where LHC data, understood as intrinsically valuable, is globally distributed. Mobach and Felt (2022) show how early anticipations of CERN and its function for European society were tied to the idea of open and free exchange of research. Europeanness was taken as a set of shared cultural ideas that did not need to be constructed but solely resurrected through an open exchange of already existing ideas (Mobach & Felt, 2022). Thus, openness can be understood as a value that is narratively connected to a European way of life and reflected in the history of CERN and its distinctly European self-understanding. Therefore, in this project, it is of particular interest to observe how openness becomes newly articulated in the context of the CERN open data portal and how this articulation is situated in an anticipation of CERN as a distinctly European research infrastructure.

5 Methods

This project is based on a multi-method approach. I initially explored the field through a short, quantitative assessment that identified actors involved in the production and use of CERN open data. Subsequently, I conducted semi-structured interviews with CERN ODP data producers and users. In order to substantiate the findings generated from the interview material, I additionally worked with openly accessible conversations about ODP development and maintenance that were held on Github (CERN open data, 2023b)[1] and Gitter (CERN open data, 2023a). I further used parts of ODP's source code (available on Github) to underscore specific aspects of the interviewees' narration. The upcoming chapter outlines these different methodological approaches and presents the quantitative findings.

5.1 Quantitative analysis

Portal development and maintenance

The development of the open data portal was realized as a joint effort of the CMS open data team, the CERN scientific information service, and the CERN IT department (Rao et al., 2019). By identifying publications that have focused on the development of the ODP, I could determine some of the key actors. As Figure 1 indicates, the collaborative structure gravitates around Kati Lassila-Perini, the Coordinator of the CMS Data Preservation and Open Access project, Thomas Mccauley, a CMS scientist, Tibor Šimko, a Computing Engineer at CERN and Sünje Dallmeier-Tiessen, the data coordinator of the Scientific Information Service at CERN.

Uploaded data

I could identify other relevant actors by analyzing the datasets available on the portal itself. Most datasets on the portal are assigned a unique digital object identifier (DOI). This makes it possible to cite them in a standardized, pre-defined way (Rao et al., 2019). Authorship of datasets is usually attributed to the respective LHC collaboration rather than an individual author. Out of 9356 datasets available on the portal, only 144 (circa 1,5 percent) are individually authored, all derived datasets produced by the CMS collaboration. Of the 144 datasets that individual researchers author, 131 are single-authored, and 13 have multiple authors. In Figure 2, the authors of these datasets are listed.

Data (re)use

Table 2 (Appendix A) lists publications that have used CERN open data. Figure 3 depicts a co-authorship network of the researchers who wrote these publications. As is visible in Figure

[1] Github is a software development platform that enables the collective generation of code. It is also possible to raise specific issues on Github and assign responsibilities for their resolution.

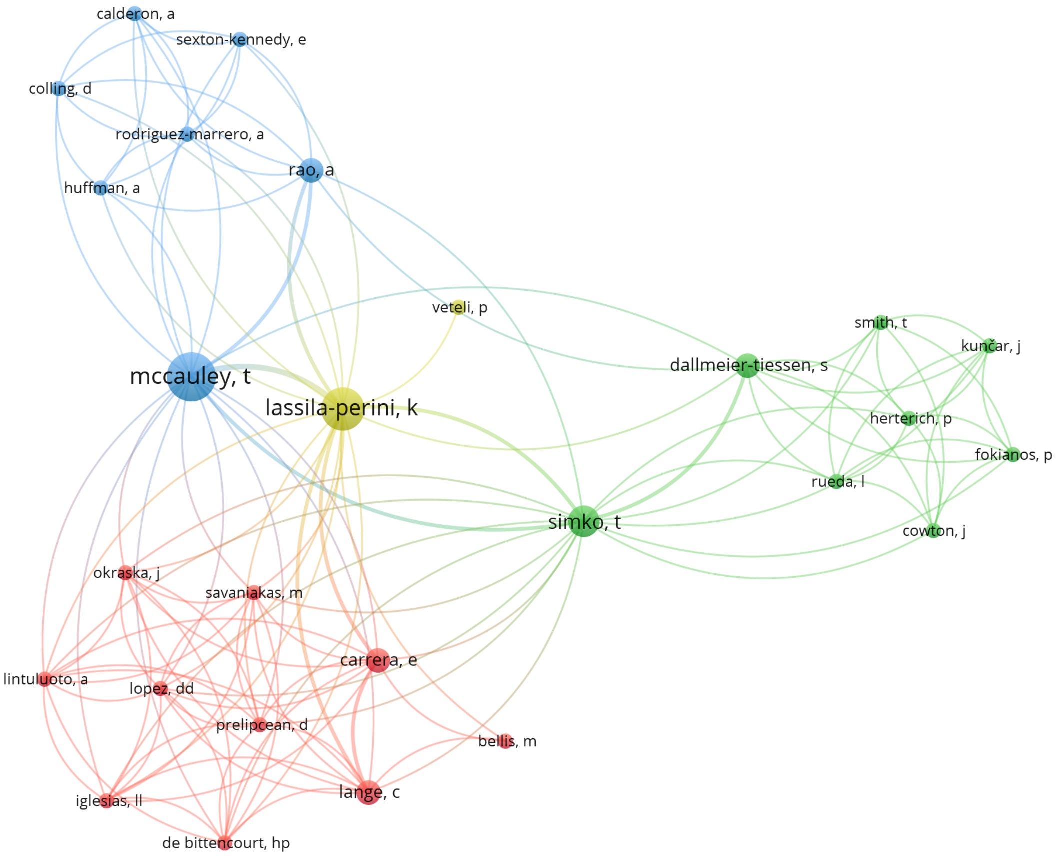
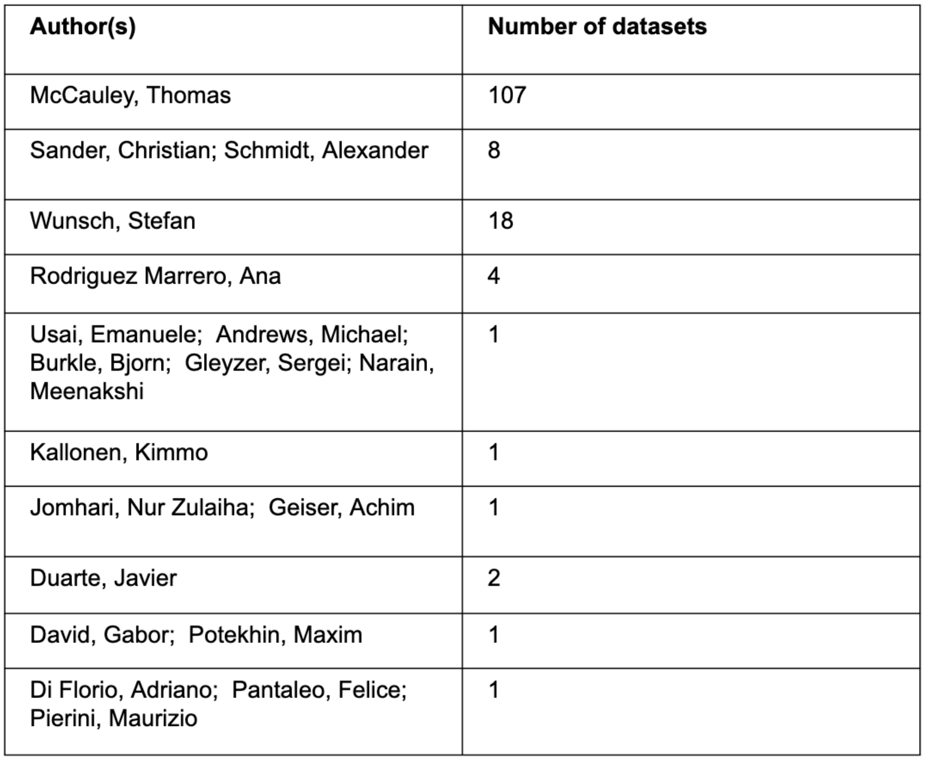


Figure 1: Co-authorship network of 16 publications concerned with the development and maintenance of the CERN open data portal (specifically focused on CMS open data); the size of the nods corresponds with the amount of publications written by the authors; non co-authoring researchers are not displayed; created with VOSviewer

3, many open data publications focus on topics situated at the intersection of High energy physics and data science. Machine learning techniques are a particularly strong focus of the research with open data. Additionally, research at the intersection of theoretical and experimental particle physics was conducted through the use of open data. The actor networks in 3 served as starting points for the qualitative part of my thesis. They were particularly useful in allocating respondents for the semi-structured interviews.

5.2 Participant observation vs. Interviews

While ethnographic methods are often considered the "golden standard" in practice-oriented research (Halkier, 2017), data practices can pose particular challenges to observational techniques. In the following section, I reflect on the problems that an analysis of the ODP poses to an ethnographic research approach.



Author(s)	Number of datasets
McCauley, Thomas	107
Sander, Christian; Schmidt, Alexander	8
Wunsch, Stefan	18
Rodriguez Marrero, Ana	4
Usai, Emanuele; Andrews, Michael; Burkle, Bjorn; Gleyzer, Sergei; Narain, Meenakshi	1
Kallonen, Kimmo	1
Jomhari, Nur Zulaiha; Geiser, Achim	1
Duarte, Javier	2
David, Gabor; Potekhin, Maxim	1
Di Florio, Adriano; Pantaleo, Felice; Pierini, Maurizio	1

Figure 2: Individually mentioned authors of CERN ODP datasets

First, the spatially distributed quality of digital research infrastructures often complicates observation in physical co-presence. The actors that are crucial in facilitating the movement of data through digital infrastructures, such as the open data portal, are distributed all around the globe. Therefore, observation in direct co-presence would only be possible for a fraction of all relevant actors.

Additionally, human-computer interactions often remain opaque when observed from an outsider's standpoint. Classic STS studies based on participant observation have focused on laboratory practices where various artifacts and machines are used and manipulated. Digital practices are mediated solely through one, often opaque nonhuman: the computer. Rather than observing researchers' interaction with the larger environment of the laboratory, during a participant observation of digital practices, one might just encounter a scientist typing cryptic phrases into a machine. In the context of the CERN ODP, the preparation and use of highly complex datasets are central research objectives. An investigator who is largely unfamiliar with data generation, maintenance, and analysis techniques in HEP is likely to gain little from observing researchers' interactions with their computers.

Next to these more general considerations, there are also practical issues to consider. While a participant observation, albeit its limitations for observing data practices, would surely yield

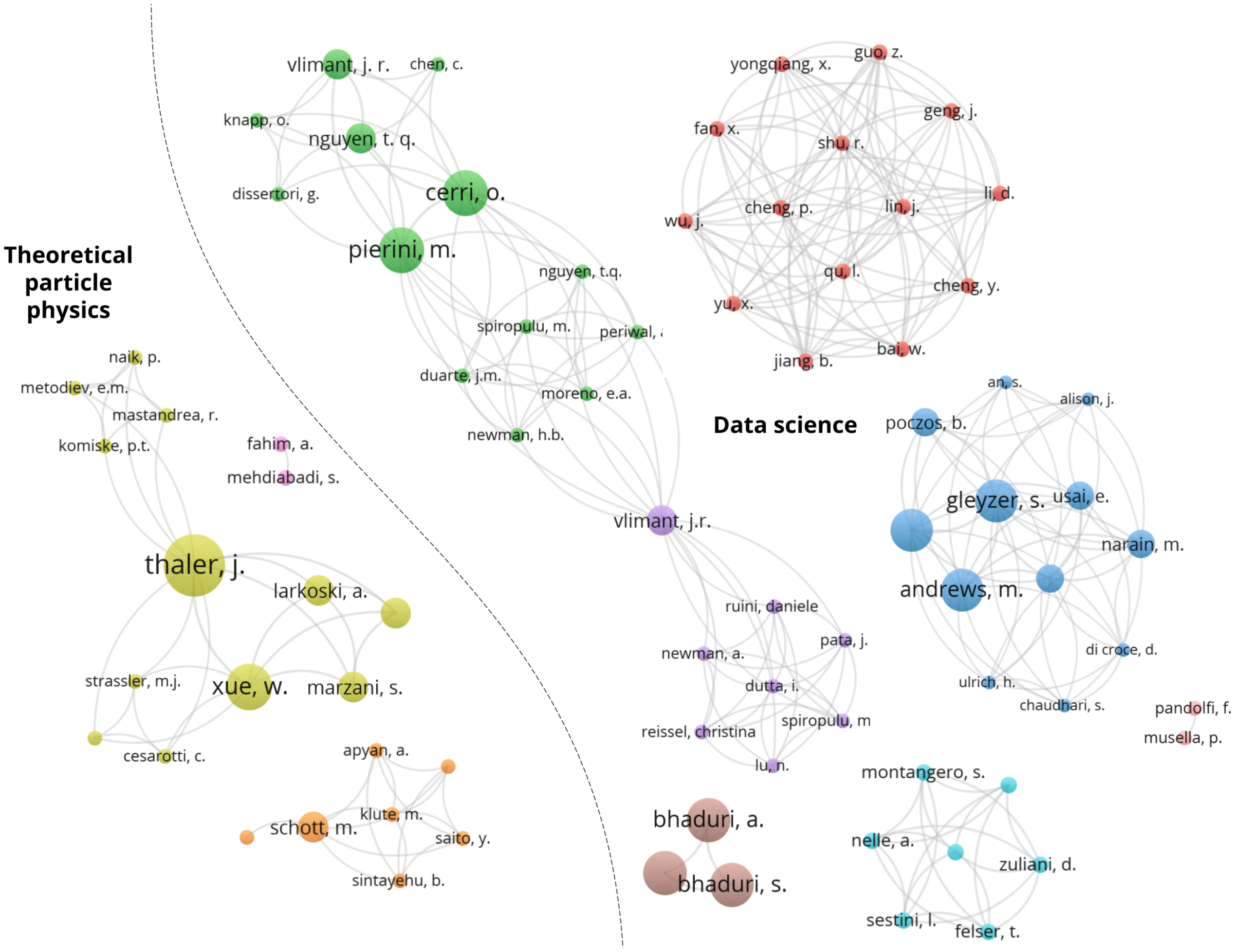


Figure 3: Co-authorship network of open data (re-)use; the authors on the left published on subjects relating to theoretical particle physics; the authors on the right on topics relating to data science (a particularly strong focus was machine learning); created with VOSviewer

interesting insights into the daily working routines of researchers, the timescales of observing major changes in curation and use practices could comprise several months, if not years, and as such succeeds the framework of this thesis. To observe major shifts in curation practices, it is thus necessary to use alternative methodologies.

As this work was conceptualized amidst an ongoing pandemic, it additionally seemed very unlikely that fieldwork would be a viable option in the near future. Therefore, for reasons of practicality, it became necessary to craft a design that enabled remote research options. Finally, this study is interested in the development phase of the portal, which took place in 2014. In order to gain access to past events, alternative analysis tools were necessary.

Following Halkier (2017), who has argued against a “gold standard” thinking towards participant observation in practice theoretical research approaches, I have decided to craft an alternative research design. Rather than understanding ethnographic fieldwork as the most viable way to analyze practices, Halkier (2017) asserts that research approaches should align with the research interest at hand. Halkier (2017) further argues that the development of multi-method approaches can help in crafting stable research designs. Following this suggestion, this work takes a multi-method approach with a focus on semi-structured interviews.

5.3 Semi-structured interviews

In order to gain a better understanding of the roles and intentions put forward by individual researchers in the construction, maintenance, and use of the CERN ODP, I conducted semi-structured interviews. The semi-structured format allows the design of interview questions around a central research interest while simultaneously enabling a dynamic adjustment of the research focus (Jensen & Laurie, 2016). I was able to progressively refine my research interest while, at the same time, remaining in my pre-defined research gap.

Jensen and Laurie (2016) argue that it is essential to consider the “collaborative nature” (p. 173) of interview settings. Researchers should ask themselves how their “interaction [with the interview respondent] produce[s a] trajectory of talk” and thus co-constructs “specific versions of reality” (Rapley, 2007, p.16). Interviewees’ accounts can not be interpreted as a set of detached statements that objectively represent past or present intentions and experiences. Rather, interviews allow the construction of mutually shared understandings within a situated frame.

While interviews do not provide direct access to facts, they nonetheless point to relations outside the interview framework. Bueger (2014) argues that narratives which arise in interview contexts relate to the practices they refer to. Narratives are both informed by the practices they describe and can be used as a way of reinforcing them (Bueger, 2014). Next to asserting the truthfulness or authenticity of a narrative, one can thus investigate which purposes are met by supplying it. I analyzed my interview material both with respect to its truthfulness, as well as its narrative quality. In order to substantiate concrete claims on data practices and collaborative routines, I consulted the open conversations and the source code of the portal on GitHub. Additionally, I viewed the interview material as a set of narratives that promote particular ways of implementing and using open data resources rather than solely describing them.

Further, I focused on the ways in which narratives are informed by common discursive repertoires that are collectively developed within particular communities of practice. The metaphors, phrases, and descriptions that are used to make sense of data-handling techniques are often shaped by the social surroundings of individual researchers. Through eliciting descriptions of data production and use, it was possible to generate insight into the relations between researchers of data production and (re-)use communities.

5.4 Interview questionnaire: Following the data journey

Since the researchers involved in the development, maintenance, and use of the CERN ODP are dispersed around the globe, a large part of the interviews was conducted via Zoom. Overall, I carried out seven interviews in a digital format. Four interview respondents were involved in the development and maintenance of the open data portal, and three respondents were open data users. The interviews lasted between 60 to 90 minutes. Additionally, seven semi-structured interviews were conducted in person with members of the ATLAS team at CERN. These interviews were carried out during a field trip to CERN in August 2022 by the METAFORIS research team. The focus of these interviews was quite general, ranging from open data aspirations of the ATLAS collaboration to the future circular collider project.

In the development of all questionnaires, I followed the general structure outlined by Galletta (2013). Galletta (2013) suggests to divide the interview into three basic segments. In the first segment - "creating space for a narrative grounded in participant experience" (Galletta, 2013, p.46) - the interviewer tries to incentivize detailed accounts of the issues at stake grounded in the participant's narrative. Here, it is important to phrase questions as open-ended as possible in order to allow the interviewee to digress beyond static explanations. The second segment of the interview concerns "questions of greater specificity" (Galletta, 2013, p.49) and builds on the narrative established in the first segment. Here, questions of clarification, specification, and broader context can be posed in order to expand the participants' narrative. Finally, the third segment of the interview is concerned with more theoretically driven questions that allow researchers to situate the interview in the conceptual context of their project (Galletta, 2013).

Generally, the questionnaire for my interviews with open data producers and users was structured along an imaginary data journey. I asked my respondents to sketch out the different steps they undertake when working with open data. The interviews with open data producers were specifically focused on the ways in which datasets are selected from CERN internal servers and how they are accessed, handled, and published on the ODP. Additionally, I covered topics such as the development phase of the open data portal, the general value of open data, and the personal interactions of researchers during data preparation and maintenance work. In the interviews with open data users, I focused my questionnaire on data selection, processing, and analysis strategies. I further interrogated how researchers perceived the quality of CERN open data, its general value, and its impact on the social structure of the HEP community.

The in-person interviews with members of the ATLAS collaboration covered more general issues relating to the potential and challenges connected to LHC open data. Additionally, they focused on LHC internal research validation strategies, such as internal peer review processes and their relation to the research products emerging from CERN compared to open data publications.

5.5 Analysis of the CERN ODP/Conversation analysis on Github/Gitter

Decuypere (2021) has developed a methodology for analyzing data practices. He outlines four overlapping topological dimensions of data practices that can be scrutinized in empirical research: "the interface of a data practice, its actual usage, its concrete design, and its ecological embeddedness" (Decuypere, 2021, p.67).

An investigation of the interface renders visible how elements such as text, data, pictures, software applications, and links materialize on a platform (Decuypere, 2021). The ODP interface displays the basic functionalities of the portal and gives an impression of how data practices of users and developers are structured. However, I did not conduct a systematic analysis of all functionalities of the ODP. Rather, I investigated them at specific points of interest, particularly when they became relevant in the narratives of my interviewees.

Scrutinizing design practices means attending to the development phase of digital applications. By analyzing conversations that were held on Github (CERN open data, 2023b) and Gitter (CERN open data, 2023a) in 2014, I could outline the (inter)actions, challenges and agendas of the portal developers at the time. However, the design dimension not only includes the development phase but additionally refers to the maintenance and care work that needs to be carried out progressively (Decuypere, 2021). For Decuypere (2021), investigating design means scrutinizing "the relational constellations that are generated behind the platform's actual interface" (p.78). In this context, the source code of the portal was a useful resource for my analysis. Together with the conversations on Github, it allowed me to make sense of particular programming choices. Due to the large amount of material available on Github and Gitter, the analysis was only carried out for issues mentioned in the interviews.

Decuypere (2021) further points to the dimension of "ecological embeddedness". Similar to Bowker and Star's (1999) conceptualization of infrastructural embeddedness, this dimension points to the situatedness of data practices in a larger context. In the case of the CERN ODP, this raises the question of how data practices connect the ODP to the LHC computing grid and CERN internal data preservation strategies. However, collaboration internal data practices are not openly documented. Therefore, they were partly covered in the interview part of the analysis.

5.6 Coding and analysis

I chose an inductive approach to coding the interview data. As a result, my empirical analysis deviates in many ways from my original research interest. I began my analysis by developing "in vivo" (Rivas, 2004) codes that were directly present in the empirical material. This turned out to be very helpful in instances where phrasing reflected collectively held knowledges and concepts. In other cases, I chose to work with codes I generated from my theoretical foci. This helped me to relate the material to the overall research question.

I followed a “zigzag approach” (Rivas, 2004) in going back and forth between data analysis and collection, which allowed me to reevaluate my choices of interviewees partners throughout the empirical analysis. For instance, after having conducted several interviews that hinted at the importance of the CERN internal peer review process in the evaluation of open data results, I decided to specifically approach researchers who were involved in the collaboration’s peer review during our on-site visit at CERN.

I chose to code my material with Atlas.ti which enabled me to keep track of and organize all my materials in one program. After a few rounds of coding, I defined overarching code groups. At this stage, I needed to let go of some of my original codes in order to sharpen my analysis and go into greater depth at particular points. In selecting codes and defining code groups, I focused mainly on recurring patterns in the material but additionally attended to striking variations in the descriptions of researchers. The group codes served as a basis for the structure of the upcoming empirical analysis.

5.7 Challenges

Conducting the empirical research for this work did not come without challenges. During the course of this project, I was sometimes confronted with situations that I did not expect, which forced me to readjust my theoretical and methodological focus. In the following, I will outline some of the central challenges I faced during the last 12 months and conclude this chapter with considerations regarding research ethics.

When I started to conduct the semi-structured interviews, I soon realized that I had defined my analytical focus too narrowly. My original proposal focused solely on the enactment of research collaboration as a form of interpersonal exchange between researchers. The first interviews, however, revealed that data practices did not only restructure direct interactions between researchers but caused a reconfiguration of more elusive socio-epistemic structures. At first, it was quite hard for me to move away from my previously defined research framework and develop a vocabulary for describing these more elusive dimensions. The fact that particle physics is an extremely well-researched area in STS helped me to make sense of the socio-epistemic dynamics that became visible in the interviews. I ultimately arrived at the term “forms of togetherness” (Felt, 2009) to describe how various aspects of the research process get re-organized through the use of data.

Another challenge arose during the coding process. I started my analysis by trying to generate codes that were closely related to the interview material. Soon I realized that many of my codes were too narrow to connect them across the material. Additionally, I noticed that the “in vivo” coding approach sometimes inhibited me from developing my own vocabulary in describing the data practices that were addressed in the interviews. After having written several pieces of analysis, I started to realize that they had, at times, taken up the narratives presented by my interlocutors by using the same terms and phrases. I had to take a step back and redefine my codes in relation to the theoretical concepts that I was working with while at the same time trying to remain grounded in the interview material. Choosing the right degree of abstraction in the coding process remained a challenge throughout the analysis.

As I analyzed conversations on Github and Gitter during the development phase of the portal, I recognized that many of the issues were very hard to make sense of due to the technical language and the fragmented character of the discussion (discussions were only partly held online and strongly related to personal meetings that I had no access to). Only together with the interview data could I begin to make sense of particular issues. The same issue applies to the source code. It was largely incomprehensible to me, and I could only very rarely, with the help of the interviews, integrate a short piece of code (e.g., the metadata schema) into my analysis. This methodological challenge relates to a more general observation about open data projects: Open access to resources does not guarantee transparency. I greatly admirable how the open data team aimed to conduct almost all of their activities openly. However, many of the resulting resources are quite incomprehensible to an outsider.

5.8 Ethical considerations

At the beginning of this research project, I did not anticipate any particular ethical challenges. I assumed that my empirical data would not be particularly sensitive since no vulnerable groups are involved in my research project. All my interview respondents were able to give informed consent freely. After conducting my first interviews, however, I noticed that some of the interview material I had gathered might, in fact, contain sensitive information. For instance, a few of my interlocutors voiced their dissatisfaction with specific processes within CERN or the LHC collaborations. Since some of my interviewees are employed by CERN or are members of an LHC collaboration, critical comments toward these institutions might result in negative consequences for their professional life. Further, due to the relatively small size of the High energy physics community, even external users of CERN open data might face negative consequences if criticism towards CERN or an LHC collaboration is voiced. This ethical challenge is heightened by the fact that some of these critiques were highly specific and could possibly be retraced to a particular person, even if their contribution would be pseudonymized in a publication.

While I presume that negative consequences for my interview partners are very unlikely since these criticisms are likely to be construed as productive feedback, it was important for me to present critical comments in the specific contexts within which they were voiced and try not to sensationalize my findings. However, it was hard for me to estimate in how far these critiques could result in negative social sanctions for my interview interlocutors. This is why I decided to rework my informed consent sheet, which subsequently offered the option to consent to the material I was planning to publish. This way, the participants themselves, who have better insights into the social and professional structure of CERN, could decide whether their statements were adequately represented. All participants that have voiced critical comments have chosen this option and were therefore informed about the material that I intended to publish. Further, I have informed interview participants that it is likely that their contributions will be identified, even if the published results are pseudonymized. Therefore, I am confident that all my interlocutors are fully aware of these issues and have full autonomy over the publication of their contributions.

As for a more general ethical consideration, I am very happy that this work is openly accessible since many of my interlocutors, who have invested time and energy in this project, are strong advocates of open science and developed and used the open data portal in order to advance the idea of free and accessible research.

6 Findings

Introduction to empirical findings

The following empirical analysis consists of four main subchapters. The first subchapter is concerned with portal development and maintenance practices that facilitate the upload of data to the CERN open data portal. In this subchapter, I mainly focus on metadata practices. The contextualization of data through metadata proved particularly important in enacting specific forms of togetherness between the members of the open data development team and the data providers in the LHC collaborations. By retracting how metadata practices have transformed since the development phase of the portal, I demonstrate how new enactments of togetherness have come into being through the implementation of a technical infrastructure for the portal.

The second and third subchapters focus on open data (re-)use from the CERN ODP. As Figure 4 suggests, open data (re-)use has resulted in the publication of various works in diverse research fields (See Table 2, Appendix A). While most of these publications are concerned with machine learning techniques[2] (see Figure 4), some focus on the physics of the standard model and beyond standard model phenomena. A few authors are theoretical physicists; others are machine learning experts. While many authors are part of an LHC collaboration (mostly CMS), others are not officially affiliated with CERN.

Due to the length of this thesis, it is only possible to investigate some dynamics of open data (re-)use implicated by this heterogeneity. In subchapters two and three, I thus focus on two particularly interesting dynamics: Research level (re-)use of CMS scientists and research level (re-)use of non-LHC affiliated theoretical physicists. In doing so, I render visible how data (re-)use facilitates the production of new forms of togetherness and apartness between data users and producers.

The fourth subchapter takes a step back and asks how members of the general physics community at CERN evaluate current open data initiatives. By outlining how researchers, who are not involved in open data projects, perceive initiatives such as the ODP, I point to contrasting conceptions of open data that prevail in parts of the physics community at CERN. The analysis will show that the ways in which physicists think about the socio-epistemic constitution of research at CERN strongly influence the ways in which they evaluate the success of open data initiatives.

Before I delve into the first main subchapter of the analysis, I present a short, exemplary

[2] Machine learning has become an “important applied research area in particle physics”(Albertsson et al., 2018, p.1) from the 1990s onward. Current machine learning approaches focus, for instance, on the improvement of the “physics performance of reconstruction and analysis algorithms” or improvements in the “execution time of computationally expensive parts of event simulation, pattern recognition and calibration” (Albertsson et al., 2018, p.4). For a detailed account of the relevance of machine learning in HEP, see Albertsson et al. (2018).

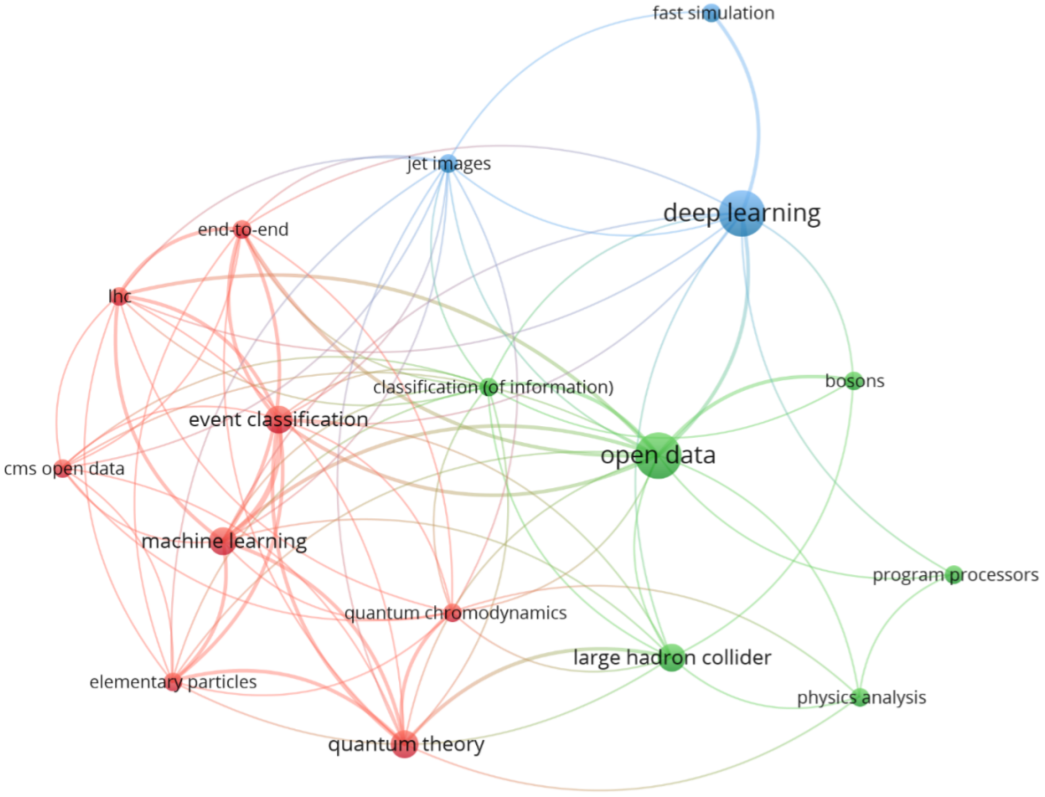


Figure 4: A Keyword co-occurrence network of publications that use CERN open data (See Table 2, Appendix A): Machine learning is a core focus area of these publications; created with VOSviewer

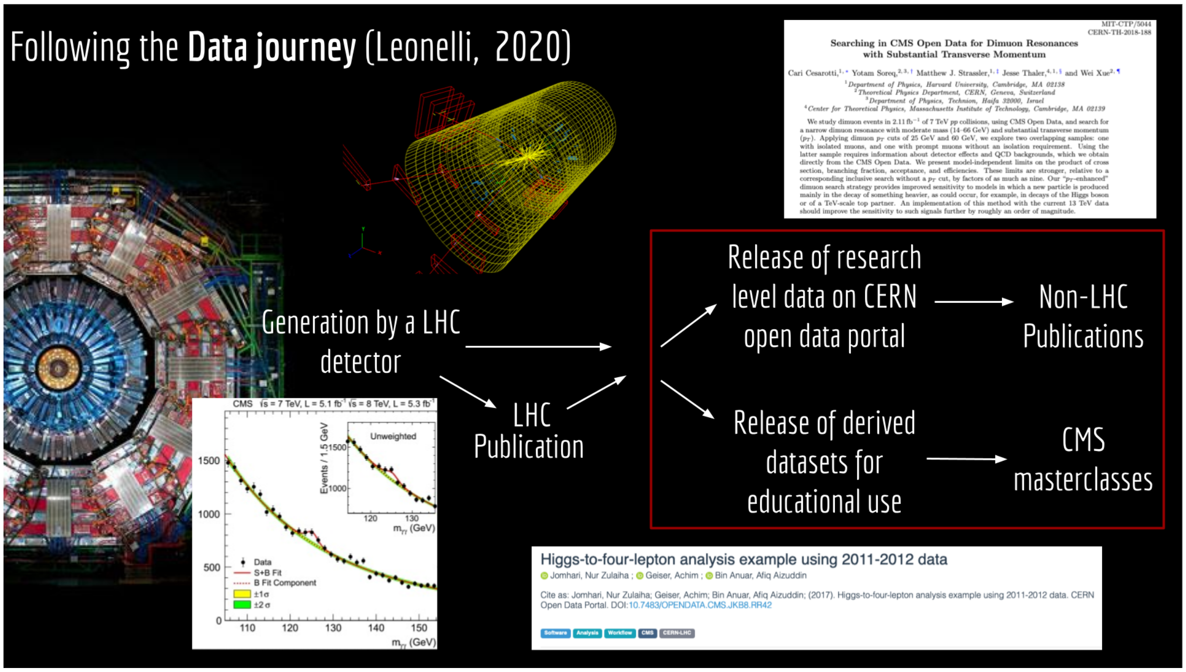
outline of a data journey through CERN.[3] Datasets produced at the LHC travel through various settings before they are made available on the CERN open data portal. In outlining such a journey, I suggest that following the movement of data in contemporary research environments is a fruitful methodology for uncovering socio-epistemic structures of research. As data is becoming the core object of knowledge production not only in particle physics but in many areas of research, the ways in which datasets travel and transform are essential in understanding the kinds of knowledge that can be generated from them and should thus be of central concern to social science research. Datasets create connections between researchers and unfold networks of interconnected practice. By following the trajectories of datasets through the research collaborations at CERN, I could observe which actors emerged at the margins of these journeys and how data was (re-)packaged, (re-)contextualized, and modified along the way. This methodological approach ultimately allowed me to focus on one

[3] As I can only refer to publicly available material in reconstructing this data journey, it will be necessarily incomplete.

particularly interesting part of the data journey; the release and (re-)use of open data on the CERN open data portal.

The data journey: Re-discovering the Higgs

Let us take as an example a dataset that was deployed for one of the most prominent publications in the field: The data used for the experimental discovery of the Higgs boson by the CMS collaboration in 2012 (Chatrchyan et al., 2012). Figure 5 outlines the major steps of this data journey. The Higgs boson dataset was produced by the CMS detector in the Large Hadron Collider (LHC) at CERN. Inside the LHC, a 26.7 kilometer long, underground tunnel, protons are accelerated in opposite directions where they reach velocities close to the speed of light. When these protons collide, new particles are created whose traces are reconstructed by particle detectors that are located at different positions inside the collider. Each detector corresponds with a research collaboration that analyzes the data generated by the detector. There are four major collaborations that use the LHC: ALTAS, CMS, LHCb, and ALICE. Inside the CMS detector, up to "one billion proton-proton interactions" (CMS Collaboration, 2022) take place each second. Not all of these events can be stored by the data infrastructure available to the experiment due to limitations in computing capacity. Therefore, they are automatically pre-selected by so-called "Triggers". Triggers are hardware and software elements



Following the **Data journey** (Leonelli, 2020)

Generation by a LHC detector → LHC Publication

Release of research level data on CERN open data portal → Non-LHC Publications

Release of derived datasets for educational use → CMS masterclasses

MIT-CTP/5044
CERN-TH-2018-188

Searching in CMS Open Data for Dimuon Resonances with Substantial Transverse Momentum

Cari Cesarotti,[1,*] Yotam Soreq,[2,3,†] Matthew J. Strassler,[1,‡] Jesse Thaler,[4,1,§] and Wei Xue[2,¶]

[1] *Department of Physics, Harvard University, Cambridge, MA 02138*
[2] *Theoretical Physics Department, CERN, Geneva, Switzerland*
[3] *Department of Physics, Technion, Haifa 32000, Israel*
[4] *Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139*

We study dimuon events in 2.11 fb^{-1} of 7 TeV pp collisions, using CMS Open Data, and search for a narrow dimuon resonance with moderate mass (14–66 GeV) and substantial transverse momentum (p_T). Applying dimuon p_T cuts of 25 GeV and 60 GeV, we explore two overlapping samples: one with isolated muons, and one with prompt muons without an isolation requirement. Using the latter sample requires information about detector effects and QCD backgrounds, which we obtain directly from the CMS Open Data. We present model-independent limits on the product of cross section, branching fraction, acceptance, and efficiencies. These limits are stronger, relative to a corresponding inclusive search without a p_T cut, by factors of as much as nine. Our "p_T-enhanced" dimuon search strategy provides improved sensitivity to models in which a new particle is produced mainly in the decay of something heavier, as could occur, for example, in decays of the Higgs boson or of a TeV-scale top partner. An implementation of this method with the current 13 TeV data should improve the sensitivity to such signals further by roughly an order of magnitude.

Higgs-to-four-lepton analysis example using 2011-2012 data

Jomhari, Nur Zulaiha ; Geiser, Achim ; Bin Anuar, Afiq Aizuddin

Cite as: Jomhari, Nur Zulaiha; Geiser, Achim; Bin Anuar, Afiq Aizuddin; (2017). Higgs-to-four-lepton analysis example using 2011-2012 data. CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS.JKB8.RR42

Figure 5: Sketch of a CMS Higgs boson dataset journey

that select collisions in accordance with pre-defined criteria.[4] The selected datasets are stored in a decentrally organized computing grid that connects the CERN data center - the "heart of CERN's entire scientific, administrative, and computing infrastructure" (CERN, 2022b) - with the data storage facilities of CMS research institutions around the globe (Lipton, 2020). This computing infrastructure not only grants CMS researchers from around the world "near real-time access to LHC data" (CERN, 2022b) but additionally provides them with the computing power to analyze the data.

In 2012, a publication titled "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC" (Chatrchyan et al., 2012) was released. Together with a paper simultaneously published by the ATLAS collaboration, it marked the experimental discovery of the Higgs boson. Co-published by no less than 2891 authors, the paper featured various graphs and tables generated with data collected by the CMS detector. As is usual for publications in HEP, these tables represent highly processed versions of the data mobilized in the original analysis (See Figure 6). The datasets that CMS scientists work with are highly complex and are generally not supplied alongside publications. Instead, researchers generate simplified graphs and tables to underscore their findings. Thus, rather than providing readers with an opportunity to reconstruct the analysis, these resources are supplied for illustrative purposes. As such, they do not allow fo further analysis of the claims made in the publication and did not elicit data (re-)use outside of the CMS collaboration.

Consequently, the journey of the Higgs boson data (outside of the CMS collaboration) did not continue until 2014, when the CERN open data portal was founded. In the first release campaign, the CMS collaboration made a selection of high-complexity data available on the ODP (Rao et al., 2019). Having educational and research-related uses in mind, the scientists, IT specialists, and bibliometrics who designed the CERN ODP aimed at supplying data at diverging levels of complexity and alongside a broad range of contextualizing resources such as software applications and example codes (Rao et al., 2019). The CMS collaboration has defined four levels of data complexity (CERN, 2020). Level 1 refers to data that is released alongside publications and is usually highly abstract. Level 2 data is defined as a "simplified data format" that is suited for educational and outreach purposes (CERN, 2023). This data is referred to as "derived data". Level 3 data includes "reconstructed collision data and simulated data together with analysis-level experiment-specific software", enabling expert users to perform a research-level analysis (CERN, 2023). Data on this complexity level is generally used for physics analysis inside the collaboration. Data level 4 refers to raw data formats that "allow[] the production of new simulated signals or even a re-reconstruction of collision and simulated data" (CERN, 2023). These data types are sometimes used for the development of machine learning techniques inside the collaborations. In contrast to publications that supply level 1 data, the ODP aims at making level 2 and level 3 data accessible to the public.

[4] Karaca (2020) provides a detailed account of these selection processes.

In the case of the Higgs boson data, a “research-level example” was made available on the ODP. Alongside this example, the open data team released several datasets on complexity level 3 (CMS Collaboration, CERN, 2016a, 2016h, 2016g, 2016f, 2016d, 2016c, 2016b, 2016e, 2017c, 2017d, 2017a, 2017b, 2017l, 2017k, 2017j, 2017h, 2017g, 2017e, 2017f, 2017i). These datasets are ”legacy versions of the original CMS datasets in the AOD format, which slightly differ from the ones used for the original publication” (Jomhari et al., 2017). In addition to these datasets, the example includes a detailed analysis approach. The authors of the analysis example note that it constitutes “a strongly simplified reimplementation of parts of the original CMS Higgs to four lepton analysis” (Chatrchyan et al., 2012). Through the example, open data users can approximate one graph from the original publication (Chatrchyan et al., 2012) (see Figure 6). The example thus only covers one of five Higgs boson decay modes that the CMS collaboration outlined in the original publication.

Additionally, derived versions (complexity level 2) of the Higgs boson data were made available as educational resources on the ODP (Wunsch, 2021a, 2021b, 2021c, 2021d, 2021e, 2021f, 2021g, 2021h, 2021i). Even though they are strongly pre-selected, these datasets allow users to perform simple analysis steps. Alongside the derived dataset, Wunsch (2021a) supplied a detailed description of the datasets and a simple analysis code. After the release of these resources on the open data portal, two publications cited datasets provided alongside the research level example (Cesarotti et al., 2019; Mehdiabadi & Fahim, 2019). Additionally, derived datasets are likely to be used in the context of yearly masterclasses where the CMS collaboration invites high school students to engage in particle physics analysis.

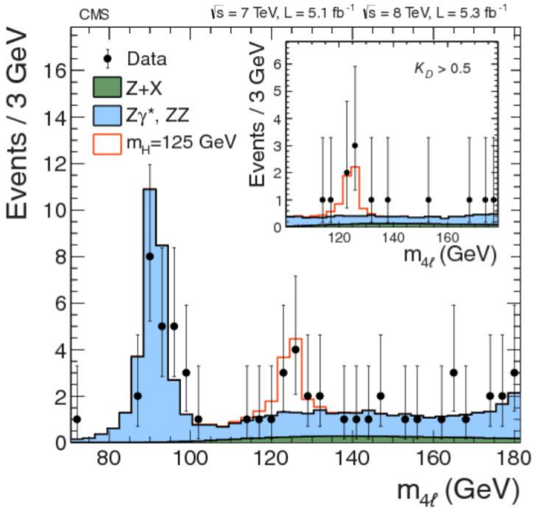


Figure 6: Graph from (Chatrchyan et al., 2012) which can be approximated through the analysis example (Jomhari et al., 2017)

During the journey outlined above, the Higgs boson data acquired different levels of complexity and was repackaged (Leonelli, 2020) in various ways. It was supplied alongside different forms of contextualizing information, for instance, in combination with the original CMS publication or as the basis of an analysis example on the ODP. Various "communities of practice" mobilized the data, such as the CMS research collaboration, the open data development team, and Non-CMS scientists. The datasets were enacted in multiple ways through the different data practices that structured its journey, fundamentally challenging the idea of data as a raw, untouched, never changing representation of reality. Movability, mutability, and usability intertwine in the context of this data journey (Karaca, 2020), since only the transformation of data allowed for its proliferation to diverse locations of research (e.g., Simplification of datasets for educational purposes). In the following, I will zoom in on data practices concerning the preparation and (re-)use of data from the CERN ODP. The selection processes that take place at the CMS detector and the socio-epistemic structures that facilitate data travel inside collaboration have been analyzed in previous works (Galison, 1997; Karaca, 2020; Knorr-Cetina, 1999). (Re-)use of CERN open data, in contrast, has gained little scholarly attention so far. The CERN ODP is highly diverse with respect to the actors involved in its construction, maintenance, and use, as well as in relation to the accessible data and contextualizing material. As such, it is an especially intriguing part of the data journey that offers to disclose how diverse forms of togetherness are enabled across various "communities of practice".

6.1 Portal development and maintenance

As elaborated in 5.1, the ODP was developed as a joint effort of the CERN IT department, the CERN scientific information service, and the CMS open data team. Members of these groups outlined a multiplicity of challenges they needed to address in the early stages of portal construction and maintenance. While the CERN IT department was responsible for developing a technical infrastructure for the portal, the CERN scientific information service focused on ensuring the searchability of resources and assigning DOIs[5] for datasets on the portal. The CMS open data team was particularly concerned with preparing datasets for open release. This task included the creation of derived datasets for education and the contextualization of simulated and collision datasets. However, the efforts and concerns of all three groups often intersected and overlapped.

In the following, I will particularly focus on the practice of metadata application. Metadata (generally understood as data about data) is material supplied alongside research data (Karasti & Baker, 2004). The implementation of metadata on the ODP not only went along with technical challenges, such as the storage of large amounts of metadata but additionally posed difficulties from an information science point of view, e.g., guaranteeing the searchabil-

[5] The CERN open data portal attaches a Digital object identifier (DOI) to each dataset. This allows researchers who work with ODP data to cite it in a standardized way.

ity of resources through metadata. Additionally, metadata application is reliant on physics knowledge about the diverse data types that the CMS collaboration decided to release. Due to the entanglement of these challenges, metadata practices are a striking example of how the different groups worked together in the making of the portal.

6.1.1 Anticipating the open data user through metadata practice

Classification practices are a crucial aspect of any infrastructural development (Bowker & Star, 1999). In the case of the CERN open data portal, datasets needed to be re-classified in relation to the new audiences they were expected to acquire. This re-classification entailed the implementation of a new metadata structure (See Appendix A, Table 1). From the perspective of the CMS open data team, designing such a structure turned out to be a challenging endeavor. The question of "metadata ontology" (CMS open data) became central as the researchers needed to cope with the complexity of the data they intended to release. This meant moving past previous understandings of metadata as only including bibliometric identifiers (author, collaboration, title) and basic information on the characteristics of the particle collision (collision type, energy, year of data taking). Next to this form of "content metadata", the CMS open data team decided to include "context metadata" (CMS open data) into the ontology of metadata. Context metadata and includes resources such as analysis examples or complementary code. According to the CMS open data team, these resources underscore "what can be done with the data" by providing an example of successful use. This demonstration is necessary since external users, in contrast to CMS members, are imagined to be less attached to the data. Since they are neither paid to analyze the data nor enrolled in data production processes, external users require an incentive to engage with ODP resources. Further, users of open data are assumed to lack access to the "knowledge infrastructure" of the CMS collaboration, "which is [...] all the people [that are] around you; supervision, meetings; all the knowledge that is around people in a collaboration" (CMS open data). Consequently, open data users are expected to be more likely to cease working with this data if they are not provided with a quick glimpse at a successful analysis.

The use of context metadata hence stems from anticipations of the open data user in contrast to the CMS member. External users are imagined to lack knowledge as embodied in the people and organizational structures of the CMS collaboration. In order to compensate for this missing socio-epistemic infrastructure, analysis examples and other contextualizing resources are provided on the portal. The missing socio-epistemic context of research was thus reframed as a techno-epistemic problem that can (at least partly) be accounted for through contextualization, such as complementary code. This dynamic shows that portal developers understand the traditional context of CMS data, the CMS collaboration, as crucial in facilitating successful data (re-)use. In designing the portal, this missing context needed to be accounted for through additional metadata elements. The use of context metadata thus underscores how infrastructuring processes involve a creative rearrangement of social, epistemic, and technological orders.

As a member of the CMS open data team pointed out, context metadata such as analysis code should not only exemplify a successful analysis but also convey a realistic understanding of the timeframes necessary for analyzing complex HEP datasets. There is a “huge gap” (CMS open data) between gaining access to data and analyzing it comprehensively. Therefore, context metadata should underscore the complexity of the analysis procedures necessary to make sense of LHC open data. Portal developers have argued that lengthy analysis processes are generally a hurdle for open data (re-)use. Because ODP users are less attached to the datasets than LHC researchers, the often month-long familiarization and analysis processes that HEP data necessitates appear as a significant hurdle to data (re-)use. In order to mitigate this tension, the CMS open data team regularly organizes tutorials and workshops where “people [have] the possibility of working hands-on with this data and trying to get an idea of what are the components that they would need in doing a real physics analysis” (CMS open data). In the context of a research-level analysis, the data processing and analysis times are thus taken as ontological constants that can only marginally be adjusted for open data (re-)use. These lengthy analysis timescales of LHC data conflict with the new context in which resources are deployed. As open data users are considered to be less attached to the data, they are likely to cease working with data that necessitates such long engagement times. Portal developers attempted to mediate this tension by supplying context metadata that foreshadows these lengthy analysis processes. Additionally, they adjusted the socio-epistemic contexts of data (re-)use. In order to immerse previously disconnected external scientists in LHC typical data processing practices, the CMS open data team offered workshops and tutorials with researchers proficient in the analysis of LHC data. Again, the CMS open data team creatively reassembled the social, technological, and epistemic orders of open data (re-)use. In order to mitigate tensions arising in anticipated open data (re-)use, the CMS open data team not only extended the techno-epistemic infrastructure of the portal but additionally initiated personal interactions between LHC scientists and CERN external researchers.

Above all, the implementation of context metadata underscores the anticipatory nature of infrastructuring practices (Nadim, 2016). The value of data for analysis always depends on the anticipated users and their socio-epistemic embedding. The anticipated open data user, understood as unattached to LHC data and without access to the socio-epistemic structure of the collaboration, was crucial in defining the implementation of context metadata. However, these anticipations and their translation into the technological infrastructure of the portal often shape the user groups they attempt to describe. Hence, in investigating (re-)use practices, I attend to the ways in which available data and metadata shape the analysis practices of users. At the same time, anticipations of open data users do not always comprehensively capture actual data (re-)use groups and practices. As I elaborate on in greater depth in subchapter 2, many users are, in fact, CMS members that use ODP resources as a way of publishing technique-oriented papers and, as such, "de-scribe" (Johnson, 1988) to the vision of the unattached open data user.

6.1.2 Enacting provenance

Next to content and context metadata, the CERN ODP provides detailed "provenance information" about simulated and collision-type datasets[6]. Provenance metadata offers information on how data has been generated, both in practical and theoretical terms. In the case of simulated data, provenance metadata includes the software versions and scripts used to generate simulated samples. As such, it "constitutes a full recipe on how the simulated data were generated, thus providing the full history of these data", including "computing environments, the configuration files and the computational procedures used in each data production step" (Simko et al., 2020).

In the case of collision-type data, the portal provides a "data provenance chain" that aims at retracting "how raw data"[7] is "processed into a format appropriate for physics analysis" (Simko et al., 2020, p.3). Consequently, raw data samples were made available on the portal, next to computing environments and scripts. While the CERN open data developers generally do not consider raw data as a useful resource for physics analysis, they argue that the development of machine learning applications could benefit from these types of data. Since collision data is generated from raw data through computational techniques, raw data could be used to test new techniques and allow for an improvement of resulting collision data.

Through metadata practice, open data developers thus conceptualize a particular relation between raw data and collision data. The application of provenance information enacts raw data as a definite outcome of the detector's read-out mechanism and collision data as dependent on the computational techniques used to construct it from raw data samples. If new machine learning techniques for raw data are developed, they hence have the capacity to transform the resulting collision data.

As such, collision data can be contested by the development of new machine-learning techniques.[8] There are hence various understandings of data that exist alongside one another: raw data as a solid, timeless entity and collision data as a more contestable product of analysis. Because raw data is the least processed data version stored by the LHC collaborations, it necessarily becomes a stable point of departure, not only for the purpose of open data but for research endeavors at the LHC more generally.

[6] The first release campaign of the portal did not include simulated data at all. In the following releases, simulated data and corresponding provenance information were released.

[7] Data taken directly from the detector's read-out mechanism.

[8] To my knowledge, there are no raw data (re-)use cases from the CERN ODP so far.

In particle physics practice, it is not possible to (re-)produce particular particle collisions and recreate raw data to test their validity. Each particle collision that is generated and recorded inside the LHC is unique and can never be fully reconstructed. Therefore, raw data constitutes the basis of all knowledge production activities. Conceptualizing raw data as a stable and timeless entity is a way of coping with the techno-epistemic constitution of High energy physics. While other research fields reproduce datasets to test their validity, particle physicists understand raw data as ultimately given. However, this does not mean that researchers think of raw data as being without flaws. On the contrary, the production and use of simulated data demonstrate that HEP physicists strongly concern themselves with what they understand as the background and the detector effects of a particle collision. Hence, rather than reproducing data in order to contest it, particle physics operates on the premise of pre-given raw data that needs to be processed into collision data and compared to simulated data to gain validity.

The application of raw data samples to datasets on the open data portal additionally underscores how portal developers distinguish between various (re-)use groups and their different interests in open data. For an experimental physicist, collision data is usually a stable point of departure from which to conduct a physics analysis. For a machine learning expert, collision data might be the outcome of a contestable computational process that needs scrutinization. As such, the various anticipated data (re-)use groups intersect with the data journey at different points. While machine learning experts might be interested in earlier representations of a dataset, experimental physicists generally work with more processed types of data. Providing provenance and context metadata can be construed as “infrastructurally inverting” (Bowker & Star, 1999; Nadim, 2016) data production procedures and allowing researchers to start their analysis at different points of the data journey.

Thus, challenges arising from the diverging epistemic interest of the anticipated ODP user groups were tackled by adding provenance elements. However, collecting provenance information serves an additional function. It allows researchers to test the reproducibility of the experiment internal data construction processes by retracing the different data production steps themselves. This makes sure that the provenance information, as well as the released data, is correct. In this way, the availability of provenance information on the portal acquires a validating function for both datasets and provenance information.

6.1.3 Mining Metadata

The CMS data preservation team needed to implement scripts for mining metadata information from CMS internal resources. This implementation, however, did not come about without difficulties. Metadata for the portal needed to be gathered from different CMS internal databases, servers, and documents. In this context, it is essential to consider that the storage and production of metadata information within CMS are geared towards local data production practices rather than data preservation. CMS internal storage infrastructures "are developing continually, and what drives the development is not data preservation but the functionality of the current or future data taking" (CMS open data). Therefore, data preservation, and in particular, the gathering of metadata, needed to be integrated into the data-taking workflows of the CMS collaboration. The embeddedness of the ODP into the broader CMS computing infrastructure hence crucially shaped data practices such as gathering and applying metadata. The primacy of data-taking objectives within the collaboration meant that the CMS open data team had to undertake considerable efforts to preserve certain metadata elements.

While internal storage procedures overall follow standardized rules, there were somewhat "volatile" (Simko et al., 2020) aspects to the ways in which resources were internally curated. In this relation, the year of data-taking was a crucial determinant of changing storage practices: Information that had been stored in a specific field in one year, was stored in another field in the next year. Consequently, the researchers needed to gain knowledge on changing storage practices. To guarantee the storage of metadata information in a more standardized form, the CMS data preservation team needed to acquire expertise on CMS internal data production procedures. This acquisition process was as much social as it was a technological endeavor. The open data development team needed to approach CMS researchers involved in the data production and ask them about the metadata elements for specific datasets. They had to "hunt down the [meta]data that was managed 7, 8, 10 years ago" (science communicator). This was challenging at times since "people had left by the time or scripts were stored differently then" (science communicator).

Before the open data development team started to gather information, knowledge on the location of metadata files was stored in the socio-epistemic structure of the collaboration; storage procedures relied on the "interlaced knowledges" (Tuertscher et al., 2011) that were held by members of the collaboration and which were informally distributed upon request. The efforts of the open data development team to extract this information and transfer it into a standardized metadata structure denote an attempt to disentangle the metadata information from the socio-epistemic structure of the collaboration. As Chapter 3 will demonstrate, this disentanglement was contested after the first data release on the ODP. Early open data users reported difficulties in analyzing data due to missing contextualization. Through feedback loops with users, the open data development team could ultimately extend and refine their understanding of what constitutes necessary metadata information.

To circumvent time-intensive metadata acquisition processes in the future, the CMS open data preservation team currently aims at deploying metadata information at the time of data-taking. To avoid a disturbance of the collaboration's workflow, metadata information can be extracted while the data is "hot" and stored for later upload on the portal. In this case, a "documentation archeology" of internal infrastructures is circumvented. A lack of standardization in the archival of metadata information was consequently reframed as a timing issue. By deciding to extract metadata at the time of data taking, as opposed to standardizing metadata storage practices within the collaboration, the open data development team attempted to leave the data practices of the collaboration undisrupted. Here, CMS internal priority settings - the data taking and analysis as opposed to long-term data preservation - become expressed in metadata practice.

6.1.4 Developing a metadata structure

The metadata information collected by the open data development team was ultimately stored in a "JSON metadata schema" (Appendix A, Table 1). Metadata schemas are classificatory systems that define particular standards and rules for implementing metadata. The schema acts as a connecting element between the open data groups within the LHC collaborations and the portal development efforts of the CERN IT department. The CERN IT department, the CERN scientific information service, and the CMS open data team collaborated closely in its design. The teams "would spend 2-3 days in a row, camping somewhere in one of the office rooms and just drawing on boards and discussing data schemas" (science communicator). Since the software development efforts of the open data development team were conducted openly on Github (and open conversations about the maintenance of the portal were held on Gitter), it is possible to retrace some of the conversations that led to the implementation of the final schema. The discussions held on these platforms particularly underscore how different professional standpoints were negotiated in the making of the ODP. In the early phases of the portal development, the "different groups had different needs and a different understanding of each other's needs" (CMS open data). From an information science point of view, improvements in the discoverability, accessibility, and usability (Wilkinson et al., 2016) of resources emerged as central concerns in the discussion. These objectives resulted in calls for the assignment of DOIs as a way of improving the readability of datasets by establishing a possibility for unique identification. In order to guarantee the discoverability of resources, the installation of a search engine was pushed forward by the scientific information service. This went along with calls for the standardization of a set of metadata elements that would allow the generation of search options. At the same time, the scientific information service was concerned with limiting the number of metadata elements that would be applied to the datasets. From their perspective, the application of too much metadata could cause users to get overwhelmed. In this relation, the scientific information service would ask "lots of nasty questions to the CMS people" in order to make sure that the anticipated users would necessarily need the information that the physicists planned on releasing.

From the perspective of the CMS researchers, it was crucial to make sure that metadata scripts were interoperable with CMS internal (and external) structures. Since they were responsible for the extraction of metadata information from CMS internal systems, they needed to ensure that the functionalities of both systems worked well together. Supplying sufficient contextualizing physics information was another central concern for the physicists. In the interviews, as well as the archived material, CMS researchers use affective language to describe data. Since they "gave birth to the datasets", "they are very attached to it and know it by heart" (CERN scientific information service). Relationships of care are in place between CMS researchers and the data (Pinel et al., 2020), and they are reflected in the physicists' practices of contextualizing and explaining data.

From the perspective of the CERN IT department, the long-term preservation of data proved to be a core objective. Software development efforts were often directed toward extending the lifetime of data. For the CERN IT team, these preservation efforts are strongly tied to an enlargement of potential user groups. By making resources accessible to a broader pool of users, feedback circles can be enabled, which might improve preservation techniques. The users and the temporalities of data become entangled in this understanding of data preservation. Metadata practices were thus not only geared towards moving past the socio-epistemic but additionally moving past the temporal structure of the CMS collaboration.

In the making of the metadata schema, the objectives and interests of the different ODP groups intersected, overlapped, and sometimes conflicted. The teams needed to negotiate a multiplicity of interests and establish priorities during the course of their collaboration. Within this process, the different groups learned about the practices of their collaborators and started to appropriate some of their working techniques:

> "When I was browsing the development version of the website, I would come across a bug, [...] a broken link, [...] and I would feed it back to the team. And I would tag one of the developers and say, this link is broken. [...] But over a few months, all of a sudden, [I was] doing a little bit of someone else's role. Because those of us on the collaboration side, and CMS's side, and the library team, we began learning how those tools themselves worked. So instead of tagging someone and saying this link is broken, I was able to fix the link and submit the change to the developer who could then just merge my changes with the main code base. So we began collaborating quite extensively." (Science communicator)

The development of the portal allowed researchers to appropriate each other's working practices across group boundaries. Each team's enrollment in specific tasks was in flux, and their collaboration created a continuously growing pool of shared knowledge. Rather than demarcating fixed boundaries and modes of interaction, the distribution of work between the open data groups largely relied on personal conversations and was in constant movement. Additionally, as outlined above, interactions between members of the open data development team and members of LHC collaborations concerning the storage of metadata elements took place personally. This stands in contrast with collaborative strategies that prevailed after the development phase of the ODP had taken place.

6.1.5 Collaborating through a metadata structure

The later use of the metadata schema shows how a more stable distribution of work (See Appendix A, Table 1) and a digitally mediated form of interaction prevailed after the development phase had taken place. Personal interactions relating to the upload of metadata are largely mediated through the schema. The flexibility of the schema is helpful in "bridging" different practices of metadata production in the collaborations. While "content" metadata is obligatory, all remaining metadata information (such as context and provenance metadata) can be applied flexibly. This high degree of freedom enables the different LHC collaborations to retain specific practices of producing metadata. Provenance information regarding simulated data, for instance, follows very distinct production processes in the different collaborations. Each LHC experiment uses its own computational techniques to generate this data, and it was difficult for the open data team to integrate those various approaches. The flexibility of the metadata schema allows researchers to refrain from aligning their metadata practices completely. For instance, CMS tends to include extensive provenance information for simulated datasets, while ATLAS usually does not offer any specifics on data production procedures.

After the development of the schema, collaboration relating to metadata implementation between the CERN IT department/the CERN scientific information service and members of the open data teams occurred to a great extent through the interface of the schema. For the CMS open data team (and the open data teams of other LHC collaborations), the IT department acted as a "third-party trusted repository". Members of an LHC collaboration would use the schema to apply metadata to their data and then upload it into a "staging area". From there on, the CMS team loses access to the data, and the CERN IT department ingests the data into the ODP. The schema emerges as a boundary object (Star & Griesemer, 1989) in this transaction. By offering the possibility of partial standardization, as well as allowing researchers to partly integrate specific data production practices, it has established a functional working relationship between the implementers of the portal (CERN IT) and the LHC open data teams. This use of the schema has consolidated boundaries between the open data teams of the collaborations and the CERN IT department. It has created a clear distribution of work in relation to metadata practices. Additionally, it allows members of the collaborations to define what counts as relevant metadata themselves; within the pre-confined classification system of the schema; without having to negotiate these meanings outside of the context of the collaboration.

6.1.6 Discussion

The analysis of metadata practices in the context of ODP development and maintenance has revealed a multiplicity of underlying assumptions that structure the journey of data through the portal. The anticipated open data user is generally understood as unattached to the datasets and hence more likely to cease working with them than the LHC scientist. This poses challenges to the application of metadata which acquires the function of contextualizing and legitimizing ODP resources in ways that create an attachment between external researchers and the data. In this relation, the portal developers differentiate between a variety of anticipated user groups, such as theoretical and experimental physicists as well as machine learning experts. These different user groups are assumed to be in need of varying types of metadata. While physicists might work with AOD datafiles and accompanying context and content metadata, machine learning experts are expected to be interested in provenance metadata such as raw data samples.

Data itself is here largely understood as a relational entity that acquires different meanings in different socio-epistemic settings. Data which is useful in the context of machine learning might be worthless to a particle physicist. Some development efforts were hence geared towards adjusting datasets to specific use contexts (e.g., the extension of metadata ontology). Further, the malleability of processed data forms, such as collision data files for physics analysis, was acknowledged by the CMS open data team through the provision of raw data samples that could be used to generate new collision samples that might contest older processed versions of this data. However, the adjustments of datasets to local use contexts were limited by an understanding of data as a partly stable entity with inherent qualities such as long analysis timescales. If such stable qualities stood in conflict with the anticipated use contexts of the data, creative solutions were sought, such as the organization of open data tutorials and workshops.

The second part of the chapter has rendered visible how not only anticipations of data and users have influenced the metadata practices on the portal but additionally the portal's embeddedness into a larger infrastructural setting. The focus of the collaboration on data taking and analysis, rather than preservation, forced the open data team to develop creative ways of bypassing internal practices in gathering metadata elements. In order to understand the functionalities of particular data practices, this finding sensitizes us to interrogate how these practices are valued and hierarchized in the contexts of their larger infrastructural embeddings.

Finally, this chapter has carved out two particular types of togetherness that evolved from the metadata practices on the ODP; one was most prevalent in the development phase of the portal, and another took shape in the maintenance phase of the ODP. The first type of togetherness involved the appropriation of knowledge and expertise between the different open data teams. Group boundaries became porous in this process, and the distribution of work and responsibility was in constant flux.

The second form of togetherness was mediated through the establishment of an agreed-upon boundary object (Bowker & Star, 1999); the metadata schema. This type is characterized by interaction via a partly standardized form, where distributions of work and responsibility are clearly delineated.

The second and third chapters of this thesis will move on to interrogate open data (re-)use dynamics and ask what forms of togetherness arise between open data users and producers.

6.2 Producing Apartness: Publishing with CERN open data as CMS member

Several publications that have mobilized CMS open data were published by researchers who are members of the CMS collaboration (See Table 1, Appendix A). Despite having access to internal resources, researchers from inside the CMS collaboration work with open data since they profit from the conditions of use that accompany the datasets on the open data portal. In order to understand why CMS researchers decide to work with open data, it is ,therefore, necessary to investigate how conditions of data (re)use transform when researchers move from CMS internal data processing structures to the CERN open data portal.

6.2.1 Formalizing informality: The CMS internal peer review system

The use of internal CMS resources is connected to a publication process that includes an extensive internal peer review (Birnholtz, 2008). It is largely shaped by the “communal” (Knorr-Cetina, 1999) structures that prevail within HEP research and promote collective forms of credit allocation. In the following subchapter, I aim to demonstrate how the timescales of this peer review process and the accompanying forms of collective authorship allocation leave the interests of some individual researchers unaccounted for. In order to meet their career goals, some CMS researchers hence resort to the use of open data where the conditions of publication are transformed in important ways.

I begin by outlining the central steps in the CMS internal publication process. In order to gain authorship rights on official CMS publications, researchers need to be members of a CMS-affiliated research institution which has to earn a pre-assigned number of “service credits” (Birnholtz, 2008). In the CMS collaboration, these service credits are allocated on the institutional as well as the individual level. By completing a pre-assigned amount of hardware maintenance and development tasks, research institutes can earn credits that ensure their membership in the collaboration (Birnholtz, 2008). Additionally, individual researchers need to complete a specific amount of maintenance obligations before they gain authorship rights[9].

[9] For a more detailed description of this allocation process, see Kasemann (2021)

In contrast to common authorship practices in other scientific fields, the order of authors on a CMS publication does not signify the researcher's contribution to the analysis. While the first position on the authorship list commonly indicates a leading role in the analysis, in CMS publications, all scientists are listed alphabetically.

Rather than indicating that a particular group of researchers has conducted a specific analysis, authorship in the context of the CMS collaboration consequently signifies a collective hardware maintenance and development effort. For CMS researchers, authorship is centered around the research institution as a unit of organization. The official representation of authorship, the author list accompanying the publication, enacts the CMS collaboration as the central agent. Following Knorr-Cetina (1999), this suggests that the individual researcher is largely replaced as an "epistemic participant" by collective entities. Typical functions of authorship, such as the attribution of credit and responsibility to individual scientists, therefore, remain largely unmet (Birnholtz, 2008). Additionally, this re-interpretation of authorship disables researchers to build a reputation based on a set of publications (Birnholtz, 2008) and consequently results in challenges regarding their personal career advancement.

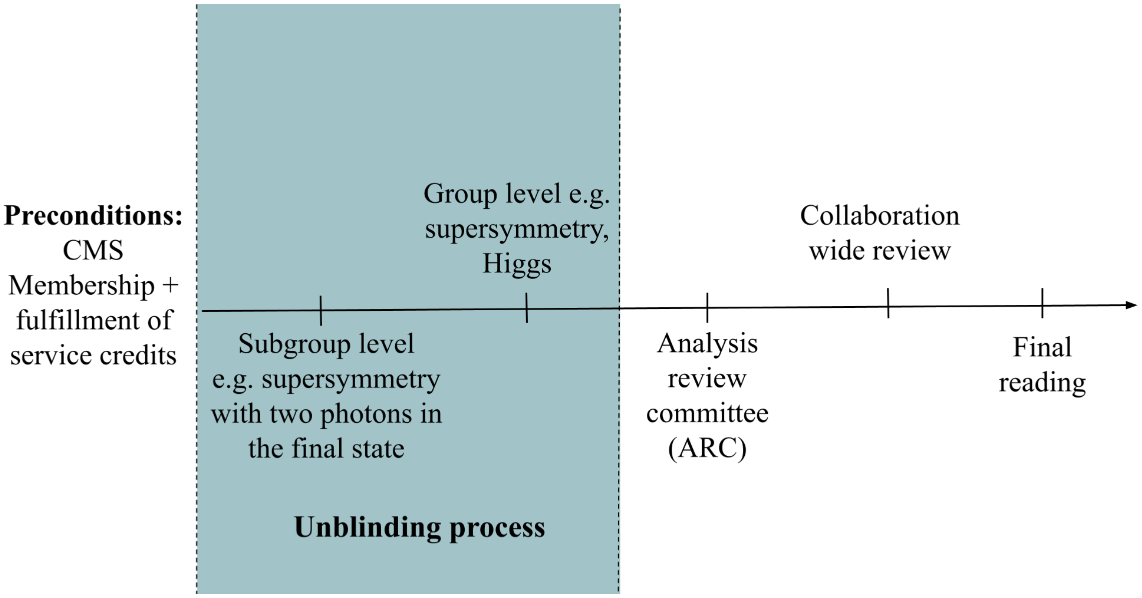

Preconditions:
CMS Membership + fulfillment of service credits
Group level e.g. supersymmetry, Higgs
Collaboration wide review
Subgroup level e.g. supersymmetry with two photons in the final state
Analysis review committee (ARC)
Final reading
Unblinding process

Figure 7: Different steps of the CMS internal peer review process

Next to gaining authorship rights, the completion of service credits enables CMS members to pursue an official CMS analysis. The development of a topic is typically coordinated by working groups that are dedicated to specific physics phenomena such as "supersymmetry", the "Higgs boson" or the "Top Quark" (CMS open data user). After a topic is approved on the group level, researchers can submit their work to the internal peer review process[10] (See

[10] The CMS collaboration describes its publishing process in CMS Experiment (n.d.)

Figure 7). The first layer of review accompanies the early stages of analysis and involves a subgroup of researchers from a specific working group. It is the "immediate layer [of review] on top of the authors" (CMS open data user) and usually concerns highly specific physics processes. For instance, a subgroup review panel in the "supersymmetry" group could be concerned with "supersymmetry with two photons in the final state" (CMS open data user). Usually, the subgroup review is conducted by a small group of scientists who study similar topics in their own work. The subgroup members supply continuous feedback during the development of the paper. Once an analysis has been approved at the subgroup level, it can be submitted for review at the group level. Group-level review again scrutinizes the physics content of a publication draft.

6.2.1.1 Blind Analysis

During these first two levels of peer review, researchers conduct a so-called "blind" analysis. In a blind analysis, the experimenter "hides some aspect of the data or result to prevent experimenter's bias" (Klein & Roodman, 2005, p.9). Depending on the type of research, a blind analysis might disguise a variety of different aspects from the researcher, such as "the signal events", the "number of events" or "a fraction of the entire dataset" (Klein & Roodman, 2005, p.9). The CMS group has described the (un-)blinding process for analyses that measure Higgs boson properties (CMS Experiment, 2022). Here, researchers seek to avoid focusing solely on mass areas where the Higgs boson and, therefore, an excess of events

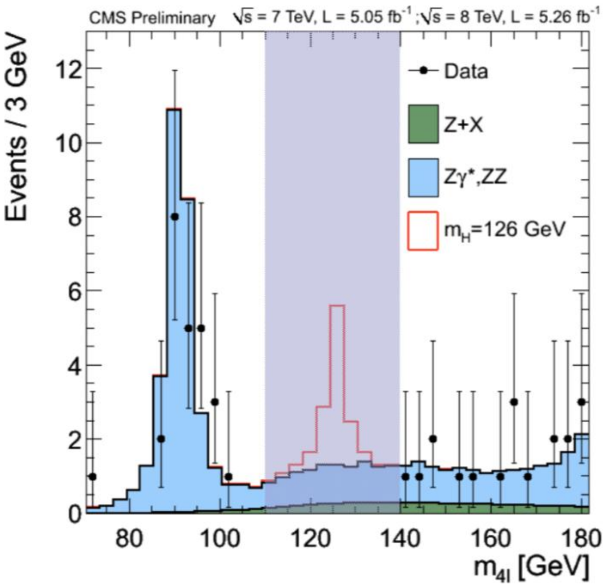


Figure 8: Graph from (Chatrchyan et al., 2012); the red line indicates an excess of events in the shaded area corresponding to the mass of the Higgs boson

is expected (see the gray area in Figure 8). Instead, researchers "draw blinds" over this region. This means that they do not interrogate how changing analysis techniques transform the characteristics of this specific data region. In particular, researchers are prohibited from investigating how the results of their analysis compare to the true number of events (y-axis of Graph in Figure 8) that occur in the shaded mass area of Graph 8. Without being able to determine the effect of their analysis techniques on the data region they are interested in, researchers need to optimize signal selection strategies and background estimation procedures by working with data from the entire mass range visible in Figure 8.

Additionally, the unblinding process involves the definition of a "complementary validation region".

> A critical step of the blinding phase is defining a complementary "validation region" that contains no signal, e.g., a final state including only 1 photon instead of the 2 photons of the signal region. Since the Higgs boson is not expected to decay to 1 photon, this validation region can be "unblinded" to validate the background estimation procedure. (CMS open data user)

Through this complementary region, researchers can validate the uncertainty and accuracy of their background estimations.

Even now, ten years after the Higgs boson discovery, researchers at CMS strive to develop analysis procedures that investigate large mass areas disinterestedly rather than only focusing on the region that corresponds to previous measurements of the Higgs boson. This way, in each analysis, the scientists seek to "re-discover" the Higgs out of the large mass regions they investigate.

By disguising aspects of the data from themselves, researchers hope to avoid that they are "unconsciously work[ing] toward a certain value" (Klein & Roodman, 2005, p.7). This is meant to ensure that they have "removed all biases or at least, [...] have uncertainties to cover any potential bias" (CMS open data user). While techniques similar to blind analysis are quite common in some areas of medicine, for instance, in the development and testing of pharmaceuticals, they are not deployed in other fields of physics and have only become a commonly used method in HEP over the last few decades (Klein & Roodman, 2005). This shift towards blind analysis practices can be considered in the light of an increasingly complex experimental apparatus that generates growing amounts of data. Even though this data is pre-selected by electronic Trigger systems, its size often necessitates further selection in the analysis process (Franklin, 2013). Franklin (2013) argues that this selectivity can lead to the production of results that rely on a specific selection strategy rather than representing an underlying reality. Similarly, the CMS collaboration argues that blind analysis "ensures objectivity when it comes to looking for much-sought-after signs of new physics" (CMS Collaboration, 2022). Changing analysis practices toward blinded procedures hence implicate a shift in the ways in which objectivity is understood by the LHC collaborations.

In their account of the evolution of scientific atlases from the eighteenth to the early twenty-first century, Daston and Galison (2007) outline how objectivity needs to be understood as a historically situated concept that has undergone a variety of discipline-specific transformations over time. Objectivity, as enacted through CMS "blinded analysis" procedures, appears as "blind sight, seeing without inference, interpretation, or intelligence" (Daston & Galison, 2007). In contrast to other understandings of objectivity, such as being "true to nature" or being based on "trained judgment" (Daston & Galison, 2007), this particular enactment decenters objectivity from the skills of individual scientists. Objectivity is performed as exactly that which the individual researcher is absent from.

The group-level convenors usually decide when an analysis has been "sufficiently understood", and the permission to unblind is granted. Hence, while the individual is erased as the conveyor of objectivity, the review structures that prevent researchers from accessing parts of their data or results gain agency. The unblinding procedures consequently do not erase human judgment from the research process. They can much rather be understood as epistemic strategies that enact HEP analysis as a collective, as opposed to an individual process.

Interviewees have additionally linked blind analysis to the notion of discovery:

> "And once you go through all the unblinding stages, now you know: Okay; Did I discover something or not? 99 percent of the time, nobody discovers anything."
> (CMS open data user)

In contrast to other scientific fields, understandings of what constitutes a "discovery" versus understandings of what is meant by an "original" or "innovative" idea are highly disparate in HEP research. Originality, denoting the mobilization of "perspectives or working on a neglected subject, an understudied area, or a noncanonical topic" (Barlösius, 2019, p.918) is a precondition, but no guarantee of a discovery. "Blind analysis" techniques enact a separation between those concepts in practice. The originality or innovativeness of a technique is evaluated at the beginning of a research project when it is approved or disapproved of on the group level. A new discovery, in contrast, reveals itself only after the unblinding process has been completed.

6.2.1.2 ARC review and collaboration wide review

If research has been sufficiently evaluated on the group level, researchers are granted permission to "unblind" and proceed to the Analysis Review Committee (ARC). Within ARC review, the research is scrutinized once more with respect to the physics processes involved. Additionally, questions regarding the language and grammar of a publication draft move to the center of attention. As one interviewee pointed out, the allocation of reviewers to these various levels of revision is not only determined in terms of expertise but additionally includes considerations such as seniority and social status:

> "The subgroup and group conveners, they're usually younger people, maybe post-docs or early professors. So they're [...] typically technically very proficient, but they may not necessarily have the wisdom of experience from doing physics for many years. And that's [...] where the ARC comes in. The ARC will typically be a panel of 4 to 5 people with more senior scientists who have some relevant experience in what it is you're publishing." (CMS open data user)

Seniority acquires a decisive role in the allocation of reviewers to the ARC and determines the ways in which a research project can advance. While earlier career scientists are involved in the heavy editing processes that often accompany research in its beginnings, in the ARC review, senior researchers are able to evaluate research projects. In this way, the social structures and hierarchies of the collaboration are accounted for in the internal peer review. The time intensiveness of the ARC review can be troubling for researchers who submit their analysis. One interviewee has described the ARC review process in relation to a machine learning paper his team intended to publish:

> "And this ARC review, it can go a while. It depends on the panel, it depends on what you're doing. [...] For this last machine learning paper that we had, one of the people in the ARC was the guy who wrote the original code that we were trying to replace. So he was very expert in this. And he initially didn't like our presentation so he requested for a lot more checks to demonstrate that the technique was as robust as we claimed it was. And that took a year, just to get all those checks done took a year." (CMS open data user)

As I will elaborate in greater depth later, this time-intensiveness of the ARC review process often stands in conflict with the interests of individual researchers who aim to publish at a faster pace.

After approval by the ARC committee, publication drafts enter collaboration-wide review. In this process, the drafts are distributed to the whole collaboration, and all members have the possibility to contribute with further comments. Additionally, five to seven research institutes are assigned to independently review the findings. After their approval, a process called "final reading" takes place within which the CMS publication committee will once more evaluate the work before it can be released to the external peer review and publication process.

6.2.1.3 CMS as "epistemic agent"

The outline of CMS internal research validation structures has revealed the particular socio-epistemic and temporal orders that structure internal peer review. Most strikingly, the collectivizing function of this process has gained visibility. Authorship is attained through the collective fulfillment of service credits and attributed to the collaboration as a whole. Further, epistemic strategies such as “unblinding” enact “originality”/”innovative” capacity as distinct from the potential of a “real physics” discovery which is only attainable through collective peer review and detached from the skills of individual researchers.

As outlined above, reviewers for the various levels of internal peer review are selected based on criteria such as fields of expertise or social factors such as seniority and status. Ultimately, a review infrastructure emerges that accounts for the socio-epistemic ordering of the collaboration. The review process consequently endows the socio-epistemic underpinnings of the CMS collaboration with the agency to shape the analyses that emerge from it, and research is enacted as a product of “all the eyes reviewing and studying the results” (CMS open data user). Ultimately, this collectivization of the research product infuses findings with credibility. Because of the diverse layers of review, research emerging from CMS is considered to be “airtight”, and “bad results” are regarded as very rare. For CMS researchers, the internal peer review process is the “scientific process” par excellence: a golden standard for the production of scientific results:

> “Why CMS and CERN have a lot of credibility is because of this review process. It’s not just the fact that they don’t trust you if you’re an individual, but that it’s gone through layers and layers of peer review.” (CMS open data user)

The peer review process enacts multiple forms of togetherness between the CMS collaboration members. The generation of new physics only becomes possible through the collective body of the collaboration, and not through individual researchers. Thus, the peer review process performs togetherness as a form of collectivity (Knorr-Cetina, 1999); the collaboration becomes the central epistemic agent and individual researchers appear as parts of a whole. This collectivity is related to the direct interactions of researchers in the process of peer review, but also to the publication of results where authorship is attributed to all members of the collaboration. Collectivity within CMS is hence both generated through direct collaboration, but also mediated through standardized practices such as authorship allocation.

Importantly collectivity is also produced through the alignment of data practices. During the CMS review process, researchers are asked to adapt their data analysis to the standards of the peer reviewers, who represent the socio-epistemic structure of the collaboration. Hence, an alignment of the epistemic result takes place. Ultimately, all results published by the CMS collaboration take the form of “airtight”, “real physics” analyses.

The collectivity that is produced inside the CMS collaboration thus does not only rely on personal interactions between researchers, but encompasses different nonhuman agents such as standards or material objects, e.g. publications. I want to suggest in this chapter that analyzing togetherness as a form of collectivity that is enacted through the alignment of different practices e.g. data practices that involve diverse arrays of nonhumans, can allow us to understand the socio-epistemic dynamics of research at CERN in greater depth.

6.2.2 The CERN open data portal as alternative publication venue

As outlined above, the CMS internal peer review process enacts a togetherness between members of the CMS collaboration, reconfigures the temporalities of publication and results in the production of a very specific epistemic result ("airtight physics analysis"). For individual researchers, these publication practices can stand in conflict with broader scientific performance evaluations that are largely based on individual, quantitative assessment strategies. A CMS researcher described this tension as follows:

> "There is value, of course, in having a CMS paper. But if you are in an industry where they judge you based on the number of papers you submit rather than the quality of these papers, you need to find another way to get your ideas out." (CMS open data user)

In the particle physics community, researchers can benefit from working on official CMS publications. While the enrollment of individual scientists in a particular analysis is not credited through authorship, "word of mouth" interactions at seminars and workshops are often used to convey individual contributions informally (Birnholtz, 2008). Since "it's a small community", "people know each other" and "have an idea of what people are working on" (CMS open data user). In the context of the particle physics community, conducting an official LHC analysis is generally highly valued and sometimes considered the only relevant determinant in research assessment.

This strict focus on official LHC publications, however, does not prevail outside of the HEP community. In many other evaluative settings (e.g. non-particle physics research) formal, quantitative performance evaluations based on authorship are dominant in research assessment. Researchers hence find themselves caught between two conflicting systems of evaluation. Since the collective and time-intensive publication practices of the CMS collaboration decelerate the production of papers and obscure the involvement of individual researchers in a particular analysis, it seems impossible to perform satisfactorily within both evaluation systems.

Generally, this underscores the situated nature of research assessment practices in HEP. The evaluation of research at CERN is entrenched in social structures (workshops, seminars, ect.) that enable the informal dissemination of information. It is only in these contexts that research contributions become de-collectivized and reassigned to individual contributors. Outside of them, the same evaluation strategies can not prevail. In order to receive credit outside the HEP community, researchers need to find other ways of "get[ting] their ideas out" (CMS open data user). Interestingly, the use of data from the CERN open data portal has become one pathway for researchers that seek to avoid the internal review process and the accompanying collectivization of results. Through working with data available on the open data portal, CMS researchers are able to transform several aspects of their research.

First, it allows the attribution of individual credit through authorship. CERN open data is released on a creative commons waiver, which enables users to process and publish results without assigning authorship to the collaboration. Researchers who work with CERN open data can hence reinterpret authorship as marking individual contribution to an analysis. Secondly, working with CERN open data reconfigures the temporalities of publication. Through circumventing the time-intensive internal peer-review process, research can be published considerably faster. This results in several advantages for the researchers. Next to improving their performance in quantitative research evaluation systems, researchers can communicate their interest in a specific topic to the scientific community at a faster pace. As has been argued by open data users, this can prevent the multiplication of research efforts and allow for a faster repurposing of results.

The new infrastructure of the portal thus reconfigures what researchers value in the data that is made available to them. Rather than its epistemic potential, open data acquires value because of the legal framework it is supplied in. This is illustrated by one researcher's approach to working with CERN open data:

> "So typically, [I] prototyped with the newer data internally, [got] everything figured out and then just pull[ed] the open data, and reproduce[d] everything there, and then publish[ed] that." (CMS open data user)

Rather than investigating new physics themes emerging from open data, the researcher decided to use internally available resources to prototype the analysis. Hence, while open data enabled new possibilities for publication, it was not used for the generation of new epistemic insights.

Waibel et al. (2021) have underscored the relevance of infrastructures in shaping processes of valuation. The mobilization of the CERN open data portal demonstrates how the new infrastructural setting shaped what CMS researchers valued in the data. These shifting valuation practices additionally resulted in a transformation of the research product and its evaluation in different settings.

First, the employment of the ODP enabled researchers to transform the conditions of publication and their performance in quantitative research evaluation systems. However, the publications that result from the use of the CERN open data portal do not solely acquire a new value in quantitative research assessment. Additionally, they gain a new status within the particle physics community.

6.2.3 "Real physics" analysis vs. "proof of concept" publication

Within the particle physics community, many researchers consider work evolving from the use of CERN open data as "proof of concept" publications rather than "real physics" results. A "proof of concept" paper is understood as a work that solely needs to demonstrate that a data analysis technique "works" on a "narrow" subset of data. In contrast, a "real physics" analysis requires a demonstration of the functionality of an analysis method over a large subset of collision data. Further, a "real physics" result needs to account for the "robustness" of an analysis technique. This includes testing an analysis on "noisier parts of the data" (CMS open data user) and data produced in different data taking years. As researchers have argued, the production of a "real physics" analysis is much more time-intensive and benefits from the scrutiny of the CMS internal peer review structure. A "proof of concept" publication, in contrast, would not be able to pass the internal peer review for physics papers[11]. Therefore, the publication of "proof of concept" results is only possible by circumventing internal peer review and using data provided on the open data portal.

For CMS researchers, the publication of "proof of concept" publications via the ODP seems to bridge a gap within the CMS publication economy. While a "real physics" analysis is assumed to profit from "all the eyes" of an internal peer review process, the open data portal allows for the acceleration of publication and the production of a "quick result".

Thus, the infrastructural shift from CMS internal data processing systems to the CERN open data portal not only transformed the temporalities and authorship practices of publishing. Additionally, the infrastructural choice determined the ontology of the resulting publication. Waibel et al. (2021) have emphasized the importance of "ontological rules" in valuation processes. This resonates with valuation practices that result in a new ontology of the open data publication. While "proof of concept" works have been considered useful in indicating the functionality of a technique, they are not assumed to improve the state of physics as such. Therefore, many particle physicists subordinate them to "real physics" results. Further, while "real physics" analyses are considered to be "fully fledged" and very comprehensive, proof of concept publications are taken as much less "rigorous". As such, some particle physicists understand open data publications only as an indication that a more comprehensive CMS internal analysis should pursue a particular topic or technique.

> "If you're really interested in pursuing it further [a particular open data analysis], you would build a full CMS analysis around this technique. And that's sort of what we did. So we started [...] with proof of concept papers first using the open data and [...] once we [...] had a strategy of how to use this in a more rigorous

[11] In most collaborations at CERN, there is also the possibility of publishing "technical papers" that are connected to different, generally less strict and time-intensive, review processes and do not include the whole collaboration in the authorship list. However, what should be treated as a technical publication is often cause for debate inside the collaboration.

> way, then that's when you do the full analysis." (CMS open data user)

In this context, CERN open data emerges as a possibility to test out an area of investigation without dedicating the time and energy that is necessary to pursue a full CMS analysis. Sometimes, researchers might decide that a full CMS physics analysis is not worth the effort. This, nonetheless leaves them with an official open data publication that credits their efforts. In addition, the new ontology of the result has consequences for the valuation practices outside the HEP community. As one interviewee explained, research mobilizing CERN open data is likely to be released in a different journal than an internal publication:

> "So for physics results you would [...] publish this in the more premier physics journals. If it's a technique result, you publish it in a more technique-oriented journal. And the physics journals tend to be more, at least within the physics community, these are more prestigious."(CMS open data user)

Publication in a different journal with a different impact factor can influence individual performance measures. Hence, the transition from real physics analysis to proof of concept publication impacts evaluation within and outside the particle physics community.

6.2.4 Publication hierarchies and transformations of togetherness

Far from challenging the CERN internal peer review process and the "real physics" analysis arising from it, open data results are assigned a very different type of epistemic worth within the particle physics community. By classifying them as "proof of concept" publications, open data works contribute to the stabilization of CMS as the sole producer of "real physics" analysis from CMS data as they enact a boundary between knowledge arising from CMS internal structures and knowledge produced outside of them (Gieryn, 1983). "Real physics" knowledge and in particular the discovery of a new physics phenomenon remain firmly tied to the LHC collaborations and their internal peer review structures. Only the internal review bears the potential to enact an innovative idea as a novel physics discovery. Without this process, research can only become a "proof of concept" publication. This classification of open data results consequently enables ODP users to argue for the expansion of publication practices that do not require CMS internal validation. "As long as [researchers] don't make a claim that [they] discovered anything" (CMS open data user), open data does not intersect with the interests of the collaboration. Even more so, based on the distinction between "proof of concept" and "real physics" analysis, researchers argue that infrastructures like the CERN open data portal are necessary for engendering innovation at a faster pace. Since it "should not be difficult [...] to publish an innovative idea", CMS researchers who "want to publish proof of concept papers using high-quality data, should have an avenue towards this that doesn't involve going through [...] the [whole internal peer review] process" (CMS open data user).

This argument shifts the perspective from the individual researcher back to the collective level. Rather than improving the h-index of a single researcher, CERN is characterized to profit from providing avenues for the publication of “proof of concept” results. In light of the increasing pressure on basic research institutions such as CERN to not only produce basic knowledge in particle physics but to act as innovators, this argument acquires particular weight. The ways in which this debate is taken up and discussed inside the collaboration constitute an interesting avenue for further research.

The production of knowledge in particle physics requires heterogeneous groups of experts (engineers, machine learning experts, experimental particle physicists) to come together at CERN. Demand for open data is particularly noticeable in the area of machine learning. The fact that researchers, who work at the intersection of particle physics and other fields of research, have started to use open data demonstrates how internal publication practices are only partially equipped to account for their interests. The internal publication avenues favor collective authorship practices and, with it, the production of “real physics” results. Many researchers who are part of the CMS collaboration hence first and foremost identify as physicists:

> “At the end of the day, you care about [...] producing physics. [...] If you are strictly a machine-learning researcher, [a proof of concept publication] is fine. But most of us working here; we are all physicists. The goal is to improve the state of physics.” (CMS member)

Research which would be developed as a “real physics” analysis inside the collaboration, becomes a proof of concept work through the use of CERN open data and enables researchers to make contributions to other areas of research. Researchers get the chance to align their data practices with those prevalent in non-particle physics research areas. In machine learning, for instance, “people pump out papers at a much faster pace” (CMS open data user), and open data enables researchers to adapt to this faster publication standard. Hence, through the use of open data, a togetherness emerges between CMS researchers and the scientific disciplines they are contributing to. At the same time, these practices enact differences in the working procedures of CMS researchers themselves. Hence, the CMS collaboration as the central epistemic agent is deconstructed in open data practice.

The use of open data further enables CMS researchers to collaborate with non-CMS members. For instance, one interviewee collaborated with Google on an open data project:

> “We had a small group inside CMS, working on this particular set of machine learning projects and it gave us an opportunity to interact with outside collaborators. Specifically, one of the collaborators we had was Google on one of our last papers. [...] We were able to demonstrate how certain hardware would run on

> their clusters [...]. And that was only possible because it was open data." (CMS open data user)

External collaborators may influence the objectives of an open data project. As such, they can contribute to the dis-alignment of open data research practices and CMS internal analysis procedures. A new togetherness-apartness configuration hence emerges through the possibility of external collaboration where research is re-oriented to fit the objectives of external collaborators rather than following the "gold standard" analysis procedures that prevail inside the collaboration.

6.2.5 Discussion

This chapter has departed from an elaboration on the CERN internal analysis and research validation procedures. I have demonstrated how the CMS internal peer review process enacts publications as collective products of the collaboration. Additionally, the time intensiveness of internal analysis and review processes has gained visibility. I have further outlined how the time-intensive, communitarian peer review practices of the collaboration stand in contrast with some of the interests of individual CMS researchers that stem from the quantitative evaluation systems that prevail outside of the HEP community. Some researchers, who aim for individual authorship and faster publication, have hence resorted to the use of CERN open data, where the rules of publication transform in important ways.

The transition from CMS internal data processing infrastructures to the CERN open data portal has gone along with important transformations in the data practices of the researchers. The collective authorship practices inside the collaboration were replaced with individual authorship practices. Additionally, researchers were able to accelerate the analysis process and publish at a faster pace. Further, the ontology of the research product transformed. Rather than producing "fully fledged", "real physics" analyses, researchers aimed for the production of "proof of concept" publications.

Ultimately, the employment of a new infrastructure for data analysis resulted in a reconfiguration of the togetherness/apartness relations - enacted by the alignment/misalignment of data practices - that prevailed between individual CMS researchers, the CMS collaboration, external fields of research and external open data collaborators. Apartness emerged through the misalignment of data practices between the CMS collaboration and individual open data users, which consequently triggered a transformation of the product of analysis. In contrast, togetherness arose between CMS researchers who worked with open data and external fields of research, in particular external open data collaborators. Chapter 3 will move on to investigate a second dynamic of open data (re-)use and its implications for the togetherness/apartness relations in HEP research.

6.3 Data (re-)use by theoretical physicists

In order to make sense of open data (re-)use by theoretical physicists, it is helpful to briefly recap the relationship between experiment, theory, and computation in particle physics.

6.3.1 Situating open data (re-)use by theorists historically

The relationship between experiment and theory has gained much scholarly attention in the past (Galison, 1997; Knorr-Cetina, 1999; Traweek, 1988). As Sharon Traweek (1988) points out, it is an essential ordering instrument in particle physics practice:

> "Besides rank, at least four other crucial distinctions are maintained in the particle physics community. One of the most fundamental is that between experimentalists and theorists." (Traweek, 1988, p. 111)

Up to the 1980s, theorists were generally seen as producers of theoretical concepts, while experimentalists tested these theories empirically (Galison, 1997). From the 1980s on, the relationship between experiment and theory started to transform (Galison, 1997). The rise of Quantum chromodynamics (QCD), which "was a notoriously difficult theory out of which to extract phenomenological predictions" (Galison, 1997, p.43), confronted experimentalists and theorists with new questions. Experimentalists needed to decide which models they would compare to detector data and ask whether they "were [...] testing the model or QCD itself?" (Galison, 1997, p.43). Thus, they started to employ their own theorists who had enough knowledge of the experimental setup to compare theoretical models to the data generated by the experiment. In the realm of theory, the profession of the "phenomenologist" started to emerge, "whose work was designed to generate experimentally testable consequences of QCD" (Galison, 1997, p.43). Through this diversification of professions, the relations between theory and experiment multiplied. Rather than retaining one constant connection between one another, different subgroups of the experimentalists and theorists started to work out situated areas of interaction (Galison, 1997). One such area was that of computation.

Experimentalists started using computational techniques to cope with the increasing size and complexity of detector technologies. Due to this complexification, it became more and more challenging to differentiate "background noise" from interesting signals in experimental data. By using so-called "Monte Carlo generations", physicists could simulate large numbers of particle interactions as they interact with the particle detector. These simulations could be compared to the detector data to delineate signals from detector effects. As Galison (1997) argues, "without the computer-based simulation, the material culture of late-twentieth-century microphysics is not merely inconvenienced— it does not exist" (p.689). Till today, Monte Carlo is an integral part of experimental practice.

For different reasons, simulation became a part of theoretical particle physics. As particle physics operates on the premise of an inherently probabilistic subatomic world, interactions of single particles are understood to be guided by coincidence and, as such, never entirely predictable by a physics theory. Only large numbers of particle collisions are assumed to describe stable probability distributions that theoretical models can predict. Through Monte Carlo simulation techniques, theorists can model their theoretical assumptions over a large number of (quasi-) random particle collisions. Consequently, these simulations allow physicists to retrieve stable probability distributions for a particular theoretical input.

For Galison (1997), the introduction of computer-based simulation into particle physics practice marked the creation of a local "trading zone" between experiment and theory. As researchers on the theoretical and experimental side of particle physics started to work with simulation, they developed a common language for discussing computational issues regarding Monte Carlo, such as random number generation. At the same time, theory and experiment retained distinct dynamics. For theorists without connections to experimental data, simulation is a way of accounting for the probabilistic nature of the subatomic world. For experimentalists, an essential part of simulation is the delineation of detector effects from particle collisions. Till today, simulations in experimental and theoretical practice assume very different forms. Simulated data produced by theorists can be construed as a "simple generation of the collision" yielding "a list of particles that in principle go to (the) detector" (ATLAS scientist). These simulations do not consider interactions between the particles and the detector. In contrast, simulated data produced by the LHC collaborations reconstructs the different steps in which particles move through and interact with the detector. In the following, I want to investigate how open data (re-)use by theoretical physicists transforms their simulated data practice. I will describe how data (re-)use by theorists enacts a new form of togetherness that exceeds shared terminologies regarding simulated data. It is a togetherness that arises through the common use of open data and has the capacity to partly align the epistemic practices of experimentalists and theorists.

6.3.2 Data friction

The first data release on the ODP in 2014 did not include simulated data, which resulted in considerable challenges for theoretical physicists who decided to analyze one of the datasets provided by the CMS collaboration. One of the theorists that worked on this analysis pointed out that the CMS detector has the potential to "smear out the measurements" or behave "more sensitive towards a particular particle than the other" (theorist working with open data). Next to describing collision processes, data generated by the CMS detector thus describes the interactions of particles with the detector. Consequently, comparing theoretical models; generated through simulations that do not include detector effects; to collision data taken by the CMS detector "is not an apple-to-apple comparison" (Theorist working with open data). Without access to “detector information, either in the form of CMS-approved fast simulation software or simulated Monte Carlo datasets” (Tripathee et al., 2017, p.22), the researchers could ultimately not determine if specific correlations between simulation and collision data were “robust or merely accidental” (Tripathee et al., 2017, p.22). The lack of detector information gave rise to a particular kind of “data friction” (Edwards et al., 2011). However, this friction did not interrupt the data journey but shaped its epistemic outcome, the theorists' publication (Tripathee et al., 2017). While the researchers were able to find ways in which they could partly test their theoretical models, they remained very cautious by claiming that some of their findings might be merely accidental. In this way, the lack of simulated datasets contributed to the establishment of a hierarchy between the results published in the theoretical paper, produced through the use of CERN open data, and papers that are published by the CMS collaboration itself, where an agreement between simulation and collision data is more confidently interpreted as a “robust” finding. The theorists missed information that goes to the core of the experimentalists' expertise: information about the detector's functionality. The collision data the theorists used appeared detached from the detector. A lack of simulated data had erased the intertwining of the detector and the data. It enacted the collision data as a universal entity rather than a local phenomenon tied to its production context. However, the data only partially managed to detach itself from the detector since the resulting findings were considered potentially accidental because of the missing detector information. Hence, the collision data was again re-localized by theoretical physicists.

In response to the theorists' recommendation, the open data development team included simulated data in the following release campaign on the ODP. As I will demonstrate, the availability of these data samples allowed the alignment of simulated data practices between the CMS collaboration and the theoretical groups that worked with this data. As such, it enabled theorists to “relocate” the generation and testing of their theoretical models inside a simlulated version of the CMS detector.

6.3.3 Epistemic data culture in experimental particle physics

In 2019, a group of theoretical physicists published a paper on beyond-standard model physics using CERN open data (Cesarotti et al., 2019). The publication used collision and simulated data provided by CMS to search for new physics phenomena. One of the publication's authors explained that the research group had previously been trying to convince LHC researchers to pursue a specific theoretical idea in their data analyses. Since none of the researchers at CERN showed any interest in the idea, the team ultimately decided to test the hypothesis themselves on data supplied by the ODP. As one physicist involved in the publication argued, this research project was only possible through the availability of Monte Carlo samples. The simulated datasets allowed researchers to conduct searches in very high precision areas. By using open data, the theorists ultimately wanted to convince the experimentalists to pursue their theoretical idea inside the collaboration by using “fresh” LHC data. In order to persuade them, they needed to demonstrate the idea’s functionality on data generated by the collaboration. In this context, the ODP can be seen as a site of struggle that allows actors to push forward a particular research interest within experimental particle physics. Through the use of CERN open data, the theorists aimed to build a case for their research in the hope that the experimentalists would ultimately acknowledge its worth and pursue it themselves more extensively. While an LHC collaboration did not officially take up this particular publication, the CMS collaboration adopted at least one theoretical open data publication. This dynamic underscores the portal’s capacity to direct research objectives inside the collaboration (CMS Collaboration, 2017h; Larkoski et al., 2017). Therefore, research infrastructures such as the ODP can be seen as sites where the power to define what counts in HEP research can be re-configured in important ways.

However, trying to convince the experimentalists of their theoretical idea came at a considerable cost for the theorists. The research group encountered several restrictions when trying to pursue their research goal. In a specific region of the data, the researchers could not simulate their model and test it against collision data because no simulated samples were available. This missing simulation ultimately disabled them from successfully investigating this area in the data. As a member of the CMS data preservation team outlined, the production of simulated samples does not follow a pre-defined logic. Instead, it is organized around the research interests within the collaboration:

> “Now we have something we can call a well-defined set of corresponding simulated data. But it’s never really the full story. [...]. And each publication is the work of a certain group of people. And for this publication, there’s a certain amount of simulated data needed. And those may be needed for the other analysis as well. But it’s always who requests the simula[ted data], it’s not someone high up in the hierarchy who says that these are the data samples that we simulate.” (CMS open data)

The group structure of the CMS collaboration and the research interests that prevail within it are hence crucially influencing which simulated data is generated and later released on the portal. This narrows down the ways in which researchers can use the data outside of the collaboration and inscribes local research interests into the selection of available simulations. Ultimately, a lack of simulated data nudged the theorists’ research into a specific direction by preventing them from exploring certain areas of the data.

For the theorists, engaging with the simulated data additionally proved challenging on another level. The timescales of data analysis were significantly enlarged through the use of detector simulations. While the development and testing of the new theoretical model took only a few months, the physicists needed to invest almost a year in “understand(ing) how the data is being collected, what the data actually means” and “what the probability is that it means what we think it means” (Theorist working with open data). This “painfully” long and complex familiarization process does not seem surprising if one considers the vast amount of information inscribed into simulated datasets. Simulations enact an imaginary journey of particles through the detector and retrace all interactions of the particles with their material surroundings.

> “Our simulated data goes one step further and really goes through all the steps of what happens to these particles when they go through the detector, and what kind of signals they generate and how do they get reconstructed afterwards with the algorithms.” (CMS open data)

Without knowledge of the LHC and the CMS detector’s inner workings, it seems almost impossible to grasp the meaning of these datasets. In order to make sense of them, the theorists had to understand intricate details about the particle’s interaction with the CMS detector. This forced them to consider the technological and computational techniques that shaped the particle journey through the LHC. In this process, the theorists became aware of how peripheral technical parts of the detector come to matter in the particle interactions described through simulated datasets:

> “So, if you have a metal structure that holds up your detector, your particle beam can bounce off the metal structure. It’s like no one expects scaffolding to be an important part of particle physics, but it can be.” (Theorist working with open data)

Thus, the theorists gained insights into a multiplicity of relevant parts of the experimental setup. They realized that they had to adapt their analysis to the local conditions of the detector. The data they worked with could only marginally be adjusted to fit their local use environment. Thus, the theorists needed to extensively study the data and adapt their analysis. Consequently, large parts of their analysis practices started to align with those of the experimentalists facing the same conditions defined by the detector’s capabilities. This shows how CMS open data can act as a powerful agent in transforming data analysis practices.

Simulated data produced by the CMS collaboration signifies a struggle to delocalize data taken by the detector. In order to perform a high-precision analysis of particle collisions that occur inside the LHC, it is necessary to inscribe detailed knowledge of the detector into simulated data. The context and content of a particle collision are intertwined on the level of research data, and this is reflected in the inextricability of simulated datasets and collision datasets. The collision data produced by the detector can not be considered epistemically valuable in itself since detector effects, the influence of reconstruction algorithms, and particle interactions are entangled in them. Simulated data is necessary to delineate these phenomena by fishing out the background noise and enabling an “apple to apple” comparison. Only by comparing simulated data to collision data are researchers able to formulate a scientific claim that can sustain itself outside of the local conditions of the detector.

Additionally, theorists used simulated data to understand the limitations of the collision data. They “needed Monte Carlo to take theory and then show what [they] can’t see beyond [the] detector” (Theorist working with open data). Thus, simulation not only allowed theorists to formulate knowledge claims but also enabled them to generate non-knowledge claims.

The high degree of contextualization necessary to successfully deploy collision data in HEP speaks to a specific experimentalist culture strongly oriented around the functionalities of the detector (Knorr-Cetina, 1999). Data can not be easily ripped out of its context, and several steps are necessary to produce knowledge not tied to a local detector. In contrast to scientific practice in other fields, data needs to be complemented by extensive simulated datasets before a comparison between theory and collision becomes possible. This specificity of HEP research then crucially shapes the communities that can successfully deploy experimental data. Data users need to familiarize themselves with local data production processes and understand the detector functionalities and the data reconstruction algorithms necessary to produce collision-type datasets. This is not an easy undertaking, as the struggles encountered by the theorists who worked with CERN open data demonstrate. However, the CERN ODP is an example where the disjointed data practices of theoretical and experimental HEP could be overcome (even if only in a few instances).

6.3.4 Producing data togetherness

In 2.3, I have conceptualized the CERN open data portal as a new digital place of research. I have argued that digital places like the ODP have become essential places for experimentation and theorization in particle physics practice. In the same way that places such as the collider and the detector gather and coordinate heterogeneous groups of researchers, digital places follow distinct dynamics and draw together new assemblages of actors. However, real-world places like the detector and digital places like the ODP are strongly intertwined. The simulated data on the CERN open data portal enacts a virtual version of the detector. As such, it transports large parts of the experimental setup into the place of the portal. This “inscription” (Johnson, 1988) of detector effects into simulated data enables a proliferation of analysis. Researchers don’t need to be in physical proximity to the detector or part of an LHC collaboration to gain access to this information. Additionally, the availability and use of simulated data underscore how the material agency of the detector extends into digital places and forces data users to adapt their practices to the restrictions defined by the detector. Several theoretical publications using CMS open data demonstrate that theorists attempted to make sense of the local data production contexts through the use of CMS-simulated data. As I have outlined above, this involved the theorists' investigation of detector effects and particle reconstruction processes within particle collisions. Adopting the experimentalists' data practices thus entailed engaging in technical questions of detector engineering and design. This created a new form of social-epistemic togetherness between the theorists and LHC researchers. Through using simulated data, theorists could retrace and make sense of the processes that take place inside the detector and hence engage in local production practices. In order to construct a delocalized scientific claim, they needed to test their theoretical concepts inside a simulated version of the CMS detector. Simulated data produced by the experimentalists thus emerges as a crucial link between those different communities of practice.

> "Theorists can use this data to test their theories and things like that. And if they can build a proof concept of why their theory is something experimentally worth exploring [...], experimentalists can pick up on that and [...] do the tests, or do more advanced tests of those theories of those models. And so, I think using this data brings the community together a little more and increases and encourages more collaboration between these different aspects of research. Theorists and experiment or people who do simulations because this acts as a binding medium for all of them to come together and put science forward together." (Theorist working with open data)

By shaping the data analysis practices of the theorists and by bringing together experimentalists as data producers and theorists as data users on the ODP, this togetherness is social and epistemic. It is a form of togetherness that re-aligns the simulated data practices of theorists and experimentalists who have become increasingly distinct. As I have outlined in this chapter's introduction, detectors have grown in size and complexity over several decades, and simulated data produced by the experiments has become similarly complex. As such, it has become increasingly distinct from the theorists' simulated data, which does not consider detector effects. The CERN open data portal unifies those disjointed practices and allows theorists to test their ideas in consideration of the limitations of the detector.

Additionally, experimental works that have picked up on theoretical open data publication suggest that open data has the potential to reshape the exchange between theory and experiment on the level of publication. However, due to the small number of theoretical publications using open data, this dynamic is not representative and could constitute an area for future investigation.

6.3.5 Discussion

This subchapter has outlined how the use of open data by non-LHC-affiliated theoretical physicists resulted in the production of a particular form of togetherness between LHC researchers and external theorists. The beginning of this subchapter underscores how friction in data (re-)use arose for theorists due to the lack of simulated data on the first ODP release. The data available on the portal lacked information on the functionalities of the detector, which the theorists considered crucial in conducting high-precision searches. A misalignment of data practices resulting from a lack of knowledge about the functionalities of the detector led the theorists to question the significance of their research results.

When the open data team supplied simulated data on the portal in the following release campaigns, theorists gained access to the experimentalists' knowledge of the functionalities of the detector. A togetherness, arising through the alignment of simulated data practices, emerged between the theorists and experimentalists. This finding strongly resonates with one of the major implications of this thesis. By using each other's data, previously disconnected areas of research can align their data practices and adjust the epistemic result of their research with respect to one another.

The alignment of simulated data practices additionally facilitated the exchange of research objectives between theory and experiment. The research of theorists who conducted a CERN open data analysis resulted in the investigation of particular phenomena within the LHC research groups. This demonstrates how the alignment of data practices had implications for the epistemic foci inside the collaboration. The use of CERN open data hence gave theoretical researchers a possibility to define what counts in experimental particle physics research and, in doing so, re-configure the power relations that prevail between theorists and experimentalists.

6.4 Open data at CERN from the perspective of ATLAS scientists

In this subchapter, I will outline how CERN scientists who are not involved in the development of the CERN open data portal perceive current open data approaches. This allows me to engage in narrations that take a more critical stance toward open data projects. In order to make sense of these narratives, it is helpful to briefly consider how current funding frameworks affect open data implementation in research institutions such as CERN.

6.4.1 Context of open data developments at CERN

Ideas of open and accessible research stand at the core of CERN's scientific endeavor. The CERN convention states that “experimental and theoretical work shall be published or otherwise made generally available.” (*Convention for the Establishment of a European Organization for Nuclear Research | CERN Council*, 1953). Over the last decades, CERN has been presenting itself as a 'laboratory for the world' that advances science through the open exchange of ideas (Mobach & Felt, 2022). Despite these longstanding concessions, what openness means in practice is an issue of ongoing contestation. The datafication of CERN's research activities during the last three decades has raised the question of how data should be made openly accessible. Considerations of opening up data are particularly driven by funding agencies that increasingly demand open data strategies from experiments. National and supranational funding structures have started to put forward guidelines for the open release of research data (Wessels et al., 2017). By calling for the development of open data strategies, funding agencies tap into a discourse around open data that has taken up speed during the last decade. While this discourse generally acknowledges the difficulties that go along with the open release of data, scholars in the field of data studies have argued that data is still often understood as a distinct entity that can seamlessly travel to various socio-epistemic settings (Leonelli et al., 2017).

The idea of data as a stable and transportable entity relies on some important presumptions. First, it is based on understanding data as an objective representation of the research objective. Thus, it assumes that data can easily be separated from the site of its production and moved to diverse research locations. This assumption is contested by some CERN scientists, as the following analysis will show. Secondly, the idea of data as easily movable to different sites implicitly assumes data (re-)use by individuals or small research groups rather than big collectives. As I will outline in the following, this conflicts with CERN internal understandings of data analysis that are tied to the collective review of the result.

In the following, I will draw on material that the METAFORIS team gathered during an on-site visit at CERN in August 2022. The scientists we interviewed have engaged critically in CERN's recent open data developments. They argued that current open data policies often fail to enable actual data (re-)use:

> "We know that it is an obligation to have an Open [...] Data Policy [...]. And that means we can refer to it now, we have one. They're part of our statutes and so forth. But basically none of us see the benefit of Open Data. The need is purely given through funding. And we don't get anything out of it and the public doesn't get anything out of it."

In the following, I will outline why scientists are skeptical about current open data developments and ask how they envision potential avenues for successful open data (re-)use in the future.

6.4.2 Boundary narratives

CERN physicists have voiced several concerns in relation to the open release of their data. In particular, they have argued that open data could be "misused", leading to the publication of results that the CERN collaborations might later have to refute. Researchers "have to understand the detector extremely well" to generate a meaningful analysis and otherwise risk producing faulty results. This observation aligns with the perception of open data users who have outlined the long and often tiresome familiarization processes they had to undergo when using open data. Scientists fear that the integrity of particle physics will be called into question by erroneous open data publications. Importantly, interview respondents argued that the danger of producing a false analysis stems from the external user's lack of access to informal knowledge that circulates in the collaboration. This knowledge is passed on from "generation to generation" and lives "in the minds of the people [and in] their experience". It is thus understood as either impossible or extremely difficult to formalize, where in the latter case, formalization would exceed the capacities of the collaborations.

Additionally, the interviewees have underscored the relevance of the CERN internal peer review process. In order to make sure that analyses are correct and coherent in terms of content and form, they undergo the extensive review procedures outlined in subchapter 2. Interestingly, interview respondents have argued that these peer review structures are not only necessary to generate correct physics analyses but additionally ways of accounting for the social structure of the collaboration:

> "Because the 3000 physicists who are on the paper afterwards are all allowed to comment. [...] Everybody who is on the paper has the right to ask questions [about it] because [...] their name is then on there. And that's why it's very complex."

Thus, authorship practices become the driver of peer review processes which in turn shape the epistemic outcomes of research, the publications. Interviewees have additionally conceptualized authorship practices as crucial in stabilizing the distribution of tasks inside the collaboration. As interview respondents have argued, the fact that authors are listed alphabetically ensures that detector maintenance, as well as data reconstruction and curation tasks are undertaken by researchers. If the listing of authors in a publication revealed which types of work they conducted, necessary maintenance work would be neglected. As such, authorship practices acknowledge the importance of all the diverse kinds of work that enable the generation of physics results in the collaboration.

Generally, the argument presented above constructs a boundary between the knowledge produced by the collaboration and knowledge resulting from the external use of the collaboration's resources. This resonates with observations made in chapter two of the empirical part, where knowledge resulting from open data (re-)use is denied the potential of constituting a physics discovery. The different types of knowledge that come into being through this argumentation are tied to the social structure of HEP research. Gieryn's (1983) account of boundary work has shown how demarcations between science and nonscience are drawn through diverging lines of argumentation that depend on the specific context of comparison. While demarcations of science from religion, for example, followed an argumentation that foregrounded the practical usefulness of scientific applications, demarcations of science from mechanics stressed the inherent supremacy of scientific thought (Gieryn, 1983). In the context of CERN open data (re-)use debates, different sites of particle physics practice are demarcated from one another. Interestingly, in the narratives of CERN physicists, the distinct feature of the CERN collaborations is one which is regularly used to characterize non-scientific domains: its dependency on social structures and tacit assumptions. As such, the physicists' arguments acknowledge the situatedness of research emerging from CERN, specifically the situatedness of data. In doing so, they underscore the co-production of social and epistemic structures.

In these arguments, the "correct" scientific result is generally framed as the driver of the social structures that seek to make these results possible. However, in the case of peer review, collective authorship practices become reasons for establishing particular review practices that influence the resulting analysis. Thus, the relation between the social and epistemic orders is understood as moving in both directions. By tying the epistemic results to the organizational structures of CERN, interviewees thus engage in boundary work that mobilizes precisely what is usually deemed absent from scientific knowledge production: situatedness. As Hine (2014) has argued, "new technologies" often constitute "new opportunities for [a] discipline to constitute itself as a discipline" (p.). Data, in the context of the physicist narratives, becomes an instrument for delineating the collaboration from other domains of particle physics. Therefore, the physicists' narratives can be conceptualized as ways of reinforcing the socio-epistemic structure of the collaboration. On the other hand, they need to be understood as concerns about the quality of the publications generated through CERN open data.

These co-productive lines of argumentation stand in contrast to understandings of research that are prevalent in funding agency requirements. In particular, there are two conflicting understandings that data producers and larger policy contexts articulate. First, a strong contrast is constituted by universal vs. situated/relational understandings of data. Funding agencies tend to understand data as easily detachable entities that acquire the same types of meaning in different socio-epistemic settings. In contrast, data producers at CERN tend to understand data as an ongoing objective of research that is still entangled with local production contexts.

Consequently, they do not think that data can easily sustain its identity outside of the context of the collaboration. This relates to the second contrast. While discourses around open data often take individual or small group (re-)use as an implicit presupposition, scientists at CERN tend to think of data as a collective product that can most successfully be processed and analyzed by the collaboration. Without the knowledge embedded in the collaboration's collective structures, scientists fear that analyses are at risk of being incorrect. Quite conversely, some interview respondents have argued that another hesitancy to open up data stems from the concern that external users could make a physics discovery before the collaboration:

> “Well, the other worry is that they would discover something. Now, it could be wrong, and that would be, of course, the analysis was incorrect. We don’t want that. But we also don’t want somebody to do a correct analysis and discover something before the collaboration does.”

If research with open data results in a physics discovery, this could undermine the efforts of the collaboration and consequently threaten its reputation. An interesting dissonance thus emerges within the narratives of the interview respondents. While the collaborative structure of CERN and the knowledge produced within it is understood as the “golden standard” for the production of physics results, the fear remains that a physics discovery could be missed by the collaboration, and external data users could capitalize on that.

However, many interviewees have argued that they deem a physics discovery with open data extremely unlikely. Further, the fear of an external discovery can be related to diverging anticipations of open data user groups. While one user group is assumed to lack the knowledge necessary to conduct a correct analysis, other user groups are assumed to be able to overcome this hurdle (maybe because of personal ties to the LHC collaborations) and discover new physics in the data.

6.4.3 Open data (re-)use by theorists

While some interviewees have eliminated the possibility of successful open data (re-)use altogether by mobilizing the boundary narrative outlined above, others have mobilized this argument only for particular (re-)use cases. These interviewees narrated theoretical physicists as the most promising (re-)use group. According to them, a growing number of theorists has expressed interest in working with open data provided by CERN collaborations over the last years. They described how theoretical particle physics is facing problems due to a lack in the confirmation of theoretical assumptions by the LHC thus far.

> “There was, I think, a very compelling picture of [...] why the LHC would be a very interesting machine. And many things were expected at the LHC and none of them turned out to be correct, other than the Higgs discovery. [...] And I think the theory community is also a little bit lost on what to do. And they need hints from the experiments to create new ideas, to go to the next steps.”

Theorists that work at the intersection of the experiment-theory divide, e.g. phenomenologists that develop models for the experimental testing of theories, are affected by this lack of confirmed ideas. Therefore, they have become increasingly interested in discussing their thoughts with experimentalists. Conversely, experimentalists are interested in receiving feedback from theorists on particular concepts and ideas. Thus, the “spirit of collaboration is very much alive” and exchanges happen regularly, though largely in an informal way. Next to informal, interpersonal exchanges, some theorists are interested in gaining access to the experimentalists' data in order to generate and test new ideas themselves. However, my interviewees have argued that their preferred way of exchanging data involves "short-term associations" where theorists officially join the collaboration for particular types of analysis:

> "Theorists or anyone who wants to work with ATLAS data can get a short-term association with ATLAS. That is, he or she presents what they want to do with our data. And then they become an ATLAS member and work with us and then they can publish with us. That's our principle. I think that's really better than open data. Because then you can be sure that [the analysis] is correct."

Short-term association thus means that theorists become part of the collaboration and get access to the "knowledge infrastructure" the collaboration embodies. Further, theorists have to adhere to the collective authorship practices defined by the collaboration. Importantly, the relevance of an analysis is judged by the collaboration, rather than the external theorist. Additionally, the finished product of analysis has to undergo the internal peer review structure and thus adhere to the collaboration's standard of correctness. In this way, short-term association accounts for the situatedness of data in the knowledge production procedures of the collaboration and, as such, reinstates the collaboration as the epistemic agent.

The concept of short-term association not only relies on an understanding of data as situated and relational in an epistemic sense. Additionally, it points to an understanding of data as being collectively owned by the collaboration. Individuals can only "rent" data for the time of their membership. In contrast to theoretical concepts or analysis techniques that "belong to" individual researchers and whom they can keep when they leave the collaboration, access rights to the data are given up. Therefore, individuals who want to work with the data, need to become part time members of the collaboration. Thus, concepts of short-term association not only rely on epistemic considerations but additionally on the idea of collaborative ownership and individual rentiership.

As one interviewee pointed out, informal types of exchange and short-term association leave theorists without social ties to the experiments at a disadvantage. Often, experimentalists and theorists discuss ideas personally "over a coffee" (e.g. with members of the CERN theory department) which is very different to impersonal email communication. These personal ties consequently play a role in short-term association. Discussing ideas regularly means being informed about the current objectives of the collaboration, which can impact the success of short-term associations. Thus, short-term association doesn't always offer all interested theorists equal opportunities.

Chapter three has demonstrated how the theoretical (re-)use of data from the CERN open data portal allowed researchers to advance particular ideas that were initially deemed uninteresting inside the collaboration. Thus, open data (re-)use from the CERN ODP is an instance where the power relations between theorists and experimentalists are transformed. Theorists are free to choose analysis models and don't have to undergo the collaboration's internal peer review. Discussions concerning open data thus become discussions about the relationship between theoretical and experimental particle physics. Should theorists be able to define data analysis techniques without consulting the data producers, the experimentalists? This is an unresolved question in the experimental community. While open data is not the preferred strategy of experimentalists, some have acknowledged the necessity of releasing particular types of data. One interviewee has argued that open data bears the potential to re-shape the experiment-theory relationship. First, it allows theorists without social ties to the collaboration to test their models on data taken by the detector. Thus, the dependency on informal forms of communication could be reduced by opening up data. Second, as one interviewee has argued, in experimental practice, it is not typical to apply more than one model to a particular set of data. Opening up data to theorists would allow for the testing of a variety of models on the same dataset and thus enable greater flexibility in comparing different analysis models. While the testing of models is also an important objective of short-term associations, open data could bring important organizational advantages:

> "[Through open data] the theorists can compare the same data with different models. [...] That I find very exciting, because that's something that wasn't the case before. There, you really worked together with the theorist, and they then sort of [...] mangled the data in a publication together with the theoretical interpretation. And you couldn't pull that apart, and if another theorist came along with a different idea, then you basically had to repeat the process again. Very inefficient and [...] very opaque."

During short-term association, experimentalists have to integrate theorists individually into the practices of the collaboration. When data is made publicly available, different theorists can apply models to the same data without extra effort on the side of the collaboration. Thus, the resources necessary to produce an analysis could be significantly reduced.

Data, in the context of theoretical (re-)use, is thus understood as both, situated and universal. It is an intermediary between the material infrastructure of the experiment and the finished result, the publication. Some researchers think that it is possible to provide data in ways that facilitate successful (re-)use. Others think that providing data should be connected to personal interactions between data users and producers.

Interestingly, the extent to which data is understood as situated seems to be dependent on the type of experiment that produces the data. As interviewees have outlined, the large LHC collaborations work with "very clear formatted" data, since the experimental apparatus, in particular the detector, is rarely changed and clearly understood. Researchers are generally aware of the data types they produce and the particles they can reconstruct from this data.

Their interest lies in processes out of which these reconstructed particles emerge. In smaller experiments at CERN, tinkering is part of daily practice. Often, parts of the experimental setup are rearranged, new elements are added, and others are removed. The data that is generated by these experiments is thus far more processual as it transforms with changes in the experimental setup. Consequently, contextualizing this data also requires accounting for the changes in the material setup of the experiment. However, this requires resources that researchers currently don't see at their disposal. This example demonstrates that even within CERN, experimental cultures diverge, and, as a consequence, data is understood in multiple ways. It is this multiplicity of data that funding agency requirements are currently failing to account for. In the following, I want to conclude by asking how diverging anticipations of data and open data could be brought into a conversation with one another.

6.4.4 Directions for concrete open data strategies

The question thus becomes how concrete open data strategies could take the various conceptions of data outlined above into account. Conversely, it is interesting to ask how general conceptions of open data and already established open data infrastructures could elicit discussions on the relation of HEP practice to larger scientific (e)valuation schemes. The physicists' narratives have underscored that open data is not valuable in its own right. Contemporary discourses on the digital often ascribe almost mythical powers of transformation to data (Hine, 2014). Data producers at CERN, however, view data as situated and contingent products of their experimental apparatus. Defining open release of data in terms of mere quantity thus can not guarantee successful (re-)use. In order to acknowledge the situated character of research data, open data implementers could think about developing social infrastructures for communities of (re-)use to develop next to providing technological open data infrastructures. This resonates with my interviewees' conception that the analysis of CERN data should be undertaken by collectives rather than single individuals.

> “I think what is needed is a community that develops around the already released data. And then, if there are interesting things that come out of it, they can be communicated back to us either through publishing or just sending an email or discussing [it] at a conference.”

Taking the co-production of the social and the epistemic seriously thus means actively attending to and shaping the social contexts of data (re-)use.

Conversely, by attending to open data (re-)use that has already taken place, researchers can gain a more comprehensive picture of the communities that are interested in CERN open data and thus develop more targeted strategies for selecting and contextualizing datasets. Studying open data (re-)use consequently helps in acquiring a better understanding of the ways in which future (re-)use could be enabled. Additionally, as subchapter two has demonstrated, open data (re-)use can point to value conflicts that prevail inside the collaborations. In the following, I conclude by summarizing the main findings and asking how they could inform future open data strategies.

7 Discussion and Conclusion

For several decades now, research in science technology studies (STS), anthropology, and philosophy of science has been interested in the ways in which knowledge is produced in High energy physics (e.g., Galison, 1997; Knorr-Cetina, 1999; Traweek, 1988). Detailed ethnographic inquiries into the world's largest laboratory for HEP, CERN, have shown how research practices in this subculture of physics are unique in many ways. Next to an exceptional socio-epistemic research infrastructure, research at CERN involved the development of unique technological applications from the onset. In search of the fundamental forces of nature, CERN researchers are designing continuously growing machines to produce and detect increasingly small subatomic particles. The research collaborations at CERN's largest particle accelerator, the Large Hadron Collider (LHC), operate detector technologies that are unmatched in size, complexity, and precision. Tremendous amounts of engineering expertise go into the development of these technologies. In order to design and operate them, particle physicists and engineering experts work closely together (Galison, 1997). This makes HEP practice an intrinsically interdisciplinary endeavor as engineers have become integrated into the research collaborations at CERN. Additionally, the increasing size and complexity of detector technologies have brought with them growing amounts of sophisticated datasets for physics analysis. Thus, the number of particle physicists who work in LHC collaborations is continuously growing. In light of this increased size and heterogeneity of the HEP collaborations, STS scholars have asked how the research groups coordinate and organize their practices.

This research project has investigated how data practices navigate the organization of research at CERN. Previous research on particle physics, such as the work of Knorr-Cetina (1999) has suggested that the "communitarian" structure of HEP research ensures that credit is attributed to all members of the collaboration. By enacting the collaboration, rather than the individual scientist as the central agent of knowledge production, researchers are subsumed under a collective whole, and the hierarchization of different tasks becomes less visible. This ensures that maintenance and development tasks, as well as data analysis practices, are undertaken by the collaboration members. Additionally, previous research has suggested that technological applications mediate the interests of different actor groups inside the collaboration. For instance, Boisot (2011) suggests that the detector acts as a "boundary object" between physicists and engineers. According to Boisot (2011), the detector is a common point of reference for collaboration members, allowing for an abstraction of group interests. Engineers, for instance, do not need to communicate a detailed description of technical problems to the physicists when they aim to optimize the precision of one of the detector's hardware elements. Instead, they may communicate an abstract representation of a problem and inform the physicists when they have achieved a particular improvement. Thus, the detector allows engineers and physicists to coordinate their efforts without communicating their complete working objectives.

The work of Boisot (2011) underscores how the detector acts as an ordering element in HEP practice that assembles an increasingly large number of researchers and coordinates their efforts. In a similar way, this research project has focused on how data infrastructures coordinate actor groups inside as well as outside of the LHC collaborations. By outlining how researchers produce and use datasets that are made available on data infrastructures at CERN, I draw attention to new sites of experimentation and theorization that have arisen through the datafication of research activities at CERN. These sites are infrastructures that distribute datasets to diverse research locations and gather a new composition of actors around them. While the detector draws together engineers and particle physicists, virtual data spaces assemble computational experts such as machine learning specialists, experimental particle physicists, and theoretical physicists. Interactions that occur via the detector enact a distribution of work, for instance, the development of hardware elements by engineers and the generation of datasets by physicists.

This can be applied in a similar way to data practices that take place on CERN's data infrastructures. Machine learning experts, for instance, generate AOD data samples from raw data files, which particle physicists later use for performing an analysis. In this way, data becomes an intermediary between the different forms of expertise that are mobilized in its production and analysis. In this work, I argue that the new pool of expertise that assembles around data infrastructures crucially impacts the epistemic results of research. Thus, in order to fully comprehend knowledge production activities in HEP research, we not only need to scrutinize how researchers design and use detector and collider technologies but additionally investigate the particularities of data infrastructures. Studying the producers and users of datasets means studying the socio-epistemic constitution of contemporary HEP research. In this project, I have focused on a particular data infrastructure that has allowed me to explore specific data practices in close detail: The CERN open data portal. Importantly, in the context of this project, I have understood the CERN open data portal as a new "place" within which research is done. This allowed me to mobilize the term "forms of togetherness" to describe how researchers get connected and disconnected in this new place through the data practices they perform.

With this project, I contribute to a body of work that has interrogated the entanglement of social, technological, and epistemic orders at CERN (e.g., Merz & Sorgner, 2022). Much previous work concerning HEP has focused on detector and collider technologies as crucial sites of experimentation. My project, however, investigates how digital sites shape the socio-epistemic condition of research (similar to (Karaca, 2020)). Thus, my work connects to research that has focused on the ways in which data infrastructures shape knowledge production practices (e.g., Hine, 2014; Leonelli & Tempini, 2020). By presenting a detailed case study of the development and use of an open data infrastructure, I additionally contribute to a growing body of research that investigates how open data projects are implemented in concrete research settings (Kitchin, 2014; Wessels et al., 2017).

In the empirical chapter, I have presented multiple forms of togetherness and apartness that emerge through data practices on the CERN open data portal. So far, I have outlined these practices separately in the different subchapters of the analysis. In the following, I draw them together and ask how they could help inform future open data practice at CERN.

7.1 Data practices at the CERN open data portal revisited

Above all, my analysis has shown how data practices at the CERN open data portal enact a variety of different forms of togetherness and apartness between the actors involved in its construction, maintenance, and use. Through the establishment of the ODP, a new space for experimenting and theorizing was created within which various forms of connection were established between a number of heterogeneous actors. The first subchapter of the empirical part has demonstrated how togetherness was enacted as a form of interpersonal collaboration between the members of the development team of the ODP. By focusing on practices of metadata application, I have shown how interpersonal exchange was essential in the early development phase of the portal, where questions regarding the ontology of metadata needed to be negotiated by the CERN IT department, the CERN scientific information service, and CMS scientists. In this phase, researchers needed to familiarize themselves with each other's working objectives and develop common understandings and agendas. Additionally, the open data team had to consult members of the LHC collaborations to extract metadata information that was stored partly non-standardized in the socio-epistemic structure of the collaboration. Ultimately, the team developed a metadata schema that started to act as an intermediary between the data providers of the LHC collaborations and the CERN IT department who assumed responsibility for uploading data to the ODP.

This metadata schema (See Appendix A, Table 1) has the form of a classificatory system that defines obligatory as well as optional metadata elements for the data providers. Through the schema, the LHC collaborations could standardize and partly align their provision of metadata for the ODP by adhering to the categories defined through the schema. At the same time, the different collaborations could retain some degree of freedom by filling in some optional fields rather than others. During the course of portal development, interpersonal forms of exchange concerning metadata application between the CERN IT department, the CERN scientific information service, and members of the LHC collaborations thus became interactions that were mediated through a pre-defined, partly standardized form and took place between members of CMS (and later members of other LHC collaborations) and the CERN IT department. The metadata schema became a "boundary object" (Star & Griesemer, 1989) that allowed the LHC teams to collaborate with the open data implementers without having to fully align or explain their practices of metadata production.

Thus, this subchapter underscores how infrastructuring practices have transformed personal collaboration into materially mediated forms of interaction. This shows that in research settings where data infrastructures become increasingly important for knowledge production and dissemination, studying direct forms of interaction between researchers is not sufficient in conceptualizing the social developments in a field. The social and the technological are always already inextricably linked in such environments and can thus only be studied together.

The second subchapter has departed from an interest in the ways in which enactments of collectivity structure knowledge production inside the research collaboration and asked how these enactments transform when collaboration members work with open data instead of internally available resources. As I have shown, collectivity inside the collaboration is produced through the enactment of the collaboration rather than the individual scientist as the producer of "real physics" results. By drawing on authorship and internal peer review practices, I have demonstrated how a "real physics" result is only considered attainable through the collective efforts of the collaboration. Only analyses that have managed to pass collective internal validation procedures are considered "real physics" results. When LHC researchers work with open data, they can not access this validation infrastructure, and their analysis acquires the status of a "proof of concept" publication rather than a "real physics" result. In this way, the use of open data produces an apartness between open data users and the collaborations that produced this data. The results of their work become de-collectivized and gain a novel epistemic status. While the first subchapter underscores the inextricability of the technological and the social, the second subchapter thus points to the entanglement of technological and epistemic orders as the infrastructure of data use becomes essential in defining the epistemic status of the result.

In the third subchapter of the empirical part, I have investigated open data (re-)use by theoretical physicists. The analysis has shown how theoretical (re-)use importantly relies on the availability of simulated data. Simulated data stores information on the functionalities of the detector and thus allows theorists to familiarize themselves with the material setup of the experiment. Through working with the experimentalists' simulated data, theorists could reorient their research with respect to the possibilities and restrictions defined through the experimental setup. Consequently, interaction through these datasets produced a new form of togetherness between theorists and experimentalists that aligned parts of their analysis techniques. Similarly to the second subchapter, the third subchapter thus points to the entanglement of technological and epistemic orders. However, in this case, (re-)use resulted in an alignment of the epistemic results rather than dis-alignment. As a result, the research of the theorists was taken up by experimentalists who started to work with observations made in open data publications.

The data practices outlined above thus underscore the entanglement of social, technological, and epistemic orders. The dynamics that take place on the portal demonstrate how these orders become inextricably linked. In the following, I want to point to three central conclusions which I draw from my work and ask what they could tell us for the future development of the ODP and open data practice more generally.

7.2 Three main takeaways

1. Data infrastructures such as the CERN open data portal are places for experimentation and theorizing, capable of shaping the socio-epistemic conditions and outcomes of research.

2. Building and maintaining these places involves constant processes of infrastructuring that are based on continuously changing anticipations of user groups that profit from the study of open data (re-)use and non-use cases.

3. The use of specific infrastructures, such as the CERN open data portal, is a way of negotiating individual and collective interests in HEP research and beyond it.

Much research at CERN concerns the preparation, simulation, and analysis of data. This project's investigation of development and (re-)use cases at the CERN ODP has shown that not only the construction and use of the detector are crucial in shaping research outcomes. Additionally, the ways in which data is processed and made available to particular communities through different infrastructures have a crucial effect on the research results. Experimentation is not only taking place at the detector but also at data infrastructures where increasingly sophisticated computational models are developed to refine and optimize data for analysis. With this thesis, I suggest that it is helpful to conceptualize data infrastructures as digital places in which proximity and distance are negotiated in new ways. Through connecting and disconnecting various actor groups in multiple ways, the CERN open data portal has led to the production of specific types of knowledge. By understanding data infrastructures such as the CERN open data portal as new places of academic practice, they acquire agency in shaping the socio-epistemic structures of research.

Subchapter two of the findings section has underscored how the use of the collaboration's internal data and peer review infrastructures enacts research as a collectively produced “real physics” result, while the (re-)use of data from the CERN open data portal enacts research as an individually produced “proof of concept” publication. Subchapter three has demonstrated how the use of open data by theoretical physicists has aligned parts of their analysis practices with those of the experimentalists. Their newly acquired access to the experimentalists' data thus reoriented the theorists' research toward the functionalities of the detector. Both examples show how the choice of a particular data infrastructure defines the epistemic constitution of the research result.

Thus, the CERN open data portal enabled different user groups to align and misalign their practices with CERN internal data processing and analysis standards. Questions on preserving data and making it accessible to different user groups should thus be central in deliberating the types of epistemic worth that HEP experiments can produce. Discussing the potentials of HEP research not only means debating the possibilities of current and future collider and detector technologies but deliberating the current and future capabilities of particular infrastructures for storing data and techniques for processing, sharing, and analyzing data. The question of how to preserve HEP data and how to make it accessible to communities outside of experimental particle physics should thus gain a more prominent place in deliberations concerned with the future of HEP practice. My analysis suggests that, in particular, contributions to research fields other than particle physics are inhibited through current open data practices that often aim to produce "real physics" results. By developing and improving infrastructures that enable contributions to other fields of research (e.g., machine learning), CERN could more clearly distinguish itself as a knowledge producer that generates epistemic worth across disciplinary boundaries.

Questions regarding the preservation and sharing of CERN data include discussions on how to build and optimize infrastructures that can disseminate data effectively. In this regard, my analysis suggests that researchers should be attentive to the processual character of developing and maintaining such infrastructures. In subchapter one of the findings section, the development of a metadata schema for the portal has underscored that the implementation of digital infrastructures involves a variety of choices and anticipations that shape its ultimate form. In the case of the CERN ODP (and this is true for many (open) data infrastructures), anticipations of the potential users were essential in defining a metadata ontology. For instance, the ODP development team anticipated the use of open data by machine learning specialists and thus decided to add raw data samples as possible metadata attributes to the schema.

Therefore, classification practices for open data infrastructures are strongly tied to the anticipation of (re-)use groups. My findings suggest that a discussion on the potential users of open data could help inform future practice. Questions concerning classification are tied to whether open data (re-)use groups and their needs can be comprehensively anticipated. No formalization can meet the needs of all (re-)use cases. Therefore, the potential of successful (re-)use needs to be discussed in reference to different types of (re-)use groups separately. While these discussions are already taking place, they usually occur between members of the ODP development team. I suggest that a broader debate in the research collaborations at CERN could help improve (re-)use potential. The collaborations should discuss: Who wants to use our data? And: What contextualization and classification work would be necessary to make (re-)use possible for this particular community? Such a discussion could help the collaborations weigh (re-)use potential against the effort connected to making data open.

However, (re-)use potential can never be fully anticipated. Therefore, it is helpful not only to discuss data (re-)use cases internally but study actual (re-)use that has already taken place. My analysis of open data (re-)use underscores how the CERN ODP, as a new "place" within which research is done, matters in multiple ways in shaping togetherness/apartness relations or, more generally, the socio-epistemic structure of research. Rather than affecting all enrolled actors similarly, the open data portal enacted socio-epistemic transformations ranging from engagement to estrangement, depending on the particular context of (re-)use. By attending to those (re-)use cases, more targeted strategies toward releasing open data could be developed. For instance, the (re-)use of open data by theoretical physicists underscores the relevance of simulated data samples. Developing strategies for familiarizing researchers more closely with these data types would be a fruitful angle for future investigation.

Thus, developing platforms for sharing, processing, and analyzing HEP data involves processes of infrastructuring in which (re-)use communities and their needs have to be continuously (re-)conceptualized. Since it is impossible to anticipate all types of (re-)use before data is made available, infrastructure developers should study how interested communities engage in data. This allows them to understand the kinds of friction that arise in particular (re-)use contexts. Additionally, feedback circles with open data users are likely to yield examples where (re-)use was inhibited completely, and infrastructuring processes could be undertaken to enable future use.

Next to gaining valuable information on future development potentials, open data (re-)use cases can point to larger struggles that persist in HEP research. By choosing particular infrastructures, researchers can pursue specific goals and interests. For instance, the use of open data by CMS members demonstrates the researchers' interest in publishing "proof of concept" publications that enable individual credit attribution and faster publication. By underscoring the relevance of the ODP as an alternative publishing venue rather than an epistemic resource, this use case points to a value struggle that persists in the LHC collaborations. While informal types of individual credit attribution prevail in HEP, formal, quantitative performance measures are common in research areas outside of HEP. Particularly young researchers, who want to perform well in both systems, struggle with these often conflicting modes of evaluation. This use case thus suggests that the LHC collaborations should discuss how more formal types of credit attribution and shorter peer review processes could be realized for particular kinds of analysis, specifically those types of research that exist at the intersection of particle physics and other areas of study. My analysis suggests that an internal publication avenue for the individual publication of "proof of concept" results would not threaten the collective production of "real physics" results. However, it could increase CERN's contributions to research fields other than particle physics.

Further, the use of open data by theoretical particle physicists underscored their interest in (re-)orienting theoretical research toward the current capabilities of detector and collider technologies. Additionally, the theorists used open data to push for particular research objectives in experimental practice. By using the experimentalists' data, theorists could pursue issues that were not considered interesting inside the collaboration to ultimately convince experimentalists to follow their idea internally. This (re-)use example thus suggests that discussions on the relation between experiment and theory could be held more explicitly in the particle physics community: Should theorists have a say in defining the research objectives of experimentalists? Should theoretical research be more closely oriented toward the current capabilities of the experiments? By making these questions more explicit, researchers could think about developing infrastructures (both social and technical) for eliciting particular types of exchange.

7.3 Areas for further research

To conclude, I want to outline areas for further investigation beyond the scope of this thesis. First, I suggest that a further investigation of data journeys through CERN internal infrastructures could yield interesting results. The data journey outlined at the beginning of the findings section was quite fragmented since I could only draw on publicly available material. With access to CERN internal databases, it could be possible to gain a better understanding of the ways in which data circulates inside CERN and how factors such as field of expertise, collaboration membership, or social status determine how researchers interact with and connect through their data practices. Generally, a study of internal infrastructures could prove helpful in explaining how particular types of data come into being and what open data infrastructures should be attentive to when releasing this data to specific (re-)use communities.

Additionally, the relevance of data practices in shaping the experiment-theory relationship could be further investigated. This research project has shown how the relationship between theory and experiment is newly articulated through data (re-)use at the ODP. As the third subsection of the findings underscored, the use of simulated data by theoretical particle physicists led to an alignment of the epistemic results between the two groups. However, the (re-)use of experimental data by theorists is quite rare. Therefore, it is not entirely clear whether theoretical research produced with open data has the capacity to substantially impact research objectives in experimental communities. This aspect could be one central focus in the study of future (re-)use cases.

Further, the interaction of experimentalists and theorists through the ODP raises the question of whether other research practices (e.g., short-term association, informal types of exchange) enact forms of togetherness between those two groups. By investigating these other types of interaction, local “trading zones” and the particular types of knowledge produced through them become visible. By comparing the knowledge produced through short-term association with knowledge produced through open data use (“proof of concept” work), a deliberation within the community on the worth of different kinds of interaction could be triggered and enable a debate on the role of data in directing and shaping the experiment-theory relationship.

While this project has attended to unanticipated open data (re-)use cases (See 6.2) and data friction arising from (re-)use (See 6.3.2), it did not attend to cases where data use was completely inhibited. However, instances, where use is disabled, are equally important in understanding the dynamics around open data infrastructures and defining future avenues for open data development. Thus, an avenue for future investigation is cases of non-use. Conversations with open data developers and maintainers are likely to reveal instances where data could not be used that will engender deeper insights into the problems that can arise when specific communities work with CERN data.

Finally, I suggest that funding agency requirements toward open data and their impact on the implementation of open data infrastructures are interesting areas of further research. Subsection four of the findings chapter suggests a misalignment between funding requirements and the ways in which scientists conceptualize successful open data (re-)use. Thus, there is a chance that data is stored and made accessible in a way that, rather than eliciting (re-)use, solely allows collaborations to fulfill their funding requirements. By investigating how funding agencies could refine their open data requirements to account for the particular user groups of HEP data, pathways to more inclusive research practices in HEP could be outlined.

References

Albertsson, K., Altoe, P., Anderson, D., Andrews, M., Araque Espinosa, J. P., Aurisano, A., ... Zapata, O. (2018, September). Machine Learning in High Energy Physics Community White Paper. *Journal of Physics: Conference Series*, *1085*, 022008. Retrieved 2023-01-27, from `https://iopscience.iop.org/article/10.1088/1742-6596/1085/2/022008` doi: 10.1088/1742-6596/1085/2/022008

Appel, H., Anand, N., & Gupta, A. (2020, December). Introduction: Temporality, Politics, and the Promise of Infrastructure. In N. Anand, A. Gupta, & H. Appel (Eds.), *The Promise of Infrastructure* (pp. 1–38). Duke University Press. Retrieved 2022-06-04, from `https://www.degruyter.com/document/doi/10.1515/9781478002031-002/html` doi: 10.1515/9781478002031-002

Aula, V. (2019, July). Institutions, infrastructures, and data friction – Reforming secondary use of health data in Finland. *Big Data & Society*, *6*(2), 205395171987598. Retrieved 2021-11-18, from `http://journals.sagepub.com/doi/10.1177/2053951719875980` doi: 10.1177/2053951719875980

Barlösius, E. (2019, November). Concepts of Originality in the Natural Science, Medical, and Engineering Disciplines: An Analysis of Research Proposals. *Science, Technology, & Human Values*, *44*(6), 915–937. Retrieved 2022-08-16, from `http://journals.sagepub.com/doi/10.1177/0162243918808370` doi: 10.1177/0162243918808370

Barry, A. (2006, May). Technological Zones. *European Journal of Social Theory*, *9*(2), 239–253. Retrieved 2021-11-16, from `http://journals.sagepub.com/doi/10.1177/1368431006063343` doi: 10.1177/1368431006063343

Bates, J., Lin, Y.-W., & Goodale, P. (2016, December). Data journeys: Capturing the socio-material constitution of data objects and flows. *Big Data & Society*, *3*(2), 205395171665450. Retrieved 2021-09-14, from `http://journals.sagepub.com/doi/10.1177/2053951716654502` doi: 10.1177/2053951716654502

Birnholtz, J. (2008, January). When Authorship Isn't Enough: Lessons from CERN on the Implications of Formal and Informal Credit Attribution Mechanisms in Collaborative Research. *The Journal of Electronic Publishing*, *11*(1). Retrieved 2022-08-09, from `http://hdl.handle.net/2027/spo.3336451.0011.105` doi: 10.3998/3336451.0011.105

Boisot, M. (2011, September). Generating knowledge in a connected world: The case of the ATLAS experiment at CERN. *Management Learning*, *42*(4), 447–457. Retrieved 2022-07-21, from `http://journals.sagepub.com/doi/10.1177/1350507611408676` doi: 10.1177/1350507611408676

Boltanski, L., & Thévenot, L. (2006). *On justification: economies of worth.* Princeton: Princeton University Press.

Bowker, G. C., & Star, S. L. (1999). *Sorting things out: classification and its consequences.* Cambridge, Mass: MIT Press.

Bressan, B., & Boisot, M. (2011, July). The Individual in the ATLAS Collaboration: A Learning Perspective. In M. Boisot, M. Nordberg, S. Yami, & B. Nicquevert (Eds.), *Collisions and Collaboration: The Organization of Learning in the ATLAS Experiment at the LHC.* OUP Oxford. (Google-Books-ID: MXrdJrzdfy0C)

Bueger, C. (2014). Narrative Praxiographie. Klandestine Praktiken und das ‚Grand Narrativ' Somalischer Piraterie. In F. Gadinger, S. Jarzebski, & T. Yildiz (Eds.), *Politische Narrative* (pp. 201–223). Wiesbaden: Springer Fachmedien Wiesbaden. Retrieved 2021-11-18, from `http://link.springer.com/10.1007/978-3-658-02581-6_8` doi: 10.1007/978-3-658-02581-6_8

CERN. (2020). *CERN Open Data Policy for the LHC Experiments.* Retrieved 2023-01-04, from `https://opendata.cern.ch/docs/cern-open-data-policy-for-lhc-experiments`

CERN. (2021). *Welcome.* Retrieved 2021-10-14, from `https://wlcg-public.web.cern.ch/`

CERN. (2022a). *Data Centre | IT Department.* Retrieved 2022-12-30, from `https://information-technology.web.cern.ch/about/data-centre`

CERN. (2022b). *The Worldwide LHC Computing Grid (WLCG).* Retrieved 2022-09-19, from `https://home.cern/science/computing/grid`

CERN. (2023). *CERN Open Data Portal.* Retrieved 2022-01-12, from `https://opendata.cern.ch/docs/about`

CERN open data. (2023a). *cernopendata/opendata.cern.ch.* Retrieved 2023-01-10, from `https://gitter.im/cernopendata/opendata.cern.ch`

CERN open data. (2023b). *CERN Open Data portal.* CERN Open Data. Retrieved 2023-01-10, from `https://github.com/cernopendata/opendata.cern.ch/blob/9fb73eeba72a691615852916975f209b79fc211a/cernopendata/jsonschemas/records/record-v1.0.0.json` (original-date: 2014-06-23T09:12:36Z)

Cesarotti, C., Soreq, Y., Strassler, M., Thaler, J., & Xue, W. (2019). Searching in CMS open data for dimuon resonances with substantial transverse momentum. *Physical Review D*, *100*(1). Retrieved from `https://www.scopus.com/inward/record.uri?eid=2-s2.0-85070269259&doi=10.1103%2fPhysRevD.100.015021&partnerID=40&md5=be0d5eff51af6d39749bf538d84b8fd6` (Publisher: American Physical Society) doi: 10.1103/PhysRevD.100.015021

Chatrchyan, S., Khachatryan, V., Sirunyan, A., Tumasyan, A., Adam, W., Aguilo, E., … Wenman, D. (2012, September). Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters B*, *716*(1), 30–61. Retrieved 2021-12-09, from `https://linkinghub.elsevier.com/retrieve/pii/S0370269312008581` doi: 10.1016/j.physletb.2012.08.021

Chimirri, D. (2021, September). Studying how tourism is done: A practice approach to collaboration. *Tourist Studies*, *21*(3), 347–366. Retrieved 2021-11-18, from http://journals.sagepub.com/doi/10.1177/1468797621998286 doi: 10.1177/1468797621998286

CMS Collaboration, CERN. (2016a). DoubleMu primary dataset in AOD format from RunA of 2011 (/DoubleMu/Run2011A-12Oct2013-v1/AOD). Retrieved 2021-12-21, from http://opendata.cern.ch/record/17 (Publisher: CERN Open Data Portal) doi: 10.7483/OPENDATA.CMS.RZ34.QR6N

CMS Collaboration, CERN. (2016b). Simulated dataset DYJetsToLL_m-10To50_tunez2_7tev-pythia6 in AODSIM format for 2011 collision data (SM Inclusive). Retrieved 2021-12-21, from http://opendata.cern.ch/record/1393 (Publisher: CERN Open Data Portal) doi: 10.7483/OPENDATA.CMS.T8RZ.D52D

CMS Collaboration, CERN. (2016c). Simulated dataset DYJetsToLL_m-50_7tev-madgraph-pythia6-tauola in AODSIM format for 2011 collision data (SM Inclusive). Retrieved 2021-12-21, from http://opendata.cern.ch/record/1394 (Publisher: CERN Open Data Portal) doi: 10.7483/OPENDATA.CMS.4475.SSXE

CMS Collaboration, CERN. (2016d). Simulated dataset SMHiggsToZZTo4L_m-125_7tev-powheg15-JHUgenV3-pythia6 in AODSIM format for 2011 collision data (SM Higgs). Retrieved 2021-12-21, from http://opendata.cern.ch/record/1507 (Publisher: CERN Open Data Portal) doi: 10.7483/OPENDATA.CMS.K9EW.KRDS

CMS Collaboration, CERN. (2016e). Simulated dataset TTTo2L2Nu2B_7tev-powheg-pythia6 in AODSIM format for 2011 collision data (SM Inclusive). Retrieved 2021-12-21, from http://opendata.cern.ch/record/1360 (Publisher: CERN Open Data Portal) doi: 10.7483/OPENDATA.CMS.CSJG.AWBA

CMS Collaboration, CERN. (2016f). Simulated dataset ZZTo2e2mu_mll4_7tev-powheg-pythia6 in AODSIM format for 2011 collision data (SM Inclusive). Retrieved 2021-12-21, from http://opendata.cern.ch/record/1382 (Publisher: CERN Open Data Portal) doi: 10.7483/OPENDATA.CMS.7G2E.4PGB

CMS Collaboration, CERN. (2016g). Simulated dataset ZZTo4e_mll4_7tev-powheg-pythia6 in AODSIM format for 2011 collision data (SM Inclusive). Retrieved 2021-12-21, from http://opendata.cern.ch/record/1648 (Publisher: CERN Open Data Portal) doi: 10.7483/OPENDATA.CMS.95DX.4BMP

CMS Collaboration, CERN. (2016h). Simulated dataset ZZTo4mu_mll4_7tev-powheg-pythia6 in AODSIM format for 2011 collision data (SM Inclusive). Retrieved 2021-12-21, from http://opendata.cern.ch/record/1651 (Publisher: CERN Open Data Portal) doi: 10.7483/OPENDATA.CMS.XWVK.M4VG

CMS Collaboration, CERN. (2017a). */DoubleElectron/Run2012B-22Jan2013-v1/AOD.* CERN Open Data Portal. Retrieved 2021-12-21, from http://opendata.cern.ch/record/6003 (Type: dataset) doi: 10.7483/OPENDATA.CMS.S0H8.LBD3

CMS Collaboration, CERN. (2017b). */DoubleElectron/Run2012C-22Jan2013-v1/AOD.* CERN Open Data Portal. Retrieved 2021-12-21, from http://opendata.cern.ch/record/6029 (Type: dataset) doi: 10.7483/OPENDATA.CMS.SINM.BV86

CMS Collaboration, CERN. (2017c). */DoubleMuParked/Run2012B-22Jan2013-v1/AOD.* CERN Open Data Portal. Retrieved 2021-12-21, from http://opendata.cern.ch/record/6004 (Type: dataset) doi: 10.7483/OPENDATA.CMS.YLIC.86ZZ

CMS Collaboration, CERN. (2017d). */DoubleMuParked/Run2012C-22Jan2013-v1/AOD.* CERN Open Data Portal. Retrieved 2021-12-21, from http://opendata.cern.ch/record/6030 (Type: dataset) doi: 10.7483/OPENDATA.CMS.M5AD.Y3V3

CMS Collaboration, CERN. (2017e). */DYJetsToLL_m-10to50_ht-200to400_tunez2star_8tev-madgraph-tauola/Summer12_dr53x-PU_s10_start53_v19-v1/AODSIM.* CERN Open Data Portal. Retrieved 2021-12-21, from http://opendata.cern.ch/record/7727 (Type: dataset) doi: 10.7483/OPENDATA.CMS.F7AB.F8SV

CMS Collaboration, CERN. (2017f). */DYJetsToLL_m-10to50_ht-400toInf_tunez2star_8tev-madgraph-tauola/Summer12_dr53x-PU_s10_start53_v19-v1/AODSIM.* CERN Open Data Portal. Retrieved 2021-12-21, from http://opendata.cern.ch/record/7728 (Type: dataset) doi: 10.7483/OPENDATA.CMS.H6M1.8471

CMS Collaboration, CERN. (2017g). */DYJetsToLL_m-50_tunez2star_8tev-madgraph-tarball-tauola-tauPolarOff/Summer12_dr53x-PU_s10_start53_v19-v1/AODSIM.* CERN Open Data Portal. Retrieved 2021-12-21, from http://opendata.cern.ch/record/7731 (Type: dataset) doi: 10.7483/OPENDATA.CMS.DRSP.TO3O

CMS Collaboration, CERN. (2017h). */SMHiggsToZZTo4L_m-125_8tev-powheg15-JHUgenV3-pythia6/Summer12_dr53x-PU_s10_start53_v19-v1/AODSIM.* CERN Open Data Portal. Retrieved 2021-12-21, from http://opendata.cern.ch/record/9356 (Type: dataset) doi: 10.7483/OPENDATA.CMS.G13X.TDSB

CMS Collaboration, CERN. (2017i). */TTbar_8tev-Madspin_amcatnlo-herwig/Summer12_dr53x-PU_s10_start53_v19-v2/AODSIM.* CERN Open Data Portal. Retrieved 2021-12-21, from http://opendata.cern.ch/record/9518 (Type: dataset) doi: 10.7483/OPENDATA.CMS.XH95.JNSE

CMS Collaboration, CERN. (2017j). */ZZTo2e2mu_8tev-powheg-pythia6/Summer12_dr53x-PU_rd1_start53_v7n-v2/AODSIM.* CERN Open Data Portal. Retrieved 2021-12-21, from http://opendata.cern.ch/record/10054 (Type: dataset) doi: 10.7483/OPENDATA.CMS.461G.HELP

CMS Collaboration, CERN. (2017k). */ZZTo4e_8tev-powheg-pythia6/Summer12_dr53x-PU_rd1_start53_v7n-v2/AODSIM.* CERN Open Data Portal. Retrieved 2021-12-21, from http://opendata.cern.ch/record/10065 (Type: dataset) doi: 10.7483/OPENDATA.CMS.PEOD.ZZJY

CMS Collaboration, CERN. (2017l). */ZZTo4mu_8tev-powheg-pythia6/Summer12_dr53x-*

PU_rd1_start53_v7n-v1/AODSIM. CERN Open Data Portal. Retrieved 2021-12-21, from http://opendata.cern.ch/record/10071 (Type: dataset) doi: 10.7483/OPENDATA.CMS.HJ1F.5U9R

CMS Experiment. (n.d.). *Publishing an Analysis.* Retrieved 2022-08-20, from https://cms.cern/content/publishing-analysis

CMS Experiment. (2022). *Blinding and unblinding analyses.* Retrieved 2022-07-20, from https://cms.cern/physics/cms-higgs-search/blinding-and-unblinding-analyses

CMS Collaboration, C. (2022). *Triggering and Data Acquisition | CMS Experiment.* Retrieved 2022-12-30, from https://cms.cern/detector/triggering-and-data-acquisition

Convention for the Establishment of a European Organization for Nuclear Research | CERN Council. (1953, July). Retrieved 2022-06-18, from https://council.web.cern.ch/en/content/convention-establishment-european-organization-nuclear-research

Daston, L., & Galison, P. (2007). *Objectivity.* New York : Cambridge, Mass: Zone Books ; Distributed by the MIT Press. (OCLC: ocn144570876)

Decuypere, M. (2021, January). The Topologies of Data Practices: A Methodological Introduction. *Journal of New Approaches in Educational Research*, *10*(1), 67. Retrieved 2022-01-19, from https://naerjournal.ua.es/article/view/v10n1-5 doi: 10.7821/naer.2021.1.650

Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C., & Borgman, C. L. (2011, October). Science friction: Data, metadata, and collaboration. *Social Studies of Science*, *41*(5), 667–690. Retrieved 2021-09-17, from http://journals.sagepub.com/doi/10.1177/0306312711413314 doi: 10.1177/0306312711413314

Felt, U. (Ed.). (2009). *Knowing and living in academic research: convergences and heterogeneity in research cultures in the European context.* Prague: Inst. of Sociology of the Acad. of Sciences of the Czech Republic.

Fochler, M., Felt, U., & Müller, R. (2016, June). Unsustainable Growth, Hyper-Competition, and Worth in Life Science Research: Narrowing Evaluative Repertoires in Doctoral and Postdoctoral Scientists' Work and Lives. *Minerva*, *54*(2), 175–200. Retrieved 2022-07-25, from http://link.springer.com/10.1007/s11024-016-9292-y doi: 10.1007/s11024-016-9292-y

Franklin, A. (2013). *Shifting standards: experiments in particle physics in the twentieth century.* Pittsburgh, Pa.: University of Pittsburgh Press. (OCLC: 870684378)

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

Galison, P. (1997). *Image and logic: a material culture of microphysics.* Chicago: University of Chicago Press.

Galison, P. (2003). The Collective Author. In M. Biagioli & P. Galison (Eds.), *Scientific authorship: credit and intellectual property in science.* New York, NY: Routledge.

Galletta, A. (2013). *Mastering the semi-structured interview and beyond: from research design to analysis and publication.* New York: New York University Press.

García-Sancho, M. (2011). From metaphor to practices: The introduction of "information engineers" into the first DNA sequence database. *History and philosophy of the life sciences*, *33*(1), 71–104. (Place: CHAM Publisher: SPRINGER INTERNATIONAL PUBLISHING AG)

Gieryn, T. F. (1983, December). Boundary-Work and the Demarcation of Science from Non-Science: Strains and Interests in Professional Ideologies of Scientists. *American Sociological Review*, *48*(6), 781. Retrieved 2022-08-22, from http://www.jstor.org/stable/2095325?origin=crossref doi: 10.2307/2095325

Gieryn, T. F. (2000, August). A Space for Place in Sociology. *Annual Review of Sociology*, *26*(1), 463–496. Retrieved 2023-01-05, from https://www.annualreviews.org/doi/10.1146/annurev.soc.26.1.463 doi: 10.1146/annurev.soc.26.1.463

Gupta, A. (2020, December). 2. The Future in Ruins: Thoughts on the Temporality of Infrastructure. In N. Anand, A. Gupta, & H. Appel (Eds.), *The Promise of Infrastructure* (pp. 62–79). Duke University Press. Retrieved 2022-11-25, from https://www.degruyter.com/document/doi/10.1515/9781478002031-004/html doi: 10.1515/9781478002031-004

Halkier, B. (2017). Questioning the 'Gold Standard' Thinking in Qualitative Methods from a Practice Theoretical Perspective: Towards Methodological Multiplicity. In M. Jonas, B. Littig, & A. Wroblewski (Eds.), *Methodological Reflections on Practice Oriented Theories* (pp. 193–204). Cham: Springer International Publishing. Retrieved 2021-11-18, from http://link.springer.com/10.1007/978-3-319-52897-7_13 doi: 10.1007/978-3-319-52897-7_13

Hallonsten, O. (2012, August). Continuity and Change in the Politics of European Scientific Collaboration. *Journal of Contemporary European Research*, *8*(3). Retrieved 2022-11-29, from https://jcer.net/index.php/jcer/article/view/366 doi: 10.30950/jcer.v8i3.366

Hallonsten, O. (2020, August). Research Infrastructures in Europe: The Hype and the Field. *European Review*, *28*(4), 617–635. Retrieved 2022-11-29, from https://www.cambridge.org/core/product/identifier/S1062798720000095/type/journal_article doi: 10.1017/S1062798720000095

Hine, C. (2014). *Systematics As Cyberscience Computers, Change, and Continuity in Science.* Cambridge: MIT Press. (OCLC: 1004846566)

Jensen, E. A., & Laurie, A. C. (2016). *Doing real research: a practical guide to social research.* Los Angeles: SAGE. (OCLC: ocn936005533)

Johnson, J. (1988, June). Mixing Humans and Nonhumans Together: The Sociology of a Door-Closer. *Social Problems*, *35*(3), 298–310. Retrieved 2022-11-14, from http://

www.jstor.org/stable/800624 doi: 10.1525/sp.1988.35.3.03a00070

Jomhari, N. Z., Geiser, A., & Bin Anuar, A. A. (2017). *Higgs-to-four-lepton analysis example using 2011-2012 data.* CERN Open Data Portal. Retrieved 2021-12-21, from http://opendata.cern.ch/record/5500 doi: 10.7483/OPENDATA.CMS.JKB8.RR42

Kahn, M. (2018, February). Co-authorship as a proxy for collaboration: a cautionary tale. *Science and Public Policy*, *45*(1), 117–123. Retrieved 2021-10-18, from https://academic.oup.com/spp/article/45/1/117/4159476 doi: 10.1093/scipol/scx052

Karaca, K. (2020). What data get to travel in high energy physics? The construction of data at the large hadron collider. In S. Leonelli & N. Tempini (Eds.), (pp. 45–58). Springer International Publishing AG. Retrieved 2021-09-22, from https://www.narcis.nl/publication/RecordID/oai:ris.utwente.nl:publications%2F1e7d4179-e7d4-4eb3-b7d6-899904d4736e

Karasti, H., & Baker, K. (2004). Infrastructuring for the long-term: ecological information management. In *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the* (p. 10 pp.). Big Island, HI, USA: IEEE. Retrieved 2023-01-02, from http://ieeexplore.ieee.org/document/1265077/ doi: 10.1109/HICSS.2004.1265077

Karasti, H., Millerand, F., Hine, C. M., & Bowker, G. C. (2016a, May). Knowledge infrastructures: Part II. *Science & Technology Studies*, *29*(2), 2–6. Retrieved 2022-11-28, from https://sciencetechnologystudies.journal.fi/article/view/55961 doi: 10.23987/sts.55961

Karasti, H., Millerand, F., Hine, C. M., & Bowker, G. C. (2016b, December). Knowledge Infrastructures: Part IV. *Science & Technology Studies*, *29*(4), 2–9. Retrieved 2022-11-26, from https://sciencetechnologystudies.journal.fi/article/view/60220 doi: 10.23987/sts.60220

Kasemann, M. (2021, April). *EprRulesExplained.* Retrieved 2023-01-23, from https://twiki.cern.ch/twiki/bin/view/Main/EprRulesExplained

Katz, J. S., & Martin, B. R. (1997, March). What is research collaboration? *Research Policy*, *26*(1), 1–18. Retrieved 2021-09-17, from https://www.sciencedirect.com/science/article/pii/S0048733396009171 doi: 10.1016/S0048-7333(96)00917-1

Kitchin, R. (2014). *The data revolution: big data, open data, data infrastructures & their consequences.* Los Angeles, California: SAGE Publications. (OCLC: ocn871211376)

Klein, J. R., & Roodman, A. (2005, December). BLIND ANALYSIS IN NUCLEAR AND PARTICLE PHYSICS. *Annual Review of Nuclear and Particle Science*, *55*(1), 141–163. Retrieved 2022-07-18, from https://www.annualreviews.org/doi/10.1146/annurev.nucl.55.090704.151521 doi: 10.1146/annurev.nucl.55.090704.151521

Knorr-Cetina, K. (1995). How Superorganisms Change: Consensus Formation and the Social Ontology of High-Energy Physics Experiments. *Social Studies of Science*, *25*(1), 119–

147. Retrieved 2021-09-16, from http://www.jstor.org/stable/285527 (Publisher: Sage Publications, Ltd.)

Knorr-Cetina, K. (1999). *Epistemic cultures: how the sciences make knowledge.* Cambridge, Mass: Harvard University Press.

Krige, J. (2003). The politics of European scientific collaboration. Retrieved from https://cds.cern.ch/record/2005108

Larkin, B. (2013, October). The Politics and Poetics of Infrastructure. *Annual Review of Anthropology*, *42*(1), 327–343. Retrieved 2022-11-25, from https://www.annualreviews.org/doi/10.1146/annurev-anthro-092412-155522 doi: 10.1146/annurev-anthro-092412-155522

Lassila-Perini, K., Lange, C., Carrera Jarrin, E., & Bellis, M. (2021). Using CMS Open Data in research – challenges and directions. *EPJ Web of Conferences*, *251*, 01004. Retrieved 2021-09-09, from https://www.epj-conferences.org/10.1051/epjconf/202125101004 doi: 10.1051/epjconf/202125101004

Latour, B., & Woolgar, S. (1979). *Laboratory Life: The Construction of Scientific Facts.* Princeton University Press. Retrieved 2021-09-22, from https://www.degruyter.com/document/doi/10.1515/9781400820412/html (Publication Title: Laboratory Life) doi: 10.1515/9781400820412

Lave, J., & Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation.* Cambridge University Press. (Google-Books-ID: CAVIOrW3vYAC)

Leonelli, S. (2009, December). On the Locality of Data and Claims about Phenomena. *Philosophy of Science*, *76*(5), 737–749. Retrieved 2021-10-13, from https://www.journals.uchicago.edu/doi/10.1086/605804 doi: 10.1086/605804

Leonelli, S. (2020). Learning from data journeys. In S. Leonelli & N. Tempini (Eds.), *Data Journeys in the Sciences* (pp. 1–24). Springer International Publishing.

Leonelli, S., Rappert, B., & Davies, G. (2017, March). Data Shadows: Knowledge, Openness, and Absence. *Science, Technology, & Human Values*, *42*(2), 191–202. Retrieved 2021-11-18, from http://journals.sagepub.com/doi/10.1177/0162243916687039 doi: 10.1177/0162243916687039

Leonelli, S., & Tempini, N. (Eds.). (2020). *Data Journeys in the Sciences.* Cham: Springer International Publishing. Retrieved 2021-09-09, from http://link.springer.com/10.1007/978-3-030-37177-7 doi: 10.1007/978-3-030-37177-7

Lipton, V. (2020). *Open Scientific Data - Why Choosing and Reusing the RIGHT DATA Matters.* IntechOpen. Retrieved 2021-09-08, from https://www.intechopen.com/books/open-scientific-data-why-choosing-and-reusing-the-right-data-matters doi: 10.5772/intechopen.87201

Mehdiabadi, S. P., & Fahim, A. (2019, September). Explicit jet veto as a tool to purify the underlying event in the Drell–Yan process using CMS Open Data. *Journal of Physics*

G: Nuclear and Particle Physics, *46*(9), 095003. Retrieved 2021-12-09, from `https://iopscience.iop.org/article/10.1088/1361-6471/ab33a9` doi: 10.1088/1361-6471/ab33a9

Merz, M., & Sorgner, H. (2022, June). Organizational complexity in big science: strategies and practices. *Synthese*, *200*(3), 211. Retrieved 2022-08-09, from `https://link.springer.com/10.1007/s11229-022-03649-3` doi: 10.1007/s11229-022-03649-3

Michael, M. (2016). *Actor network theory: trials, trails and translations* (1st edition ed.). Thousand Oaks, CA: SAGE Ltd.

Miller, D., & Slater, D. (2000). *The Internet: an ethnographic approach.* Oxford ; New York: Berg. (OCLC: ocm44772097)

Misa, T. J., & Schot, J. (2005, March). Introduction: Inventing Europe: Technology and the hidden integration of Europe. *History and Technology*, *21*(1), 1–19. Retrieved 2022-11-29, from `http://www.tandfonline.com/doi/abs/10.1080/07341510500037487` doi: 10.1080/07341510500037487

Mobach, K., & Felt, U. (2022, July). On the Entanglement of Science and Europe at CERN: The Temporal Dynamics of a Coproductive Relationship. *Science as Culture*, *31*(3), 382–407. Retrieved 2022-10-17, from `https://www.tandfonline.com/doi/full/10.1080/09505431.2022.2076586` doi: 10.1080/09505431.2022.2076586

Nadim, T. (2016, October). Data Labours: How the Sequence Databases GenBank and EMBL-Bank Make Data. *Science as Culture*, *25*(4), 496–519. Retrieved 2021-11-18, from `https://www.tandfonline.com/doi/full/10.1080/09505431.2016.1189894` doi: 10.1080/09505431.2016.1189894

Nicolini, D. (2012). *Practice Theory, Work, and Organization: An Introduction.* OUP Oxford. (Google-Books-ID: 0lBoAgAAQBAJ)

Pickering, A. (1984). *Constructing quarks: a sociological history of particle physics.* Chicago: University of Chicago Press.

Pinel, C., Prainsack, B., & McKevitt, C. (2020, April). Caring for data: Value creation in a data-intensive research laboratory. *Social Studies of Science*, *50*(2), 175–197. Retrieved 2021-11-18, from `http://journals.sagepub.com/doi/10.1177/0306312720906567` doi: 10.1177/0306312720906567

Rao, A., Dallmeier-Tiessen, S., Lassila-Perini, K., McCauley, T., & Simko, T. (2019, July). Early Experience with Open Data from CERN's Large Hadron Collider. In *Open Innovation: Bridging Theory and Practice* (Vol. 04, pp. 227–245). WORLD SCIENTIFIC. Retrieved 2022-02-08, from `https://www.worldscientific.com/doi/abs/10.1142/9789813271647_0008` doi: 10.1142/9789813271647_0008

Rapley, T. (2007). Interviews. In C. Seale, G. Gobo, J. F. Gubrium, & S. David (Eds.), *Qualitative research practice* (pp. 15–33). London ; Thousand Oaks, Calif. :: SAGE,.

Rivas, C. (2004). Finding themes in qualitative data. In C. Seale (Ed.), *Re-*

searching Society and Culture (pp. 431–453). London: Sage. Retrieved 2021-10-20, from https://scholar.google.com/citations?view_op=view_citation&hl=de&user=EmixvvEAAAAJ&citation_for_view=EmixvvEAAAAJ:dTyEYWd-f8wC

Schatzki, T. R., Knorr-Cetina, K., & Savigny, E. v. (Eds.). (2001). *The practice turn in contemporary theory.* New York: Routledge.

Simko, T., de Bittencourt, H., Carrera, E., Lopez, D., Lange, C., Lassila-Perini, K., … Savaniakas, M. (2020). Open data provenance and reproducibility: a case study from publishing CMS open data. In C. Doglioni, D. Kim, G. Stewart, L. Silvestris, P. Jackson, & W. Kamleh (Eds.), (Vol. 245). doi: 10.1051/epjconf/202024508014

Star, S. L. (2010, September). This is Not a Boundary Object: Reflections on the Origin of a Concept. *Science, Technology, & Human Values*, *35*(5), 601–617. Retrieved 2021-09-23, from http://journals.sagepub.com/doi/10.1177/0162243910377624 doi: 10.1177/0162243910377624

Star, S. L., & Griesemer, J. R. (1989, August). Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science*, *19*(3), 387–420. Retrieved 2021-09-22, from http://journals.sagepub.com/doi/10.1177/030631289019003001 doi: 10.1177/030631289019003001

Star, S. L., & Ruhleder, K. (1996, March). Steps toward an ecology of infrastructure: design and access for large information spaces. *Revue d'anthropologie des connaissances*, *4*(1), 114–161.

Suchman, L. A. (2007). *Human-machine reconfigurations: plans and situated actions* (2nd ed.. ed.). Cambridge ; New York :, Cambridge :: Cambridge University Press,. Retrieved 2021-10-13, from http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&AN=178889

Traweek, S. (1988). *Beamtimes and lifetimes: the world of high energy physicists.* Cambridge, Mass: Harvard University Press.

Tripathee, A., Xue, W., Larkoski, A., Marzani, S., & Thaler, J. (2017, October). Jet substructure studies with CMS open data. *Physical Review D*, *96*(7), 074003. Retrieved 2021-09-09, from https://link.aps.org/doi/10.1103/PhysRevD.96.074003 doi: 10.1103/PhysRevD.96.074003

Tuertscher, P., Garud, R., Nordberg, M., & Boisot, M. (2011, July). The Concept of an ATLAS Architecture. In M. Boisot, M. Nordberg, S. Yami, & B. Nicquevert (Eds.), *Collisions and Collaboration: The Organization of Learning in the ATLAS Experiment at the LHC.* OUP Oxford. (Google-Books-ID: MXrdJrzdfy0C)

Vertesi, J. (2014, March). Seamful Spaces: Heterogeneous Infrastructures in Interaction. *Science, Technology, & Human Values*, *39*(2), 264–284. Retrieved 2022-11-25, from http://journals.sagepub.com/doi/10.1177/0162243913516012 doi: 10.1177/

0162243913516012

Waibel, D., Peetz, T., & Frank Meier. (2021, April). Valuation Constellations. *Valuation Studies*, *8*(1), 33–66. Retrieved 2022-07-22, from https://valuationstudies.liu.se/article/view/397 doi: 10.3384/VS.2001-5992.2021.8.1.33-66

Wessels, B., Finn, R. L., Wadhwa, K., & Sveinsdottir, T. (2017). *Open data and the knowledge society.* Amsterdam: Amsterdam University Press. (OCLC: ocn974978358)

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... Mons, B. (2016, March). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. Retrieved 2022-05-30, from https://www.nature.com/articles/sdata201618 (Number: 1 Publisher: Nature Publishing Group) doi: 10.1038/sdata.2016.18

Wunsch, S. (2021a). *Analysis of Higgs boson decays to four leptons using data and simulation of events at the CMS detector from 2012 using ROOT's RDataFrame.* CERN Open Data Portal. Retrieved 2021-12-21, from http://opendata.cern.ch/record/12360 doi: 10.7483/OPENDATA.CMS.F7HD.P3K4

Wunsch, S. (2021b). *Run2012B_doubleelectron dataset in reduced NanoAOD format for education and outreach.* CERN Open Data Portal. Retrieved 2021-12-21, from http://opendata.cern.ch/record/12367 (Type: dataset) doi: 10.7483/OPENDATA.CMS.YRUF.MEMI

Wunsch, S. (2021c). *Run2012B_doublemuparked dataset in reduced NanoAOD format for education and outreach.* CERN Open Data Portal. Retrieved 2021-12-21, from http://opendata.cern.ch/record/12365 (Type: dataset) doi: 10.7483/OPENDATA.CMS.04XV.ESBR

Wunsch, S. (2021d). *Run2012C_doubleelectron dataset in reduced NanoAOD format for education and outreach.* CERN Open Data Portal. Retrieved 2021-12-21, from http://opendata.cern.ch/record/12368 (Type: dataset) doi: 10.7483/OPENDATA.CMS.60ZD.OJK3

Wunsch, S. (2021e). *Run2012C_doublemuparked dataset in reduced NanoAOD format for education and outreach.* CERN Open Data Portal. Retrieved 2021-12-21, from http://opendata.cern.ch/record/12366 (Type: dataset) doi: 10.7483/OPENDATA.CMS.86ZS.7G78

Wunsch, S. (2021f). *SMHiggsToZZTo4L dataset in reduced NanoAOD format for education and outreach.* CERN Open Data Portal. Retrieved 2021-12-21, from http://opendata.cern.ch/record/12361 (Type: dataset) doi: 10.7483/OPENDATA.CMS.8FLU.UIQJ

Wunsch, S. (2021g). *ZZTo2e2mu dataset in reduced NanoAOD format for education and outreach.* CERN Open Data Portal. Retrieved 2021-12-21, from http://opendata.cern.ch/record/12364 (Type: dataset) doi: 10.7483/OPENDATA.CMS.XRWZ.IX16

Wunsch, S. (2021h). *ZZTo4e dataset in reduced NanoAOD format for education and outreach.* CERN Open Data Portal. Retrieved 2021-12-21, from `http://opendata.cern.ch/record/12363` (Type: dataset) doi: 10.7483/OPENDATA.CMS.BZ3O.K8Q5

Wunsch, S. (2021i). *ZZTo4mu dataset in reduced NanoAOD format for education and outreach.* CERN Open Data Portal. Retrieved 2021-12-21, from `http://opendata.cern.ch/record/12362` (Type: dataset) doi: 10.7483/OPENDATA.CMS.FBFX.VI9P

Appendix A

metadata element	**content**	**obligatory**	**Example 1 10.7483/OPENDATA.CMS.G13X.TDSB**	**Example 2 10.7483/OPENDATA.CMS.8FLU.IQJ**
abstract	A summary of the resource	No	Simulated dataset SMHiggsToZZTo4L_M-125_8TeV-powheg15-JHUgenV3-pythia6 in AODSIM format for 2012 collision data. See the description of the simulated dataset names in: About CMS simulated dataset names. These simulated datasets correspond to the collision data collected by the CMS experiment in 2012.	Dataset in reduced NanoAOD format for education and outreach derived from the primary dataset in AOD format linked below.
accelerator	The accelerator involved in the production of the data (e.g. CERN-LHC)	No	CERN-LHC	CERN-LHC
authors	The name of the author(s) and corresponding affiliations	No	None	Wunsch, Stefan
categories	Primary category of the simulated dataset or related asset; The secondary category of the simulated dataset of related asset; the authority who attributed the category	No		
cms_confdb_id	the cms_confdb_id (applies to CMS records)	No	not visible	not visible
collaboration	The name of the group the author is part of, The name of the collaboration the	No	CMS	CMS

	author is part of, The recid containing a collaboration's author list within COD. E.g. for the CMS collaboration, the author list is record 453 for 2012 data			
collections	The name of the collection the record belongs to (internal)	Yes	not visible	not visible
collision_information	collision energy; for instance: 7Tev	No	8Tev	none
dataset_semantics	relevant variable(s), The unit of the variable, The type of the variable, Description or explanation for a variable	No	none	Table; see 10.7483/OPEN DATA.CMS.8FLU.IQJ
date_created	The data-taking year during which the collision data or for which the simulated data, software and other assets were produced	No	2012	2012
date_published	The year of publication on the portal	Yes	2017	2021
date_reprocessed	The year the resource was reprocessed	No	none	none
distribution	The total size of the files in bytes, The total number of files attached to the record, The total number of events, The total number of entries, All the file formats included in the record (e.g. aodsim, root), specifies if dataset is on-demand or online in the future	No	299973 events. 34 files. 106.2 GB in total., the rest is not visible	299973 events. 34 files. 106.2 GB in total. the rest is not visible
doi	The Digital Object Identifier that has been registered for the resource	No	10.7483/OPENDATA.CMS.G13X.TDSB	10.7483/OPENDATA.CMS.8FLU.UIQJ
experiment	The name of the experiment the author is part of	Yes	CMS	CMS

keywords	“Relevant keywords for the record” (e.g. keywords on ODP: datascience, education, external resource, heavy-ion physics masterclass teaching)	No	Higgs Physics, Standard model	none
language	"The language of the resource, based on ISO 639-2 (e.g. 'eng' for English)"	No	not visible	not visible
license	"The license this resource is published under. Most of the content in the portal is released under the Creative Commons CC0 waiver and, if the resource is software, under another open-source license (usually GNU General Public License or Apache)"	No	Creative Commons CC0 waiver	Creative Commons CC0 waiver
links	"The URL for the link associated with this record (for an external URL)", "The title for the link", "If the link is to another CERN Open Data portal record, the record ID of that record", "A brief description of the link"	No		
methodology	"A description of the methodology used for the production of this data/software"	No	none	none
pileup	“Note about pile-up events (applies to simulated records)”	No	To make these simulated data comparable with the collision data, pile-up events are added to the simulated event in this step. The pile-up dataset is: /MinBias_TuneZ2star_8TeV-pythia6/Summer12-START50_V13-v3/GEN-SI	none

			M	
prepublication	"The report number","The name of the publisher", “the place of the publication”, "The prepublication date, based on ISO 8601 (YYYY-MM-DD)"	No	none	none
publisher	“The name of the publisher (internal)”	Yes	not visible	not visible
recid	"The record ID for this record (internal)"	Yes	not visible	not visible
relations	"The type of relation between the records (most common is 'isChildOf' for linking derived datasets to their parent)", "The title of the related record", "The internal ID of the related record, if it is another Open Data record", "The DOI assigned to the related dataset", “a Description about the related data set”	No	none	This dataset was derived from: /SMHiggsToZZTo4L_M-125_8TeV-powheg15-JHUgenV3-pythia6/Summer12_DR53X-PU_S10_START53_V19-v1/AODSIM
run_period	"The data-taking run period during which the collision data or for which the simulated data or software was produced, in a format such as 'Run2011A'"	No		none
signature	"The final state particles in a dataset (e.g. electron)"	No	none	none
source code repository	"The URL of the source code repository", “a description text for the source code repository”	No	none	none
system_deils	"The software version (e.g. CMSSW_5_3_32)", "E.g. FT_53_LV5_AN1"	No	CMSSW_5_3_32	none

title	"The title for this resource"	Yes	/SMHiggsToZZTo4L_M-125_8TeV-powheg15-JHUgenV3-pythia6/Summer12_DR53X-PU_S10_START53_V19-v1/AODSIM	SMHiggsToZZTo4L dataset in reduced NanoAOD format for education and outreach
title_additional	"A more descriptive, human-readble title"	No	Simulated dataset SMHiggsToZZTo4L_M-125_8TeV-powheg15-JHUgenV3-pythia6 in AODSIM format for 2012 collision data	none
type	"The primary category this resource belongs to (what appears in the UI facets)", "The secondary category this resource belongs to (what appears in the UI facets)"	Yes	Dataset, simulated	Dataset, derived
usage	"Instructions on how this resource can be used/accessed":	No	You can access these data through the CMS Virtual Machine. See the instructions for setting up the Virtual Machine and getting started in How to install the CMS Virtual Machine Getting started with CMS open data	none
use_with	"Information regarding other resources that can be used alongside this one"	No	none	This dataset can be used with the following analysis: Analysis of Higgs boson decays to four leptons using data and simulation of events at the CMS

				detector from 2012 using ROOT's RDataFrame
validation	"Information regarding the validation process this resource has undergone"	No	The generation and simulation of simulated Monte Carlo data has been validated through general CMS validation procedures.	These data were derived from the primary datasets linked below using only the validated runs. No further validation was done for the output. /SMHiggsToZZTo4L_M-125_8TeV-powheg15-JHUgenV3-pythia6/Summer12_DR53X-PU_S10_START53_V19-v1/AODSIM

Table 1: *Metadata scheme of CERN open data portal; from (CERN open data, 2023b)*

Publication	**Authors**	**Topic**	**Datasets used**
Quantum-inspired machine learning on high-energy physics data	Felser T., Trenti M., Sestini L. (CERN-LHC-LHCB), Gianelle A. (CERN-LHC-LHCB), Zuliani D. (CERN-LHC-LHCB), Lucchesi D. (CERN-LHC-LHCB), Montangero S.	Machine learning (Data science)	10.7483/OPENDATA.LHCB.N75T.TJPE
Analysis-Specific Fast Simulation at the LHC with Deep Learning	Chen C. (CERN-LHC-ATLAS), Cerri O., Nguyen T.Q., Vlimant J.R., Pierini M.(CERN-LHC-CMS)	Machine learning (Data science)	10.7483/OPENDATA.CMS.HBBW.LTT4

Accelerating End-to-End Deep Learning Workflow with Codesign of Data Preprocessing and Scheduling	Cheng Y. (CERN-LHC-CMS?), Li D., Guo Z., Jiang B., Geng J., Bai W., Wu J., Xiong Y.	Machine learning (Data science)	10.7483/OPENDATA.CMS.2DSE.HYDF 10.7483/OPENDATA.CMS.7RZ3.0BXP 10.7483/OPENDATA.CMS.ARKO.6NV3 10.7483/OPENDATA.CMS.REHM.JKUH 10.7483/OPENDATA.CMS.DELK.2V7R 10.7483/OPENDATA.CMS.HHCJ.TVXH. 10.7483/OPENDATA.CMS.DELK.2V7R 10.7483/OPENDATA.CMS.KAYE.XLAH
Data analysis with GPU-accelerated Kernels	Pata J. (CERN-LHC-CMS), Dutta I. (CERN-LHC-CMS), Lu N. (CERN-LHC-CMS, CERN-LHC-ATLAS), Vlimant J.R.(CERN-LHC-CMS), Newman H. (CERN-LHC-CMS), Spiropulu M. (CERN-LHC-CMS), Reissel C.(CERN-LHC-CMS), Ruini D. (CERN-LHC-CMS)	GPU acceleration (Data science)	10.7483/OPENDATA.CMS.42GY.2VJI 10.7483/OPENDATA.CMS.2DSE.HYDF 10.7483/OPENDATA.CMS.7RZ3.0BXP 10.7483/OPENDATA.CMS.ARKO.6NV3 10.7483/OPENDATA.CMS.REHM.JKUH 10.7483/OPENDATA.CMS.DELK.2V7R 10.7483/OPENDATA.CMS.HHCJ.TVXH 10.7483/OPENDATA.CMS.DELK.2V7R 10.7483/OPENDATA.CMS.KAYE.XLAH
Adversarially Learned Anomaly Detection on CMS open data: re-discovering the top quark	Knapp O., Cerri O.(CERN-LHC-CMS), Dissertori G. (CERN-LHC-CMS) , Nguyen T.Q.(CERN-LHC-CMS), Pierini M.(CERN-LHC-CMS), Vlimant J.R. (CERN-LHC-CMS)	Machine Learning (Data science)	10.7483/OPENDATA.CMS.IYVQ.1J0W 10.7483/OPENDATA.CMS.REHM.JKUH 10.7483/OPENDATA.CMS.DELK.2V7R 10.7483/OPENDATA.CMS.HHCJ.TVXH 10.7483/OPENDATA.CMS.TCAX.E3IO

			10.7483/OPENDATA.CMS.5ABQ.6SIS 10.7483/OPENDATA.CMS.RYNC.1VIB 10.7483/OPENDATA.CMS.5XRD.X0BY 10.7483/OPENDATA.CMS.XCTC.0OXC
End-to-End Physics Event Classification with CMS Open Data: Applying Image-Based Deep Learning to Detector Data for the Direct Classification of Collision Events at the LHC	Andrews M. (CERN-LHC-CMS), Paulini M. (CERN-LHC-CMS) , Gleyzer S. (CERN-LHC-ATLAS, CERN-LHC-CMS), Poczos B.	Machine learning (data science)	10.7483/OPENDATA.CMS.WQ7P.BZP3 10.7483/OPENDATA.CMS.WV7J.8GN0 10.7483/OPENDATA.CMS.2W51.W8AT
End-to-end jet classification of quarks and gluons with the CMS Open Data	Andrews M .(CERN-LHC-CMS), Alison J. (CERN-LHC-ATLAS, CERN-LHC-CMS), An S. (CERN-LHC-CMS), Patrick Bryant (CERN-LHC-ATLAS, CERN-LHC-CMS), Burkle B. (CERN-LHC-CMS), Gleyzer S. (CERN-LHC-ATLAS, CERN-LHC-CMS), Narain M. (CERN-LHC-CMS) , Paulini M.(CERN-LHC-CMS), Poczos B., Usai E. (CERN-LHC-CMS)	Machine learning (data science)	10.7483/OPENDATA.CMS.2W51.W8AT

Probing resonance states in high-energy interaction: a novel approach using complex network technique based on symmetry scaling	Bhaduri S., Bhaduri A., Ghosh D.	Data science	10.7483/OPENDATA.ALICE.Y4KJ.8HZC;
			10.7483/OPENDATA.CMS.FZ5U.TTXP;
			10.7483/OPENDATA.CMS.ZCFQ.Q557;

Interaction networks for the identification of boosted H →b b ¯ decays interaction networks for the identification of	Moreno E.A., Nguyen T.Q. (CERN-LHC-CMS), Vlimant J.-R. (CERN-LHC-CMS), Cerri O. (CERN-LHC-CMS), Newman H.B.(CERN-LHC-CMS), Periwal A.(CERN-LHC-CMS), Spiropulu M. (CERN-LHC-CMS), Duarte J.M., Pierini M. (CERN-LHC-CMS)	Data science	10.7483/OPENDATA.CMS.JGJX.MS7Q
Opportunities and challenges of Standard Model production cross section measurements in proton-proton collisions at s=8 TeV using CMS Open Data	Apyan A. (CERN-LHC-ATLAS, CERN-LHC-CMS), Cuozzo W., Klute M. (CERN-LHC-CMS), Saito Y.(), Schott M. (CERN-LHC-ATLAS), Sintayehu B.	cross section measurement (High energy physics)	10.7483/OPENDATA.CMS.9A4E.7SIR 10.7483/OPENDATA.CMS.IYVQ.1J0W 10.7483/OPENDATA.CMS.BAKP.W6TP 10.7483/OPENDATA.CMS.8XN1.J5N7
Testing non-standard sources of parity violation in jets at the LHC, trialled with CMS Open Data	Lester C.G. (CERN-LHC-ATLAS), Schott M. (CERN-LHC-ATLAS)	analysis of particle jets (high energy physics)	10.7483/OPENDATA.CMS.VZSR.LYZX; 10.7483/OPENDATA.CMS.7Y4S.93A0; 10.7483/OPENDATA.CMS.P2XT.ZX19, 10.7483/OPENDATA.CMS.7RZ3.0BXP; 10.7483/OPENDATA.CMS.FZCE.MBDW; 10.7483/OPENDATA.CMS.B91N.86OR, 10.7483/OPENDATA.CMS.V2C6.O1P4, 10.7483/OPENDATA.CMS.71R9.VLZA, 10.7483/OPENDATA.CMS.QGC3.PTZ9,
Symmetry-Scaling Based Complex Network Approach to Explore Exotic	Bhaduri S., Bhaduri A., Ghosh D.	Data science	10.7483/OPENDATA.ALICE.Y4KJ.8HZC

Hadronic States in High-Energy Collision			10.7483/OPENDATA.CMS.ZCFQ.Q557
Explicit jet veto as a tool to purify the underlying event in the Drell-Yan process using CMS Open Data	Mehdiabadi S.P. (CERN-LHC-CMS), Fahim A. (CERN-LHC-CMS)	particle jets (High energy physics)	10.7483/OPENDATA.CMS.TXT4.4RRP 10.7483/OPENDATA.CMS.RZ34.QR6N
Searching in CMS open data for dimuon resonances with substantial transverse momentum	Cesarotti C., Soreq Y (CERN- theory department)., Strassler M.J., Thaler J., Xue W. (CERN-theory department?)	Theoretical high energy physics	10.7483/OPENDATA.CMS.RZ34.QR6N 10.7483/OPENDATA.CMS.TXT4.4RRP 10.7483/OPENDATA.CMS.T8RZ.D52D 10.7483/OPENDATA.CMS.U6JT.SMMC 10.7483/OPENDATA.CMS.4G5S.44WQ 10.7483/OPENDATA.CMS.D5KD.U5MR 10.7483/OPENDATA.CMS.UUG7.4NHT

End-to-end particle and event identification at the Large Hadron Collider with CMS open data	Andrews M. (CERN-LHC-CMS), Alison J. (CERN-LHC-ATLAS, CERN-LHC-CMS), An S. (CERN-LHC-CMS) , Bryant P., Burkle B. (CERN-LHC-CMS), Gleyzer S. (CERN-LHC-ATLAS), Narain M. (CERN-LHC-CMS), Paulini M.(CERN-LHC-CMS) , Poczos B., Usai E. (CERN-LHC-CMS)	Machine learning (data science)	10.7483/OPENDATA.CMS.Q3BX.69VQ,
			10.7483/OPENDATA.CMS.84VC.RU8W;
			10.7483/OPENDATA.CMS.PUTE.7H2H;
			10.7483/OPENDATA.CMS.QJND.HA88;
			10.7483/OPENDATA.CMS.WKRR.DCJP;
			10.7483/OPENDATA.CMS.X3XQ.USQR;
			10.7483/OPENDATA.CMS.BKTD.SGJX;
			10.7483/OPENDATA.CMS.S3D5.KF2C;
			10.7483/OPENDATA.CMS.96U2.3YAH;
			10.7483/OPENDATA.CMS.RC9V.B5KX;
			10.7483/OPENDATA.CMS.CX2X.J3KW;

Fast and Accurate Simulation of Particle	Musella P. (CERN-LHC-CMS),	Machine learning	10.7483\/OPENDATA.CMS.Q3BX.69VQ

Detectors Using Generative Adversarial Networks	Pandolfi F. (CERN-LHC-CMS)		10.7483\/OPENDATA.CMS.84VC.RU8W 10.7483\/OPENDATA.CMS.PUTE.7H2H 10.7483\/OPENDATA.CMS.QJND.HA88 10.7483\/OPENDATA.CMS.WKRR.DCJP 10.7483\/OPENDATA.CMS.X3XQ.USQR 10.7483\/OPENDATA.CMS.BKTD.SGJX 10.7483\/OPENDATA.CMS.EJT7.KSAY 10.7483\/OPENDATA.CMS.S3D5.KF2C 10.7483\/OPENDATA.CMS.96U2.3YAH 10.7483\/OPENDATA.CMS.RC9V.B5KX 10.7483\/OPENDATA.CMS.CX2X.J3KW
Jet substructure studies with CMS open data	Tripathee A., Xue W., Larkoski A., Marzani S., Thaler J.	High energy physics/ Machine learning	10.7483/OPENDATA.CMS.3S7F.2E9W, 10.7483/OPENDATA.CMS.UP77.P6PQ 10.7483/OPENDATA.CMS.6BPY.XFRQ
Exposing the QCD Splitting Function with CMS Open Data	Larkoski A., Marzani S., Thaler J., Tripathee A., Xue W.	High energy physics/ Machine learning	10.7483/OPENDATA.CMS.3S7F.2E9W
Pion Fluctuation Study in Pb–Pb Collision at 2.76 TeV per Nucleon Pair from ALICE Experiment with Chaos and Complex Network-Based Methods	Bhaduri S., Bhaduri A., Ghosh D.	Data science	10.7483/OPENDATA.ALICE.Y62S.E7UR

Table 2: Publications using CERN open data