



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA



Ph.D. Dissertation - Tesis doctoral

**A causal model to explain data reuse in science:
a study in health disciplines**

by

M^a Inmaculada Aleixos Borrás (author)

Valencia (Spain), March 2020

PhD supervisors:

Pablo D'Este
María Fernanda Peset Mancebo
Richard Woolley

Als meus. Als qui estàn i als qui han marxat.
A qui no ha deixat d'anirmar-me mai.
A qui ha activat la meua *susceptibilitat* (Sayer, 2000, 2010) de ser valenta i creativa per a teoritzar.

A los míos. A los que están y a los que se han ido.
A quien no ha dejado de animarme nunca.
A quien ha activado mi *susceptibilidad* (Sayer, 2000, 2010) de ser valiente y creativa para teorizar.

To mine. To those who are, and those who have left.
To who has never stopped encouraging me.
To who has activated my *liability* (Sayer, 2000, 2010) of being brave and creative to theorize.

Acknowledgments

Agraïments – Agradecimientos - Remerciements

First, I want to thank all my participants, who, in order of appearance in this dissertation, are David Cook (PhD candidate), Joan Ballester Climent (PhD), Jaume Forés Martos (PhD), Deshayne Fell (PhD), Mary Smith (PhD), Sarah Wilson (PhD), Nicole Langlois (MSc), Claire Johnson (PhD), and Areti Angeliki Veroniki (PhD). Thanks for all your time answering my questions during the interviews and for responding my emails. Without you, this dissertation would not exist.

Second, I want to thanks my three PhD advisors for their intellectual and emotional support. Their different backgrounds have enriched my learning process, and contributed to the completion of this dissertation in varied ways. Choices and imperfections in this dissertation are my own. Thanks also to Ian Graham (PhD) and Sylvie Grosjean (PhD) for being my host mentors with the ERASMUS MUNDUS NOVADOMUS CHEMEDPHO scholarship (2016-2017), which funded me in order to conduct most of my empirical research in Canada. Thanks to Amparo Cortés Lucas (PhD), who was the coordinator of the NOVADOMUS CHEMEDPHO for all her support during my application, and my stay in Canada. Thanks to Ángela Villena Aleixos, for her patience and hard work in drawing with AUTOCAD all the versions of some images included in this dissertation and some other images, which I have finally decided to leave out.

Thanks to all assessors and jury members who have accepted pleasingly to assess this dissertation, as either principals or substitutes. I am sure that they will do a very good job providing me with thoughtful feedback. Thanks to all administrative staff (Mila, Beatriz, Rosa) of the *Departamento*

de Comunicaci3n Audiovisual, Documentaci3n e Historia del Arte (UPV) for all their administrative support of all bureaucratic issues for the submission and defense of my dissertation.

I am also grateful to very generous people, who I have met during my PhD and have contributed in varied aspects along the dissertation journey. I was very fortunate to find all of them on my way. They are my colleagues at INGENIO, especially –in no especial order– Isabel Piqueras, Ester Planells Aleixandre, Marisa Rodr3guez, Juanjo P3rez, Antonio Guti3rrez (PhD), Julia Olmos Peñuela (PhD), Elena Mas Tur (PhD), Michelle Bandeira (PhD), Teresa Escrich, Javier Ortega Colomer (PhD), Carmen Corona Sobrino (PhD), Vicky Pellicer (PhD), Ismael Ràfols (PhD), Jordi Molas Gallart (PhD), Franois Perruchas (PhD), Carlos Benito Amat (PhD), Elena Castro Mart3nez (PhD), Adri3n Arias Di3z-Faes (PhD), Mar3a Luz L3pez Terrada (PhD), and J. F3lix Lozano Aguilar (PhD).

I was also very fortunate to meet very generous researchers or professionals during my dissertation journey from other institutions. So, in no special order, thanks to Grit Laudel (PhD), Jochen Gl3sser (PhD), Taran Mari Thune (PhD), C. Scott Findlay (PhD), Sandra Dunn (PhD), Daniel Pare (PhD), Jaya S. Peruvemba (PhD), Doug Coyle (PhD), Kednapa Thavorn (PhD), Jordi Pardo Pardo, David Moffete (PhD), Jessica McEwan, Susan Morrison, Kelly Cobey (PhD), Michael McBurney (PhD), Vicente Mart3nez Tur (PhD), Francisco Gracia (PhD), and Ram3n Mart3nez M3ñez (PhD).

A dissertation, although a challenging task, is merely one aspect of my life. Therefore, I want to thank some people for being part of other aspects of my life while doing this dissertation. In no special order, 3ngel Villena Garc3a, Quique Lanuza (PhD), Carmen Picazo (PhD), Lucile Danvier (PhD), Gabriela Perdomo P3ez (PhD candidate), Mariana Dominguez (PhD candidate), Isaac Nahon Sefarty (PhD), Salvio Digesto (PhD), Mich3le Wilson, Hermann Franke, Andr3s di Masso Tarditti (PhD), Carles Xavier Sim3 Noguera (PhD), Ram3n V. Cirilo (PhD), Marina Canellas Fontanilles (PhD), Fco. Javier P3rez Areales (PhD), Marina Romeo Delgado (PhD), Nina Baranowska, Ricardo Hern3ndez Sandoval, Maur3cio Reis Nothen (PhD candidate), Laura Garc3a Vargas (PhD candidate), Katrina Pondemira, Karen Paterson, Franois Lortie (PhD), Silvia, Maeve Moriarty, Matthew Wallace (PhD), Pablo Gil Anaya (PhD), Carlos P3rez L3pez, and Mar3a Mart3nez Ant3n (PhD).

Thanks also to Lucas and Mab. Where is Gamin?

Abstract

Investments in data infrastructures, data management, data repositories, and *Open Data* sharing policies and recommendations are viewed as increasingly important for scientific knowledge production. One of the underlying assumptions justifying these investments is that the more available *Open Data* becomes, then the greater the possibilities for creating new knowledge that can advance both science and human wellbeing. Yet efforts and investments in *Open Data* and other ways of data sharing only have value if data are actually reused. Recent scholarly efforts have brought forth some of the challenges and facilitators related to the reuse of data, in order to inform current and future policies and investments. However, despite these efforts, we still do not know why and how some researchers are successful in reusing data, despite the challenges they face, and why some researchers abandon the process of reusing data when facing such challenges. This dissertation aims to fill this gap by focusing on a causal explanation of the data reuse process, which it understands as being nested in broader patterns of researchers' motivations, scientific goals and decision-making strategies.

The dissertation is comprised of three main elements. First, it proposes a heuristic model of the scientific actor, the *bounded individual horizon* (BIH) model, which understands that, on the one hand, researchers' work and careers are structured by their motivation to produce scientific contributions and rewards systems that prioritizes certain types of contributions. On the other hand, researchers' struggles to achieve their objective of creating new findings that accrue recognition and rewards occur within a frame of limited information and resources, conditioned by multiple institutional, social, and other factors. Second, the study proposes a mechanistic causal theoretical explanation that enables us to understand the data reuse process and its effects (outcomes). The *data-reuse mechanism* as it is called, enables us to understand how the satisficing behavior that characterizes scientific decision-making applies to the specific conditions and processes of data reuse. Third, a set of ten empirical case studies of data reuse in health research were conducted and are reported in the dissertation. These cases are analyzed and interpreted using the complementary theoretical lenses of the bounded individual horizon and the data-reuse mechanism approaches.

The main findings explain that there is an apparent association between the extent and types of efforts required to reuse data, researchers' contextualized motivations, and broader goal-setting and decision-making frames. Access to data is a necessary condition for the reuse of data, yet is not sufficient for the reuse to happen. Characteristics of available data, including the context of their production, the extent of the preparation and stewarding of these data and their potential value in relation to researchers' motivations to make new scientific claims or generate background knowledge are found to be essential elements for understanding why some data reuse processes persist and succeed, while others do not. The thesis concludes that efforts and investments designed to reap the benefits of data reuse should also be expanded to include training researchers in data reuse, including to efficiently recognize opportunities, navigate the challenges of the reuse process, and be aware of and acknowledge the limitations of the use of secondary data. Without such investments, the promises and expectations linked to emerging data infrastructures, data repositories, data management guidelines and open science practices are argued to be far less likely to reach their full potential.

Resum

Les inversions en infraestructures de dades, gestió de dades, repositoris de dades i polítiques i recomanacions d'intercanvi de Dades Obertes (Open Data) es consideren cada vegada més importants per a la producció del coneixement científic. Un dels supòsits subjacents que justifiquen aquestes inversions és que com més disponibles siguen les Dades Obertes, majors seran les possibilitats de crear nou coneixement que pugui fer avançar tant la ciència com el benestar humà. No obstant això, els esforços i les inversions en les Dades Obertes i altres maneres de compartir dades només tenen valor si les dades es reutilitzen realment. Recents investigacions acadèmiques han posat de manifest alguns dels reptes i dels factors facilitadors relacionats amb la reutilització de les dades, a fi d'informar les polítiques i inversions actuals i futures. No obstant això, encara desconeixem per què i com alguns/es investigador(e)s aconsegueixen reutilitzar les dades, malgrat els reptes als quals s'enfronten, i per què altres investigador(e)s abandonen el procés de reutilització de les dades quan s'enfronten a aquests reptes. La present tesi té com a objectiu omplir aquest buit centrant-se en una explicació causal del procés de reutilització de dades, que s'entén que està associada amb pautes més àmplies derivades de les motivacions, els objectius científics i les estratègies de presa de decisions d'els/les investigador(e)s.

La tesi consta de tres elements principals. En primer lloc, proposa un model heurístic de l'actor científic, el model de l'horitzó individual delimitat (BIH pel nom anglès, bounded individual horizon), que entén que, d'una banda, el treball i la carrera d'els/les investigador(e)s s'estructuren en funció de la seua motivació per a produir contribucions científiques i dels sistemes de recompensa que prioritzen determinats tipus de contribucions. D'altra banda, els esforços d'els/les investigador(e)s per aconseguir el seu objectiu d'obtenir nous resultats que acumulin reconeixement i recompenses es produeixen en un marc d'informació i recursos limitats, condicionats per múltiples factors institucionals, socials i d'altra índole. En segon lloc, aquesta tesi proposa una explicació teòrica causal mecanicista que permet comprendre el procés de reutilització de les dades i els seus efectes (resultats). El mecanisme de reutilització de dades (data-reuse mechanism), com es denomina, ens permet comprendre com el comportament satisfactori que caracteritza la presa de decisions científiques s'aplica a les condicions i processos específics de reutilització de dades. En tercer lloc, aquesta tesi inclou l'estudi empíric d'un

conjunt de deu estudis de casos de reutilització de dades en ciències de la salut, així com també els resultats d'aquest estudi.. Aquests casos s'han analitzat i interpretat utilitzant les lents teòriques complementàries de l'horitzó individual delimitat i els enfocaments del mecanisme de reutilització de dades.

Les principals conclusions expliquen que existeix una aparent associació entre l'abast i els tipus d'esforços necessaris per a reutilitzar dades, les motivacions contextualitzades d'els/les investigador(e)s i els marcs més amplis de fixació d'objectius i presa de decisions. L'accés a les dades és una condició necessària per a la seua reutilització, però no és suficient perquè aquesta es produeixi. Es considera que les característiques de les dades disponibles, inclòs el context de la seua producció, el grau de preparació i administració d'aquestes dades i el seu potencial valor en relació amb les motivacions d'els/les investigador(e)s per a fer noves afirmacions científiques o generar coneixements de base, són elements essencials per a comprendre per què alguns processos de reutilització de dades persisteixen i tenen èxit, mentre que uns altres no. Aquest estudi conclou que els esforços i inversions destinats a aprofitar els beneficis de la reutilització de dades també haurien d'ampliar-se per a incloure la capacitació d'els/les investigador(e)s en matèria de reutilització de dades, en particular per a reconèixer eficientment les oportunitats, superar els problemes del procés de reutilització i ser conscients i reconèixer les limitacions de la reutilització de dades secundàries. Sense aquests esforços i inversions, les promeses i expectatives vinculades a les infraestructures, repositoris i directrius de gestió de dades i les pràctiques científiques obertes tenen moltes menys probabilitats d'aconseguir el seu ple potencial.

Resumen

Las inversiones en infraestructuras de datos, gestión de datos, repositorios de datos y políticas y recomendaciones de intercambio de Datos Abiertos (Open Data) se consideran cada vez más importantes para la producción del conocimiento científico. Una de las razones que justifica estas inversiones es que cuanto más Datos Abiertos haya, mayores serán las posibilidades de crear nuevo conocimiento que pueda hacer avanzar tanto la ciencia como el bienestar humano. Sin embargo, los esfuerzos y la inversión en Datos Abiertos y otras formas de compartirlos sólo tienen valor si se reutilizan realmente. Recientes trabajos académicos han puesto de manifiesto algunos de los retos y factores facilitadores relacionados con la reutilización de los datos, a fin de asesorar las políticas e inversiones actuales y futuras. Sin embargo, a pesar de esos esfuerzos, todavía desconocemos por qué y cómo algunos/as investigadores/as logran reutilizar los datos, a pesar de los retos a los que enfrentan, y por qué otros/as investigadores/as abandonan el proceso de reutilización de los datos. La presente tesis tiene por objeto llenar este vacío centrándose en una explicación causal del proceso de reutilización de los datos, que se entiende está inmersa en pautas de conducta más amplias que se relacionan con las motivaciones, los objetivos científicos y las estrategias de toma de decisiones de los/as investigadores/as.

Este estudio consta de tres elementos principales. En primer lugar, propone un modelo heurístico del actor científico, el *modelo del horizonte individual delimitado* (BIH por su nombre en inglés, *bounded individual horizon*). En él se entiende que, por una parte, el trabajo y la carrera de los/as investigadores/as se estructuran en función de su motivación para producir contribuciones científicas y de los sistemas de recompensa que dan prioridad a determinados tipos de contribuciones. Por otra parte, los esfuerzos de los/as investigadores/as para lograr su objetivo de crear nuevos hallazgos que acumulen reconocimiento y recompensas se producen en un marco de información y recursos limitados, condicionados por múltiples factores institucionales, sociales y de otra índole. En segundo lugar, esta tesis propone una explicación teórica causal mecanicista que permite comprender el proceso de reutilización de los datos y sus efectos (resultados). El mecanismo de reutilización de datos (*data-reuse mechanism*), como se denomina, nos permite comprender cómo la toma de decisiones científicas está caracterizada por una conducta que tiende a satisfacer esos objetivos en unas condiciones y

procesos específicos de reutilización de datos. En tercer lugar, este estudio incluye los resultados del estudio empírico de diez estudios de casos de reutilización de datos en ciencias de la salud. Estos casos se han analizado e interpretado utilizando el modelo teórico del horizonte individual delimitado y los enfoques del mecanismo de reutilización de datos.

Los resultados principales explican que existe una aparente asociación entre el alcance el alcance y tipo de esfuerzo requerido para reutilizar datos, las motivaciones contextualizadas de los/as investigadores/as y marcos más amplios de fijación de objetivos y toma de decisiones. El acceso a los datos es una condición necesaria para su reutilización, pero no es suficiente para que ésta se produzca. Para comprender por qué algunos procesos de reutilización de datos persisten y tienen éxito, mientras que otros no, son elementos esenciales: las características de los datos disponibles, incluido el contexto de su producción; el grado de preparación y administración de esos datos; y su potencial valor en relación con las motivaciones de los investigadores para hacer nuevas afirmaciones científicas o generar conocimientos de base. Este estudio concluye que los esfuerzos e inversiones destinados a aprovechar los beneficios de la reutilización de los datos también deberían ampliarse para incluir la capacitación de los/as investigadores/as en materia de reutilización de datos. En particular, debe insistirse en la capacidad para reconocer eficientemente las oportunidades, sortear los problemas del proceso de reutilización y ser conscientes y reconocer las limitaciones de la utilización de datos secundarios. Sin estas inversiones, las promesas y expectativas vinculadas a las emergentes infraestructuras de datos, los repositorios de datos, las directrices de gestión de datos y las prácticas científicas abiertas tienen muchas menos probabilidades de alcanzar su pleno potencial.

Contents

Acknowledgments – Agraïments – Agradecimientos – Remerciements	iv
Abstract.....	vi
Resum.....	viii
Resumen	x
Contents.....	xiii
Chapter 1. Introduction	1
Chapter 2. Literature review.....	6
Chapter 3. Conceptual framework and theoretical strategy	25
3.1. Definition of <i>data</i> , <i>primary data</i> , <i>secondary data</i> and <i>reuse of data</i>	25
3.2. Rational choice theory, bounded rationality and procedural rationality	31
3.3. A model of the scientific actor’s behavior and decision-making: the <i>bounded individual horizon</i> (BIH) model	38
Chapter 4. Methodology and methods	42
4.1. Research questions and an approach to answer them.....	44
4.2. Researchers’ decisions based on the BIH model when reusing data: the data-reuse mechanism.....	46
The researcher’s structure and causal powers and liabilities.....	51
Condition C1 – The researcher knows that secondary data exist.....	52
Condition C2 – Secondary data are obtained	53
Condition C3 – Particular secondary data are an initial satisficing option	53
Condition C4 – The idea of collecting particular primary data is not an initial satisficing option.....	54

Condition C5 – An expected scientific contribution exists and the researcher finds its potential rewards initially satisficing	54
Potential events of the data-reuse mechanism	55
4.3. Justification of a multi-case study approach	59
4.4. The search process of case studies	65
4.5. Description of data collecting methods and instruments.....	67
The interview	68
Authors’ publications and their complementary role	69
Visual representation of each participant’s data reuse process	70
Follow-up messages	71
4.6. Data analysis methods.....	71
Within-case analysis.....	73
Cross-case analysis.....	74
4.7. A diachronic process of data collection and data analysis	74
4.8. Ethics protocols and data sharing agreements.....	79
4.9. A small exercise on reflexivity.....	79
Chapter 5. General overview of cases and data sources collected for each case	82
5.1. Case studies reusing <i>released data</i>	85
5.1.1. Collected empirical data and collection dates in case study #1 (GTEX data repository)	88
5.1.2. Collected empirical data and collection dates in case study #2 (GEO Profiles repository)	89
5.1.3. Collected empirical data and collection dates in case study #3 (TCGA data repository)	90
5.1.4. Collected empirical data and collection dates in case study #4 (GEO Profiles and TCGA repositories).....	91
5.2. Case studies reusing <i>stewarded data</i>	92
5.2.1. Collected empirical data and collection dates in case study #5 (BORN Ontario data & ICES data)	95
5.2.2. Collected empirical data and collection dates in case study #6 (BORN Ontario data)	96
5.2.3. Collected empirical data and collection dates in case study #7 (BORN Ontario data)	97
5.3. Case studies reusing <i>proprietary data</i>	97
5.3.1. Collected empirical data and collection dates in case study #8 (IPD MA)	100
5.3.2. Collected empirical data and collection dates in case study #9 (IPD NMA)	101
5.3.3. Collected empirical data and collection dates in case study #10 (IPD NMA)	102

Chapter 6. Empirical analysis: testing the data-reuse mechanism	103
6.1. Case studies reusing <i>released data</i>	105
6.1.1. Case study #1 (GTEx data repository)	
6.1.2. Case study #2 (GEO Profiles repository).....	
6.1.3. Case study #3 (TCGA data repository).....	
6.1.4. Case study #4 (GEO Profiles and TCGA repositories).....	
6.2. Case studies reusing <i>stewarded data</i>	129
6.2.1. Case study #5 (BORN Ontario data and ICES data).....	
6.2.2. Case study #6 (BORN Ontario data).....	
6.2.3. Case study #7 (BORN Ontario data).....	
6.3. Case studies reusing <i>proprietary data</i>	155
6.3.1. Case study #8 (IPD MA).....	
6.3.2. Case study #9 (IPD NMA).....	
6.3.3. Case study #10 (IPD NMA).....	
6.4. Summary of analysis of the ten case studies	176
6.4.1. When condition C4 of the data-reuse mechanism is met	
6.4.2. When condition C4 of the data-reuse mechanism is not met.....	
6.5. How does the data-reuse mechanism work?	182
6.5.1. When are secondary data used as evidence of scientific claims (outcome c or #3)? ..	
.....	
6.5.2. When some conditions of the data-reuse mechanism are partially met	
6.5.3. When are secondary data used for the creation of background knowledge (outcome	
b or #2)?.....	
6.5.4. When are secondary data not finally used (outcome a or #1)?.....	
6.5.4.1. After having tried using secondary data.....	187
6.5.4.2. The use of secondary data does not happen at all	188
Chapter 7. Findings	190
Chapter 8. Discussion: contributions, limitations, and opportunities for further research	195
Chapter 9. Conclusion.....	199
Bibliography.....	201

Appendix.....	216
Annex 1. Interview script vs3 IAB Nov 2016.....	217
Annex 2. Data collection instruments and dates. Case study #1	222
Annex 3. Data collection instruments and dates. Case study #2.....	223
Annex 4. Data collection instruments and dates. Case study #3.....	224
Annex 5. Data collection instruments and dates. Case study #4.....	225
Annex 6. Data collection instruments and dates. Case study #5.....	226
Annex 7. Data collection instruments and dates. Case study #6.....	227
Annex 8. Data collection instruments and dates. Case study #7.....	228
Annex 9. Data collection instruments and dates. Case study #8.....	229
Annex 10. Data collection instruments and dates. Case study #9.....	230
Annex 11. Data collection instruments and dates. Case study #10.....	231
Annex 12. Workflow diagram of data reuse process. Case study #1	232
Annex 13. Workflow diagram of data reuse process. Case study #2.....	233
Annex 14. Workflow diagram of data reuse process. Case study #4.....	234
Annex 15. Situating the reuse of data within a larger research enquiry. Case study #4	235
Annex 16. Workflow diagram of data reuse process. Case study #5.....	236
Annex 17. Workflow diagram of data reuse process. Case study #6.....	237
Annex 18. Situating the reuse of data within a larger research project. Case study #6.....	238
Annex 19. Workflow diagram of data reuse process. Case study #7.....	239
Annex 20. Workflow diagram of data reuse process. Case study #8.....	240
Annex 21. Workflow diagram of data reuse process. Case study #9.....	241
Annex 22. Workflow diagram of data reuse process. Case study #10.....	242
Annex 23. Structure of the data-reuse mechanism.....	243
Annex 24. Literature about factors affecting the reuse of data by IS scholars.....	244
Annex 25. Process-tracing of the data-reuse mechanism when data are reused as the only evidence of scientific claims	250
Annex 26. Process-tracing of the data-reuse mechanism when secondary data are used to support scientific claims done with primary data.....	251
Annex 27. Process-tracing of the data-reuse mechanism when condition C5 is not met in time 2.....	252

List of Tables

Table 1 - Two possible initial combinations A and B of conditions of the data-reuse mechanism, and their respective hypothesized outcomes.....	58
Table 2 - Key of the images and symbols used in the visual representation of data collection instruments and dates.....	84
Table 3 - Variability of case studies reusing "released data"	87
Table 4 - Variability of case studies reusing "stewarded data"	94
Table 5 - Variability of case studies reusing "proprietary data"	99
Table 6 - Summary of outcome and conditions of findings in case study #6	176
Table 7 - Summary of outcomes and conditions of case studies #4 and #5.....	176
Table 8 - Summary of outcome and conditions of case studies #8, #9 and #10	177
Table 9 - Summary of outcome and conditions of case study #7	178
Table 10 - Summary of outcome and conditions of case study #1	179
Table 11 - Summary of outcome and conditions of case studies #2 and #3	180
Table 12 - Summary of analysis of the ten case studies	181
Table 13 - Several combination of conditions C3, C4 and C5 that do not lead to the reuse of data.....	189

List of Figures

Figure 1 - Definition of data in this dissertation	29
Figure 2 – Visual representation of the definition of reuse of data or use of secondary data	30
Figure 3 - Source: Figure 7 – The structures of causal explanation (Sayer, 2010, p. 74)	47
Figure 4 - The data-reuse mechanism and its structure and potential events.....	57
Figure 5 - Four case studies of reuse of “released data”. Same researcher in case studies #1, #2, #3 reuses data from three different repositories (A, B, C). Researcher in case study #4 reuses data from repositories B and C.....	86
Figure 6 - Data collection instruments and dates. Case study #1	88
Figure 7 - Data collection instruments and dates. Case study #2.....	89
Figure 8 - Data collection instruments and dates. Case study #3.....	90
Figure 9 - Data collection instruments and dates. Case study #4.....	91
Figure 10 - Three case studies of reuse of "stewarded data". Three different researchers (case studies #5, #6, #7) reuse data from the same repository (BORN Ontario). One researcher (case study #5) reuses also data from ICES repository.....	93
Figure 11 - Data collection instruments and dates. Case study #5.....	95
Figure 12 - Data collection instruments and dates. Case study #6.....	96
Figure 13 - Data collection instruments and dates. Case study #7.....	97
Figure 14 – Three cases studies of reuse of “proprietary data”. Three different researchers, rather a research team, (case studies #8, #9, #10) reuse individual participant data (IPD) from different data sets in three different health problems	98
Figure 15 - Data collection instruments and dates. Case study #8.....	100
Figure 16 - Data collection instruments and dates. Case study #9.....	101
Figure 17 - Data collection instruments and dates. Case study #10.....	102
Figure 18 - Workflow diagram of the data reuse process of case study #1	107
Figure 19 - Workflow diagram of the data reuse process of case study #3	115

Figure 20 - Workflow diagram of the data reuse process of case study #4 123

Figure 21 - Situating the reuse of data within a larger research inquiry. Case study #4 125

Figure 22 - Dimensions, elements, and sub-elements of BORN’s Data Quality Framework. Source: BORN’s DQF 130

Figure 23 - Workflow diagram of the data reuse process of case study #5 134

Figure 24 - Situating the reuse of data within a larger research project. Case study #6 142

Figure 25 - Workflow diagram of the data reuse process of case study #6 143

Figure 26 - Workflow diagram of the data reuse process of case study #7 150

Figure 27 - Workflow diagram of the data reuse process of case study #8 157

Figure 28 - Workflow diagram of the data reuse process of case study #9 164

Figure 29 - Number of eligible studies and participants. Case study #9. Source: (Welch et al., 2019, p. 12)..... 166

Figure 30 - Workflow diagram of the data reuse process of case study #10 171

Figure 31 – Process-tracing of the data-reuse mechanism when data are reused as the only evidence of scientific claims 183

Figure 32 - Process-tracing of the data-reuse mechanism when secondary data are used to support scientific claims done with primary data. 183

Figure 33 - Process-tracing of the data-reuse mechanism when condition C5 is not met in time 2 .186

Figure 34 - Association between high effort and the goal of making a scientific contribution in the ten case studies..... 192

Chapter 1

Introduction

Social scientists tend too easily to assume that the sociopolitical importance of an object is in itself sufficient warrant for the importance of the discourse that addresses it.

(Bourdieu & Wacquant, 1992, p. 220)

Two contextual events are of a high sociopolitical relevance nowadays. On the one hand, *we are entering into the dataverse* (Bowker, 2013) or into the (big) data society (Kitchin, 2014). Data¹ are currently regarded as the “new oil²” of the economy in all aspects of our lives (Lopez de Vallejo, Scerri, & Tuikka, 2019). On the other hand, there have been some confluent legal, social, technological and economic factors at the end of the 20th century, which have introduced a new way to share data openly, which stems from ideas of *freedom of information, accountability, transparency, and openness*, especially in government issues and in public institutions. The practice of science has not

¹ In this dissertation, I use the plural form of verb tenses for *data*, since the term *data* is the plural form for *datum* (singular).

² https://ec.europa.eu/commission/presscorner/detail/en/SPEECH_13_261 ; <https://www.youtube.com/watch?v=9Jq4Qy1UeAE> [2 February 2020]

escaped from these ideas³, which have crystalized in the concept of *Open (research) Data*⁴, among other concepts. Consequently, researchers are encouraged, and sometimes required in some disciplines, to share their data in data repositories or in thematic data infrastructures to, ultimately, make their data publicly available for any prospective user.

The term Open Data is older than *Open Access* to publications. However, it is still difficult to find an unambiguous authoritative or broadly accepted definition of it (Borgman, 2015). Open Data have been proposed to be data that are *accessible*⁵, *usable*⁶, *assessable*⁷ and *intelligible*⁸ (Boulton et al., 2012). However, in trying to develop a definition of Open Data, Murray-Rust⁹ warns us about the complexity of the term, and about the impossibility of completely applying the principles of Open Access of publications to data. Thus, a new term, rather acronym, seems to compensate this complexity, namely *FAIR*¹⁰ data or *findable, accessible, interoperable and reusable* data, though FAIR data do not necessarily imply openness (Gregory, Groth, Scharnhorst, & Wyatt, 2019, p. 25). As a consequence of the not-so-recent high value awarded to scientific data (Borgman, 2015), some funding institutions, such as the European Commission, have adopted the FAIR principles as the basis for data management (Open Research Data Pilot¹¹) for its funded projects within its current funding program *Horizon 2020*.

³ “*The need to share and reuse data is an important topic in almost every high-level report or discussion concerning contemporary science. There are two overarching reasons for this emphasis. First, there is a belief that these activities are necessary to advance scientific research and solve important global problems. Second, there is a move to make the products of research available to a broad audience to support transparency, participation in the scientific process, and decision-making.*” (Faniel & Zimmerman, 2011, p. 65)

⁴ *Open Data* and *data sharing* are sometimes used interchangeably in both policy and scholarly discourse. However, there are some use trends. While the European Commission uses and promotes the term *Open Data* more frequently, US funding agencies and scholars prefer the term *data sharing*. Within the European Commission, Open Data is not used in an isolated way, but it is accompanied by other broader constructs such as, for instance, Open Science or Responsible Research and Innovation (RRI).

⁵ *Data must be located in such a manner that it can readily be found. This has implications both for the custodianship of data and the processes by which access is granted to data and information.*

⁶ *Data should be able to be reused, often for different purposes. The usability of data will also depend on the suitability of background material and metadata for those who wish to use the data. They should, at a minimum, be reusable by other scientists.*

⁷ *Recipients need to be able to make some judgment or assessment of what is communicated. [...] Assessability also includes the disclosure of attendant factors that might influence trust in the research. For example, medical journals increasingly require a statement of interests from authors.*

⁸ *Data must provide an account of the results of scientific work that is intelligible to those wishing to understand or scrutinise them. Data communication must therefore be differentiated for different audiences. What is intelligible to a specialist in one field may not be intelligible to one in another field. Effective communication to the non-scientific wider public is more difficult, necessitating a deeper understanding of what the audience needs in order to understand the data and dialogue about priorities for such communication.*

⁹ “*The label “Open Access” is a weak tool when describing access to, and re-use of, data. I and others have promoted the term “Open Data” (http://en.wikipedia.org/wiki/Open_data and references therein) to describe the need to consider data as a critical resource which needs political and legal activity. The use of Creative/Science Commons licenses is extremely valuable but will need refinement as the principles of Open Access and Open Source do not translate automatically to data*” In <https://blogs.ch.cam.ac.uk/pmr/category/berlin5/page/3/>

¹⁰ The term FAIR data or FAIR principles to data was launched by the scholarly community FORCE11 <https://www.force11.org/group/fairgroup/fairprinciples>

¹¹ https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm [2 February 2020]

Some of the underlying arguments of the benefits of Open Data or FAIR data include: more reliable research findings through the process of replication of original analyses; the discouragement of fraud; enabling scrutiny of research findings; the avoidance of duplication of data collection and thus, the reduction of research costs. In addition, the answer to novel research questions, training in research, and policy formulation and evaluation with existing data are also arguments for sharing research data (Irwin & Winterton, 2011; Norman, 1985; Thanos, 2017; Zimmerman, 2008).

It is argued that one of the most benefited actors of available-for-reuse data are researchers, because they can save money, time, and other resources such as personnel or equipment, especially when budgets are scant and low (Castle, 2003; Curty, 2015; Dale, Gilbert, & Arber, 1983; Hyman, 1972; K J Kiecolt & Nathan, 1985; Law, 2005). Time and resources are saved considerably when research studies require ethics clearance, data sharing agreements between institutions, or recruitment of large amounts of participants, for example. Available data for training and education are also frugal assets for academics (Fienberg et al., 1985). In sum, the scientific community can advance knowledge from secondary data by *understanding change, examining problems comparatively, improving general knowledge through replication and enlargement, and elevating and enlarging theory* (Hyman, 1972) with great benefits to research, innovation, education, and the citizenry (Borgman, 2012).

Apart from these practical benefits for the research community, there are other social benefits. Funding institutions can save money by not funding twice the collection of the same data, which, in turn, benefits the whole society because taxes, at least in the case of public funding, are better managed and distributed (Fecher & Friesike, 2014; Hyman, 1972¹²). Furthermore, the use of available data can benefit people from whom data were already collected, since primary data collection can aggravate situations where tensions exist. The use of available data can avoid awakening these tensions (Hyman, 1972).

The aforementioned potential benefits of the use of available scientific data have led to the emergence of all kinds of national and disciplinary initiatives related to research data infrastructures, data repositories, data sharing policies, and research data management (RDM) recommendations all around the world. Funding for these initiatives, which governments or public institutions mainly provide, is mostly perceived as insufficient, at least in Europe¹³, but still means a great spending. However, investment in these data initiatives, and to reap the benefits of available data, can be only recovered if the data are reused (Borgman, 2015; Pasquetto, Randles, & Borgman, 2017). Therefore, the reuse of data has recently attracted the attention of scholars from several disciplines, but mainly from scholars in Information Science (IS), who have suggested that there is a need for systematic investigations on

¹² “*These economies serve not only the private interests of sponsors and researchers but also the public good. Money, competent personnel, and time are scarce resources, and if they can be allocated to new research that is essential rather than wasted to duplicate data that are already available, so much the better for everyone.*” (Hyman, 1972, p. 7)

¹³ https://www.scienceurope.org/media/uuqf0i03/se-ke_briefing_paper_funding_rdm.pdf [2 February 2020]

data reuse practices in all fields of knowledge (Zimmerman, 2007, 2008) in order to design adequate initiatives and infrastructures for data sharing and management.

This context and the interest shown by scholars in Information Science motivates this dissertation. However, unlike previous IS scholars' studies on data reuse, this dissertation does not focus on the reuse of data from the perspective of the available data, or of the adequate research data infrastructures or data repositories. Instead, it focusses on the scientific reusers' broader decisions to achieve research goals with the use of secondary data. There are two main underpinning reasons for this focus. On the one hand, the reuse of data is not a goal per se. On the other hand, the reuse of data is a practice that has long existed before entering the *(big) data society* or *dataverse* and before the era of research data infrastructures (RDI) and repositories.

This dissertation does not question the importance of the data, or the fact that they are or should be easily accessible (which does not necessarily mean that they are *open*). Neither does it question the usefulness of adequate RDI or data repositories. What this dissertation questions is whether, perhaps, the fact that data is finally reused might not depend so much on how data are shared or made available, or on the availability of hugely expensive infrastructures and data repositories, but rather on the individual who reuses data and her motivations. Therefore, investments and efforts made in data sharing and research data infrastructures could be allocated better on the side of the reuser, and on encouraging researchers to reuse data as the Economic and Social Research Council in UK does (Heaton, 2008). In fact, some studies show that data reuse is less practiced than data sharing (i.e., Nahar & He, 2016).

The main objective of this dissertation is to uncover the linked causal forces that lead to the reuse of data. This study makes several contributions. First, it makes two theoretical contributions underpinned by decision-making theories. It proposes a heuristic model for understanding and explaining the scientific actor's behavior and decision-making, and a causal mechanism, which explains why and how the reuse of data happens. Second, it makes an empirical contribution by testing the causal mechanism in ten case studies of data reuse in health sciences, in which data reuse has been scarcely studied. Last, but not least, methodologically, this dissertation uses a diachronic study of the data reuse process based on the assumption that the value of conditions or factors affecting the process of reusing data changes over time, and that the sequence of the conditions may also affect the process.

Following this introductory chapter, Chapter 2 is a literature review of factors affecting researchers when using secondary data, which rather than being comprehensive, aims to draw attention to two different points of view regarding these factors. Chapter 3, Conceptual framework and theoretical strategy, provides the definitions of main concepts used in the dissertation and expounds the model of the scientific actor. Chapter 4, Methodology and methods, presents the theorized causal mechanism

for which I have drawn on decision-making concepts and theories. It also includes the research questions, and both the data collection and data analysis methods used to test the theorized causal mechanism in ten case studies. In Chapter 5, General overview of cases and data sources collected for each case provides a general overview of the ten case studies, as well as the data sources I have used in each case study to search for evidence. This chapter also includes information about the variability of different aspects of the ten case studies. Chapter 6, Empirical analysis: testing the data-reuse mechanism includes the analysis of each of the ten case studies following the structure and conditions of the theorized causal mechanism. Chapter 7, Findings, includes the main findings from the analysis. Chapter 8 discusses the main findings, and mentions some limitations of this study, and some opportunities for further research. Chapter 9 includes conclusions of this study as a whole.

Some chapters contain tables and figures. Most of the figures are in full size in annexes for a better visualization. There are also numerous footnotes, which I consider important as they contain relevant clarifications or additional information.

Chapter 2

Literature review

Some special style and skill seem to be required to reap the benefits even from a source as attractive and adaptable as the social survey. To exact the benefits from other buried and less pliable sources surely demands those special attributes. As the archives continue to grow in number and in the size and diversity of their holdings, it becomes ever more imperative to teach secondary analysis and increase the number of successful practitioners. Otherwise, we shall fall farther and farther behind, and the gap between our opportunities and our accomplishments will become bigger.

Herbert H. Hyman, 1985¹⁴ (1918-1985)

One of the grand challenges of data-intensive science is to facilitate knowledge discovery by assisting humans and machines in their discovery of, access to, integration and analysis of, task-appropriate scientific data and their associated algorithms and workflows. Here, we describe FAIR¹⁵ - a set of guiding principles to make data Findable, Accessible, Interoperable, and Reusable.

FORCE 11¹⁶

¹⁴ The Wesleyan edition of the book *Secondary analysis of sample survey* was published in 1987, but Herbert H Hyman wrote the introduction to this edition in 1985 shortly before his death.

¹⁵ The term FAIR was launched at a Lorentz workshop in 2014, the resulting FAIR principles were published in 2016.

¹⁶ Source: <https://www.force11.org/group/fairgroup/fairprinciples>

The heterogeneity of disciplines, and, thus of, research topics about the use of secondary data poses particular challenges in integrating the literature in a single review. Therefore, I have not aimed to carry out a comprehensive and critical review of the literature on the use of secondary data since this literature is extensive and scattered over a large number of, mainly, social sciences and health disciplines, and published in different scientific media, including grey literature. I rather focus in on some contributions by scholars in Information Science (IS) –my own discipline– and draw from them some issues, which differ from how some literature from other disciplines approach or report on the same issues. Such a difference can be summarized with the two 29-year-apart introductory paragraphs that precede this review. While I suggest that Hyman puts the onus on the *reuser* of the actual consummation of the use of secondary data, FORCE puts the onus on the *sharer*, *data* and *data infrastructures*. Hyman’s and FORCE’s different points of view are the ones that guide this literature review in order to reflect on these issues, namely on cognitive factors that affect researchers when reusing data.

In the following paragraphs, first, I present briefly selected IS scholars’ studies’ general characteristics and findings, which are mainly related to factors affecting researchers when they use secondary data. Second, I focus on findings about three of these factors, which have been reported by IS scholars to be essential for reusing data –*quality of the data*, *understanding and interpreting the data*, and *fitness of the data with the research question*, which ultimately need *contextual information* about the data. At the same time, I provide different points of view from scholars from other disciplines (sociology, nursing, philosophy of science, etc.) regarding these three factors and the contextual information needed by researchers to assess these factors. Third, I summarize these two different approaches and introduce the main research gap that guides this dissertation.

The value of this literature review lies mainly in drawing attention to these two different points of views regarding these three factors, which have been addressed previously (e.g., Irwin & Winterton, 2011; Moore, 2006), and which, I suggest, can ultimately shed some light into how and why some researchers decide to use secondary data and others do not.

As mentioned above, there is ample literature that addresses myriad issues related to the *use of secondary data* or *secondary analysis*. This literature is highly dispersed among many empirical disciplines, including nursing, criminology, policy studies, psychology, marketing, evolutionary science, among others¹⁷. The definition of *use (or analysis) of secondary data* varies across all these disciplines, and there is a lack of consensus on the meaning of the term (Cheng & Phillips, 2014;

¹⁷ Nursing (Doolan & Froelicher, 2009), criminology (Riedel, 2000), policy studies, (Young & Ryu, 2000), psychology (Trzesniewski, Donnellan, & Lucas, 2011), marketing (Lesch & Hazeltine, 2012), evolutionary science (Sidlauskas et al., 2010).

Smith, 2008¹⁸), even sometimes within the same discipline. The literature that deals with the topic of the *use of secondary data* has increased (Faniel & Zimmerman, 2011)¹⁹ in the last two decades, mainly, though not exclusively (for instance, Heaton, 2008), due to work by scholars in the field of Information Science (IS). IS scholars have mainly focused on factors affecting researchers when reusing research data, and on researchers' data reuse practices.

The literature by IS scholars differs from the literature that has traditionally dealt with the use of secondary data in two main aspects. First, IS scholars tend to use a different term when referring to the use of secondary data, namely *reuse of data* or *data reuse*. Few of these scholars define explicitly the concept *reuse of data* in their publications (Faniel & Jacobsen, 2010, van de Sandt, Dallmeier-Tiessen, Lavasa, & Petras, 2019)²⁰, but their preference for this term seems to be pointing out that their research focus is different²¹ from other disciplines. Most of the latter's literature focuses on the practicalities and methodological issues arising when analyzing data, and applying sound statistical inference to the available secondary data when studies are based on a quantitative approach. Occasionally, some of this literature also includes factors affecting researchers when using secondary data, though in a marginal way, and rarely deals with the process of searching and accessing the data (see for example Doolan & Froelicher, 2009; Jacobson, Hamilton, & Galloway, 1993; and Kiecolt &

¹⁸ "Numerous definitions of secondary data analysis appear in the literature, many with subtle differences, which together suggest a lack of consensus about what is meant by the term. Some definitions emphasise the usefulness of secondary data analysis for exploring new research questions: 'the study of specific problems through analysis of existing data which were originally collected for another purpose' (Glaser, 1963, p. 11). However, such definitions appear to disregard the potential of secondary analysis in re-analysing existing data sets with novel statistical or theoretical approaches in such a way that: 'secondary analysis is the re-analysis of data for the purpose of answering the original research questions with better statistical techniques, or answering new research questions with old data' (Glass, 1976, p. 3). One apparent area of consensus among those looking for a definition of secondary analysis is that it should involve the analysis of someone else's data: 'a collection of data obtained by another researcher which is available for re-analysis' (Sobal, 1981, p. 149); but this can be disputed as: 'even re-analysis of one's own data is secondary data analysis if it has a new purpose or is in response to a methodological critique' (Schutt, 2007, p. 4127). Given the differences in the definition and interpretation of secondary analysis that we see here, it seems likely that neat distinctions between primary and secondary data will not always be possible (Dale et al. , 1988). Such lack of consensus might leave one wishing to adopt a very general definition of secondary analysis such as that offered by Jary and Jary (2000): as 'any inquiry based on the re-analysis of previously analysed research data' (p. 540) or one such as Hakim's: secondary data analysis is any further analysis of an existing dataset which presents interpretations, conclusions or knowledge additional to, or different from, those produced in the first report on the inquiry as a whole and its main results. (Hakim, 1982, p. 1) Whichever definition one favours, secondary analysis should be 'an empirical exercise carried out on data that has already been gathered or compiled in some way' (Dale et al. , 1988, p. 3). This may involve using the original, or novel, research questions, statistical approaches and theoretical frameworks; and may be undertaken by the original researcher or by someone new." (Smith, 2008, p. 323-324)

¹⁹ "Until recently, scholarly investigations related to data sharing and reuse were sparse. They have become more common as technology and instrumentation have advanced, policies that mandate sharing have been implemented, and research has become more interdisciplinary". (Faniel & Zimmerman, 2011, p. 58)

²⁰ "Few studies on scientific data reuse formally define reuse but generally agree that it includes the secondary use of data for a purpose other than originally intended (Karasti and Baker 2008; Zimmerman 2008)" (Faniel & Jacobsen, 2010, p. 357)

²¹ "The term "data reuse" tends to be applied more consistently in the literature with a more empirical perspective. In other words, studies on data reuse have primarily focused on the investigation of data reuse among scientists (e.g.; Faniel, Barrera-Gomez, Kriesberg & Yakel, 2013; Faniel, Kriesberg & Yakel, 2012; Faniel et l. 2013; Kriesberg et al., 2013; Sands, et al., 2012; Wallis, Rolando & Borgman, 2013). Additionally, these studies are situated in the information science and technology field and approach data reuse as a process, instead of only as a methodological approach to data. This justifies the preference for the term "data reuse" throughout this document, which is considered a broader term and which is more aligned with previous empirical research terminology" (Curty, 2015, p. 36)

Nathan, 2012). However, IS scholars' research concerns are more related to the design of data infrastructures, data repositories, data sharing policies, data curation and preservation policies and best practices, and data management methods and tools. Therefore, IS scholars' research goals are to understand researchers' data reuse practices and factors and both the challenges and facilitators affecting researchers when searching for, accessing and using secondary data.

Second, most²² IS scholars have focused on the reuse of a specific type of data, namely *research data*²³, which are circumscribed to data collected, produced or used by academics or researchers within academic or scholarly settings, and used as evidence of scientific claims. Thus, the term *reuse of research data* in their investigations implies that data being reused or targeted for reuse are only produced in academic or scholarly settings and used as evidence of scientific claims. This definition excludes the study of the reuse of government statistics, business records, archival records, administrative data, hospital and health care data, that is, data collected in institutional settings other than research or scholarly contexts. An example of an implicit definition of research data as data collected by academics or scholars is that of Zimmerman, who "investigated the processes by which ecologists locate data that were initially collected by other[...] [ecologists]" (Zimmerman, 2007, p. 5). In Chapter 3. Conceptual framework and theoretical strategy, I elaborate more on both the terms *data*, *primary data* and *secondary data* by drawing on both the philosophy of science and the sociology of science, and provide the conceptual definitions that guide the empirical work of this dissertation.

²² One exception, for instance, is Niu's research (Niu, 2009b)

²³ Some of these definitions are:

"Research data represents data obtained by scientists through systematic investigations, including surveys, observations, experiments, and simulations. Based on Given and Porter (2008) and Heaton (2008), this study makes a distinction between non-naturalistic and naturalistic data. For the purpose of this study we only consider the non-naturalistic type of data, which consists of self-reported or researcher-manipulated, quantitative or qualitative primary data generated with a research purpose, gathered utilizing different instruments (e.g. questionnaires, video recording, voice recording, etc.) and data collection techniques (e.g. surveys, experiments, observations, interviews, etc.). Thus, the term research data in this study does not include data that were produced independently from the actions of scientists and were not elicited by a research action (e.g. log files, audit trails, transaction user logs, navigation history, location tracking, autobiographies, personal diaries, letters, official documents, photographs, third parties' e-mails, tweets, online reviews), regardless if they are amendable to inductive or deductive forms of inquiry." (Curty, 2015, p.12)

"We can then think that the defining trait of research data – as opposed to any other kind of data – is that they perform as evidence for certified knowledge, which means knowledge that is peer-reviewed and published." (Pasquetto, 2018, p.11)

"By research data, we mean scientific or technical measurements, values calculated, and observations or facts that can be represented by numbers, tables, graphs, models, text, or symbols and that are used as a basis for reasoning or further calculation [7]. Such data may be generated by various means, including observation, computation, or experimentation. Scientists regard data as accurate representations of the physical world and as evidence to support claims [8]." (Thanos, 2017, p. 2)

"Research data sometimes are distinguished from resources such as government statistics or business records (Open Knowledge Foundation 2015). Here we rely on a definition developed earlier, in which data refers to 'entities used as evidence of phenomena for the purposes of research or scholarship' (Borgman 2015, p. 29). The above definition is useful in determining the point at which some observation, record, or other form of information becomes data." (Pasquetto et al., 2017, p.2)

Despite the recent growth of literature on the *use of secondary data* or *reuse of data* (used interchangeably from now on), especially among IS scholars, there are still few studies on this topic (Curty, 2015)²⁴ compared to those about *data sharing* (Curty & Qin, 2014; Pasquetto et al., 2017).

IS scholars' interests in different aspects of the data reuse process, possibly together with scholars' ontological views, have led them to choose different research questions²⁵, to use different conceptual frameworks and theories –if any²⁶–, and thus to employ different methodologies. In addition, since the research concerns about the use of secondary data by IS scholars are relatively incipient within the *Open Science* context described in the introduction, consequently most of the studies are exploratory and descriptive. Studies by IS scholars, are predominantly discipline-based, may be due to the assumption that data reuse practices are partly contingent on broader disciplinary or epistemic practices (Borgman, 2007, 2015). The majority of the studies have been conducted in social sciences disciplines²⁷, although there are a few studies comparing data reuse across disciplines (Faniel, Barrera-Gomez, Kriesberg, & Yakel, 2013; Faniel, Frank, & Yakel, 2019; Gregory et al., 2019; Yoon & Kim, 2017). See Annex 24 to see the different research questions addressed, as well as the theoretical frameworks, empirical fields and methodologies used to answer the former, and the main findings of these studies. However, comparison of findings and conclusions from these studies is rather difficult.

Despite the aforementioned heterogeneity among studies on data reuse by SI scholars, there is a common thread. These studies aim to understand researchers' data reuse practices and factors, both challenges and facilitators, which affect researchers during the process of reusing data. Within the social sciences, Curty found up to twenty-five different factors affecting researchers in social sciences disciplines when reusing data (2015). Twenty-four of these factors were identified in previous studies of secondary use, not exclusively in social science disciplines and not only regarding data reuse, but also regarding knowledge²⁸.

²⁴ *Although various arguments about the importance of data reuse have been put forward, the literature on research data reuse remains relatively scarce with few systematic and empirical investigations.* (Curty, 2015, p. 35)

²⁵ Most of the research questions are exploratory and descriptive.

²⁶ For example, the one by Federer et al. (2015) is completely atheoretical, and despite being the term “reuse” in the title, the article hardly address it. The one by Fear (2013) does not explicitly refer to any theory or concept either. Her research on scholarly impact of data reuse is built on bibliometric analysis of data citations.

²⁷ Some of the studied empirical fields are archaeology (Daniels, 2014; Faniel, Kansa, et al., 2013), experimental geomorphology (Hsu, Martin, McElroy, Litwin-Miller, & Kim, 2015), ecology (Zimmerman, 2003, 2007, 2008), molecular biology (Pasquetto, 2018; Pasquetto et al., 2017; Pasquetto, Sands, Darch, & Borgman, 2016), earthquake engineering (Faniel & Jacobsen, 2010) and social sciences (Akmon, Zimmerman, Daniels, & Hedstrom, 2011; Curty, 2015; Curty, Crowston, Specht, Grant, & Dalton, 2017; Curty & Qin, 2014; Curty, Yoon, Jeng, & Qin, 2016; Faniel, Kriesberg, & Yakel, 2012, 2016; Fear, 2013; Niu, 2009a, 2009b; Yakel, Faniel, Kriesberg, & Yoon, 2013; Yoon, 2014b, 2016a; Yoon & Kim, 2017)

²⁸ I would suggest being cautious when comparing data reuse with knowledge reuse since *data* and *knowledge* are different concepts. There have been many attempts from different disciplines (psychology, philosophy, economics, management, sociology, etc.) and from different perspectives to explain and set up differences between knowledge, information, and data (Case, 2007). The DIKW (data, information, knowledge, wisdom) pyramid is an example of these situated socially-constructed attempts (Kitchin, 2014). However, depending on the disciplines and approaches, boundaries among these elements are so blurry that making a clear distinction among them is nearly impossible. I argue that in some disciplines, the difference between data, information, and knowledge is not relevant, but it is in the context

The findings regarding factors affecting researchers in disciplines other than social sciences are quite similar to those found by Curty (2015). Results show that there are differences in how researchers overcome challenges due to researchers' different strategies and resources. For instance, in order to assess the quality and trustworthiness of the data, ecologists consider the reputation of the data producer and the familiarity with artifact collection processes (Zimmerman, 2007). Earthquake engineers assess quality and trustworthiness by analyzing the experimental processes which created the data (Faniel & Jacobsen, 2010). Faniel & Jacobsen (2010) point out that these differences in strategies and resources may not be necessarily rooted only in epistemic issues, but in other issues such as the type of data or the method²⁹. Therefore, these authors suggest considering factors other than epistemic issues to find explanations about factors that affect researchers when making decisions about data reuse (Faniel & Jacobsen, 2010)³⁰.

Factors affecting the process of reusing data have been usually classified into two main groups – challenges³¹ and facilitators³²– according to how factors have been self-reported as positive or negative

of this dissertation. The difference between these two concepts is important to the extent that, of the two concepts, data are the ones being used or aimed to be used as evidence of scientific claims. Comparison between findings about data reuse and findings about data knowledge might result in unfruitful, if not problematic, conclusions despite similarity of findings. As Sayer (2010) suggests, when doing research we have to be very cautious in knowing very precisely what kinds of objects –and their abstractions, parts of it, and attributes– we are studying.

²⁹ “Take the findings from our study as an example. It is likely that reusability assessments are influenced by not only our respondents membership in a particular community (EE researchers), but also their reuse of a particular data type (experimental), for a particular reuse purpose (model validation)” (Faniel & Jacobsen, 2010, p. 371)

³⁰ “Without further study, it is impossible to say whether and to what degree scientists in other communities appeal to the same factors when making data reuse decisions” (I. Faniel & Jacobsen, 2010, p. 373)

³¹ a) Reusing other people’s data in research can be perceived as less valuable, and thus have fewer pay-offs than conducting research based on new data; b) Hesitation to reuse data which was obtained through consent to a particular study and/or unwary violating aspects of confidentiality, copyright and data protection; c) Misinterpretation, incorrect association, or misuse that might occur while reusing other people’s data; d) The susceptibility to faulty data given the difficulty of identifying potential errors on data collected by others; e) The effort of identifying original contributions from second-hand data and exploring different issues not yet explored or overlooked by primary researchers, as well as other reusers; f) Refers to data accessibility. The effort associated with obtaining access and retrieving data; g) Refers to data discoverability. The effort associated with data discovery and the identification of relevant and potentially reusable datasets; h) The effort of working with data that was generated based on different research questions and/or hypotheses, using particular instruments or techniques for data collection, in a particular context and time-frame, and having specific variables, constructs, and measurements. It also accounts for the effort associated with resigning initial ideas and reframing the study design and goals in order to accommodate the existing data; i) Refers to the effort to get data ready for reuse and manipulation, including: screening and cleaning processes, dealing with missing data, adding/complementing data, and putting it in an appropriate format, sorting, recoding etc.; j) The effort associated with making sense of the data and thoroughly comprehending the original study; k) Whether the supplementary documentation provided along with the data is sufficient, easy to understand and clearly explains the methodology, the rationale of the study, etc. to support reuse; l) Whether the topic, level of analysis, and type of data are compatible with the purpose of reuse; m) How consistent and complete data are perceived to be; n) How well-designed and executed the study was (Curty, 2015, p. 73-74).

³² a) Data can be reused to answer different questions other than the ones covered by primary studies or for replication/validation; b) Ways to circumvent data collection problems associated with time and cost (money) to minimize duplicated efforts or the need to develop data collection skills; c) Data available for reuse are considered to some extent credible and reliable, otherwise they would not be shared and available to the public; d) How trustful and credible data producers (institutions or individual authors/contributors) are; e) The availability of comprehensive and detailed data documentation improves chances for data reuse; f) The existence of repositories and their capability to organize, self-guard, and facilitate access to reusable data improves conditions for reuse; g) Communication with primary investigators helps reusers to obtain additional information about the data and the study; h) Having institutional support and assistance from the data repository personnel or at the university level (e.g. statistical center,

in studies of data reuse. However, the categories “challenge” and “facilitator” can be misleading because one factor can be both a challenge and a facilitator. For example, the factor *having supplementary documentation* about data and a study, which may have initially facilitated the collection of the data, can subsequently be a challenge if the documentation does not provide all the information needed or is ambiguous. On the contrary, the factor *having supplementary documentation* can only be simply a facilitator if it is complete, accurate and understandable in every way and requires no resource consuming supplementary work on the part of the data re-user. The latter situation is, of course, quite an unlikely scenario.

Bringing all these factors together in one study, as was done by Curty (2015), or simply presenting them in a comprehensive list can therefore be misleading for two main reasons. On one hand, it may lead us to think that all of these factors occur within a single process of data reuse. Although this is plausible, it is doubtful that this happens very often. On the other hand, even if all these factors converge in a single process of data reuse, they do not do so simultaneously, as Faniel and colleagues (Faniel et al. 2019) found. Data reuse is a process, which in some cases can last months or even years. These two issues appear to have been neglected in the empirical approaches used in previous studies.

Of all the factors identified in previous studies, three have been consistently reported as being very influential with regard to researchers’ use of secondary data. A certain level of *quality of the data*, the fact that the researchers need to *understand the research context* from which these data originated, and the fact that a *compatibility or fitness between the data and the research question* posed by the secondary user must exist. It can be argued that there is one crucial common aspect of these three factors, in that their *value* depends exclusively on the secondary user’s subjective cognitive and intellectual assessment.

All findings related to these three factors state, with no discrepancy, that secondary users need to have contextual information about the data for understanding the data, interpreting the data, assessing the quality of the data, and judging if the data fit their research question, in order to decide whether to use or not use the data for their own research (Faniel et al., 2013; Faniel, Frank, & Yakel, 2019; Faniel et al., 2012; Fear, 2013; Frank, Yakel, & Faniel, 2015; Kim & Yoon, 2017; Yakel et al., 2013). Researchers need to know about the specific context of data production, the producer’s research methods, how the research is carried out, how variables are defined and measured and which measurement devices are used (Faniel et al., 2013). This information should be available in various kinds of documentation (e.g., codebooks, metadata, etc.) or be accessible by other means (e.g., reaching original collectors or producers of the data as some reusers do (Yoon, 2017). Previous studies

library, IT center, advisors); i) Importance of training on secondary analysis for skill development. Expertise in secondary analysis will lead to more reuse of data; j) Disciplinary tradition or perceived acceptance of the reuse of data. Some disciplines are more prone to data reuse than others; k) The acceptance or habitual practice of data reuse among colleagues and peer recommendations to reuse particular datasets (Curty, 2015, p. 73-74).

have conclusively found that the availability of data documentation, about both the original study and its data, is a key facilitating factor for data reuse (Niu, 2009a, 2009b; Zimmerman, 2007, 2008) because provides contextual information about the original study. Contextual information includes all kinds of information for each of the steps that happen in the context of discovery. This is because the information in the context of justification is usually not enough in order to provide the secondary analyst with information for a full judgement of the data (Reichenbach, 1938).

Researchers know about contextual information mainly through *sufficient*, *accurate*, and *easy-to-use* documentation, including codebooks (Niu, 2009a; Niu & Hedstrom, 2009), and metadata (Pasquetto, 2018). Faniel, Frank, & Yakel (2019) identified up to twelve types³³ of relevant contextual information in a study involving quantitative social scientists, archaeologists, and zoologists. Yet, researchers deploy different strategies in order to assess the quality of the data. For example, ecologists may use their tacit knowledge to *reconstruct* in their minds the process of collection of the data (Zimmerman, 2007). They also assess the reputation and the competency of the original data collectors or producers (Zimmerman, 2007). Researchers may also contact the original data collector or producer (Niu, 2009a) or assess the trustworthiness of data curators and repositories (Donaldson & Conway, 2015; Frank, Chen, Crawford, Suzuka, & Yakel, 2017; Yakel et al., 2013; Yoon, 2014a).

Researchers may decide to reuse data despite not having been able to assess these data fully. They may also decide to reuse the data despite concluding that the data are of insufficient quality (Curty, Crowston, Specht, Grant, & Dalton, 2017)³⁴. If this latter situation happens, then researchers should subsequently explain their decision with common honest research reporting. Failure to “evaluate completeness and accuracy, limitations associated with invalid or unreliable data should be clearly addressed in the research report” (Garmon Bibb, 2007b, p. 98).

Contextual information is also claimed to be necessary to interpret data. Some IS scholars have reported that researchers need contextual information in order to interpret data “correctly”, as they fear misinterpreting the data.

³³ “Findings show researchers mentioned twelve types of context information across three broad categories: 1) data production information (data collection, specimen and artifact, data producer, data analysis, missing data, research objectives), 2) repository information (provenance, reputation and history, curation and digitization), and 3) data reuse information (prior reuse, advice on reuse, terms of use).” (Faniel, Frank, & Yakel, 2019, p. 2)

³⁴ A common theme in the literature on data reuse has been the difficulty of being able to trust or even understand data produced by others. In contrast, the most striking result of our study is that expressed lack of trust in reused data was not a factor explaining a lack of data reuse. Indeed, many respondents agreed with questions about the lack of trustworthiness of data and still reported reusing data. This effect does not change with development of data management practices. It is difficult to unpack this result with the available data. It appears, however, that while respondents are aware of the potential pitfalls of reusing data, they apparently feel that they can overcome them in their own practice. (Curty et al., 2017)

Without [contextual information] E[arthquake] E[ngineer] researchers know they are less likely to interpret the data properly and thus are less likely to trust that they can reliably recreate the data; Without such context information, EE researchers may not be able to reconcile the data with what they understand about the experiment. In these cases they tend to reject the data rather than risk misinterpreting the data or drawing the wrong conclusions (Faniel & Jacobsen, 2010);

[U]nderstanding and interpreting the unstructured data collected in the social sciences often requires detailed contextual information (Yoon & Kim, 2017)

[C]omprehensive metadata is needed to support the correct interpretation of the data. Scientists need to feel confident that they have enough information about the data to minimize the chances of making wrong assumptions and unintentionally misuse it [8, 14, 45]. (Curty et al., 2017)

However, these findings contrast with Pasquetto's statement regarding the role of contextual information for interpreting the data, at least in the form of metadata and ontologies.

[M]etadata and ontologies cannot inform a data reuser on the potential for a certain dataset to contain novel information worth a scientific publication (Pasquetto, 2018, p. 217).

Pasquetto's statement is closer to what some other scholars suggest regarding the availability of contextual information in order to assess the quality of the data, understand and interpret the data, and assessing the fitness between the data and the research question. For instance, Irwin & Winterton (2011) provide three examples of how the interpretive capacity of the secondary user has to be put into action for reusing data. For example:

Bishop's investigation (2007) into contemporary eating practices involved (re)using data from two previously conducted research projects, Blaxter & Patterson's (1982) Mothers and Daughters and Thompson's (1975) The Edwardians (both as cited by Bishop 2007). In her own analysis Bishop prioritised aspects relating to „convenience“ foods and family eating practices that were contained (respectively) in these two archived studies. In neither study was convenience food or family eating practices the focus of the research. (Irwin & Winterton, 2011, p. 4)

Irwin and Winterton (2011) do not challenge the claims that contextual information is needed to understand how and why the data were produced. Rather, what they emphasize, in citing Mason, is

that it is the reuser's *reflexivity rather than their proximity to the original context in which data were produced which is the key to [reusing data successfully]* (2011, p. 8).

If we consider a commonsense meaning of the verb *interpret*³⁵, when we interpret something, we decide what its meaning or significance is. In other words, we give or attach meaning to it. This implies that when reuse happens for the same or other purpose other than the original purpose, we are reinterpreting the data, and thus attaching a new meaning to the data. (Re)interpretation of the data depends on the secondary user's theories and socio-cognitive factors, as Vinck (2010)³⁶ explains, but not on the data themselves. In the philosophy of science, Leonelli (2016) also recognizes that having access to the contextual information of the data is important, yet this information is not necessary to attach a new meaning –or interpretation– to the data in such a way that data become evidence of new scientific claims by a new user. Reusers should be able to (re)interpret data in novel ways according to *their own backgrounds and interests* (Leonelli, 2016, p. 32). Re(interpretation), though, is not an easy cognitive task. It requires expertise not only in the field (Leonelli, 2016) or more specifically in a specific domain (Borgman, 2015)³⁷, but also training and skills (Hyman, 1972; Irwin & Winterton, 2011) and the data must fit with researchers' cognitive structures or frames of references (Ansbacher 1950; Vinck 2010).

Regarding *data fitness* or simply *fitness*³⁸, all studies where the issue of the fitness has arisen, report it as a condition for data to be reused (e.g., Curty, 2015; Curty & Qin, 2014; Niu, 2009b). This is consistent no matter how other factors are affecting researchers in the process of reusing the data, for instance, the availability of proper documentation.

³⁵ See for example the online American Heritage Dictionary [<https://ahdictionary.com/word/search.html?q=interpret>] or the online Collins Dictionary [<https://www.collinsdictionary.com/dictionary/english/interpret>], both consulted August 19, 2019.

³⁶ *Facts in themselves rarely lead to proof of hypothesis or to their refutation. Scientists are wary of the apparent and illusory 'evidence of facts'. Facts are only deemed to be significant or valid once various interpretation, evaluation and qualification processes have taken place and through their connection to prior knowledge. As a result, observation loses its primary role, and is instead assigned to the interpretative framework that allows facts and data to be qualified. Observation is dependent on accepted theories and the socio cognitive factors used to interpret them (accepted conventions, the language used and background knowledge). Fact is also indissociable from the way in which it is expressed (linguistically speaking, in particular), which carries both meaning and interpretative elements. Organising and classifying facts requires there to be a concept in place. The identification and isolation of a phenomenon or object from the flow of sensory perception also implies that the observer has concepts at his/her disposal. Categories of thought therefore make their own imprint on observations. Experimentation is always accompanied by interpretation of the phenomenon. Raw data already constitute an interpretation. In addition, experimenters carry out adjustments so as to obtain satisfactory data. These corrections play an important role in the production of 'raw' data precisely because they are guided by the interpretation of the phenomenon. Interpretation, far from following on from observation, actually precedes it.* (Vinck, 2010, p. 158)

³⁷ *Collecting, using and interpreting data –big or little- usually depends heavily on expertise in the domain; Collins and Evans 2007) [...] Pritchard, Carver, and Anand (2004) found that groups cluster around data collection practices rather than field.* (Borgman, 2015, p. 60)

³⁸ Other terms have been also used for the idea of the *fitness of the research question with the primary data* –or simply *data fitness* or *fitness*–. For instance, *relevance* (Faniel, Barrera-Gomez, et al., 2013), *level of difference* between the research purpose of the secondary analysis and the primary study (Hinds, Vogel, & Clarke-Steffen, 1997), or *fit for purpose* or *utility* of the data (Palmer, Weber, & Cragin, 2011).

However, users' incentives to use secondary data mostly depend on how well the data fit their information needs rather than documentation quality. A well-documented dataset will not be used if it doesn't answer users' research questions. Users will not give up using a dataset simply because it is poorly documented. (Niu, 2009b, p. xiii)

Fitness is usually understood as the suitability of the reuser's research question with the data. In the literature, we can find several types of fitness, although no specific terms have been used to refer to them. It is important to distinguish clearly between these types of fitness here, as they may provide us with a better understanding of the challenges for reusing data.

The most common type of fitness reported in the literature is related to concepts or themes, in which theories, concepts, constructs and variables have to be carefully considered (Curty, 2015; Orsi et al., 1999) (from here forth *conceptual fitness* or *thematic fitness*). However, conceptual or thematic fitness might not be sufficient to guarantee fitness between the data and the reuser's research question, since other types of fitness have been identified that may intervene. For instance, based on Orsi and colleagues (1999)³⁹, we can also talk of *measurement fitness* and *technological fitness*, although these authors refer to compatibility or fitness between two data sets. Curty's findings refer to *technological fitness* when she reports that data have to be available in the format that secondary users need (Curty, 2015; Yoon, 2016b). Participants in her study also stated that primary data have to be at the same level of analysis of the new study, so we could talk of *level-of-analysis fitness*. Both *conceptual fitness* and *measurement fitness* are suggested by Secrist (1920)⁴⁰ as also to be considered when reusing data.

The use of different instruments or protocols during the collection of primary data, compared with the instruments and protocols that the secondary user would or could use, is also important to consider. These differences can be due to geographic⁴¹ differences as Borgman suggests (2015), but also to time differences since different instruments or protocols could be used in the same place, but in different moments (Ribes & Jackson, 2013). For example, two or several ocean measurements may be made with the same instruments, protocols and exact location, but 20 years apart. When the time span is

³⁹ "Combining data sets does not assure that the two data sets are compatible. That is, the two data sets need to be combined technically to determine if they are compatible. Prior to this point, the data sets cannot be compared. The major technical tasks prior to combining data sets include translating data into one statistical package, adding a variable in each data set to distinguish data sets when they are merged, and creating new identification numbers if they are the same in each data set." (Orsi et al., 1999, p. 138)

⁴⁰ "A second consideration relates to the applicability of data to the problems being considered. Are the facts germane? Do the units of measurements in which they are expressed admit of use for the particular problem in mind? Many statistical data having only a general application may, if used with discrimination, substantiate or lend support to a contention which they would not be sufficient to uphold *de novo*. The bearing of these tests assumes importance only by detailed study of the uses to which one desires to put data and the conditions surrounding their collection. No single rule or principle is sufficient to cover all cases." (Secrist, 1920, p. 20)

⁴¹ However, local data taken at one site are not necessarily consistent with local data taken at other sites, because sites may use different instruments and different protocols for data collection. (Borgman, 2015, p. 58-59)

longer, the issue of fitness can be worsened (McAllister, 2018). I use the terms *instrument or protocol fitness* to describe these types of fitness issues.

As mentioned earlier, in order to know whether there is a fitness of the novel research question with the secondary data, the researcher needs to have as much information as possible about the data collection or creation circumstances (McAllister, 2018; Parsons et al., 2011). However, contextual information to assess the quality of the data can become a secondary factor affecting researchers in comparison to the *fitness* of the research question with the data. From their findings, Niu (2009c)⁴² and Garmon Bibb (2007a)⁴³ argue that researchers can, under specific circumstances, prioritize data fitness over quality of data, or lack of information to assess the quality of the data properly.

Nevertheless, and far from undervaluing the challenges secondary users face with fitness, fitting the question with the data is not exclusively a challenging task in the process of using secondary data. Fitting research questions, methods, and theoretical frameworks with data is also a challenge, although not always acknowledged, when a researcher uses their own primary data (Clarke & Cossette, 2000⁴⁴; Zimmerman, 2007⁴⁵). The acknowledgment of both the general messiness of a research process and the unfitness of original research plans with data can be barely identified in scientific publications

⁴² “Inadequate documentation increases use cost and may turn users away in some situations. However, users’ incentives to use secondary data mostly depend on how well the data fit their information needs rather than documentation quality. Users would not use a dataset if it doesn’t answer their research questions, no matter how well documented it is. When data and documentation do not fit users’ needs perfectly, users need to decide whether to compromise their information needs or give up using the data. Their decision-making is not necessarily changed by documentation quality. With inadequate documentation, users need to seek outside information to supplement documentation and reduce uncertainty. Their decision to use or not depends on how much they can benefit from using the data, the cost of overcoming inadequate documentation and the potential cost to them to collect the same data. As a result, many users do not want to use small secondary datasets because the potential cost of collecting the same data is not greater than the costs of using secondary data, which are caused by uncertainties, information seeking and the compromization of information needs. On the other hand, even though many administrative records are regarded as messy and poorly documented, users still often choose to use those data because there is no way they can collect the same data.” (Niu, 2009b, p. 71)

⁴³ “However, data sets available from nonstate or nonfederal sources, from unpublished primary studies, and from local health care organization clinical, administrative, and health survey databases may not have sufficient supporting documentation to appropriately assess the quality of a data set. If supporting documentation is not available or does not contain sufficient information to conduct this initial assessment, one should consider using a different data set. If a decision is made to acquire and use the data set for research, without sufficient documentation to evaluate completeness and accuracy, limitations associated with invalid or unreliable data should be clearly addressed in the research report.” (Garmon Bibb, 2007, p. 98)

⁴⁴ “However, in primary research as well, compromises and deviations from methodological ideals, simultaneous rebalancing of multiple theoretical and design factors, and repeated recasting of research questions are nearly always required.” (Clarke & Cossette, 2000, p. 112)

⁴⁵ “The findings from this study make clear that the reuse of data to create new knowledge is “doing science,” and is therefore subject to the same norms, requirements, and challenges that affect [the use of primary data]. Therefore, it should not be surprising that the use of existing data to create new knowledge is a complex, difficult, and iterative process. While there is much progress to be made in making it easier for scientists to find, retrieve, aggregate, and understand data, the practice of science can never be completely automated.” (Zimmerman, 2007, p. 14)

because the *black box* of the research process remains closed (Marshall & Rossman, 2011)⁴⁶, or if opened, researchers may falsify retrospectively their research process (Vinck, 2010)⁴⁷.

No researcher denies that the reuse of data is or can be both a challenging and daunting task. However, in general, while some scholars think that the task is insurmountable, others propose different strategies to reuse data successfully. For instance, Doolan and Froelicher suggest three different approaches or strategies to deal with the unfitnes of the research question with the data: first, adjust the research question to fit the data; second, disregard the data and search for others; or third, disregard the whole study and design a new one (Doolan & Froelicher, 2009). These strategies presuppose that the research question comes first, and the search for the data comes afterwards. However, only the first strategy – make adjustments – leads to the reuse of targeted data, since with the other two strategies the initial data target is abandoned. Regarding the adjustment of the research question to the data, other scholars have reported this same reframing strategy, which in some cases may imply *unacceptable compromises*. For instance,

This involves a sound approach to conceptualizing the problem to be studied, having a theoretical framework, carefully delineating the research questions to be answered, identifying concepts and how they are operationalized, matching questions with data sets that can answer these questions, an interplay with the data to recast the research questions to fit the data, devising new coding systems, recoding data to fit with the new research questions, and applying rigorous analysis to answer the questions. (Rew, Koniak-Griffin, Lewis, Miles, & O’Sullivan, 2000, p. 226)

If the data set is of sufficient quality, the PI then determines if the sample and measures used in the study are a good fit. It is important that the measures are both

⁴⁶ “Quite unlike its pristine and logical presentation in journal articles – “the reconstructed logic of science” (Kaplan, 1964) – real research is often confusing, messy, intensely frustrating, and fundamentally nonlinear. In critiquing the way journal articles display research as a supremely sequential and objective endeavor, Bargar and Duncan (1982) describe how, “through such highly standardized reporting practices, scientists inadvertently hide from view the real inner drama of their work, with its intuitive base, its halting time-line, and its extensive recycling of concepts and perspectives” (p.2)” (Marshall & Rossman, 2011, p. 55)

⁴⁷ “Reference has been made to the case of a scientist who lacked sufficient quantities of the enzyme he was studying, but decided to pursue his work using another enzyme that came to hand. Incidentally, he observed that the ears of the rabbits he used in his experiments softened, before returning to their usual stiffness. Nevertheless, he continued with his original project. Seven years later, during a discussion with a colleague, he remembered the incident and went on to ask his students to think about the problem. This type of factor, which clearly affects the course of research, is erased from the reports published. Barber and Fox (1958) talked about the ‘retrospective falsification’ that explains the differences between the way research is actually performed and the way it is presented in publications. Thus, Feltz (1991) reported on the case of a biologist who, when faced with difficulties in acquiring sheep foetuses, decided to use hamster foetuses, considered to constitute a good alternative model. This change prompted the scientist to alter both her line of questioning and her work methods, but also to produce results of another type. Confronted with a lack of resources [that is the primary data she expected to have], she reorganised her work and the structure of the problem she wanted to address. In her thesis and publications, however, the story was reconstructed.” (Vinck, 2010, p. 153-154)

appropriate and collected with a sufficient level of accuracy and detail for the proposed research. This may require that research questions and hypotheses be refined to better fit the available data set. The research question may need to be modified depending on the data available. (Doolan & Froelicher, 2009, p. 210)

Some feel that secondary analysis bypass the long process of framing research questions and that “ready-made” data sets enable researchers to skirt the issues of definition and operationalization of major concepts (Kasl, 1995). In avoiding the many difficulties that plague original data collection [...], Kasl states, secondary analysts are too often forced to make unacceptable compromises. (Clarke & Cossette, 2000, p. 111)

Two other strategies have been proposed by Garmon Bibb (2007). One of these strategies is more related to the quality of the data, albeit it may be also applied to fitness. First, Garmon Bibb (2007) proposes to proceed in the reverse direction, with the first step being to find a research question that fits an existing data set or database. In this case, it may be possible that a complete fitness is established from the very outset of the study. Second, researchers may end up reusing the data despite not having a complete fitness, but have to acknowledge these limitations in their findings and conclusions of their publications.

There are some examples of Garmon Bibb's first strategy (2007) in previous studies on data reuse by IS scholars. For instance, Curty (2015)⁴⁸ reports that some of her interviewees first formulated general research questions or had in mind theoretical frameworks or hypotheses before approaching the available data, and that later they reframed their original research questions in order to fit the data into the former. Indeed, research does not happen in a linear way, and the formulation of a research question is constantly tweaked with the data during the process of a study (Abbott, 2004). Research happens in a constant looping as Clarke and Cossette (2000)⁴⁹ explain.

⁴⁸ “Interviewees initiated the reuse process with a given initial theoretical framework, hypotheses, or general research questions, which eventually had to be reframed to adjust to the available data they aimed to reuse. This pattern is illustrated in Heidi and Ivan’s descriptions:

(...) so (I started) just trying to see what would be of interest, what [researchers] would be already working with, so I found very, several variables I was interested in, so I started going through..., figuring out how they would be of use to my research questions, which I was looking at resilience in emerging adulthood. So, this is...you know...being a national survey with families obviously they had, they had quite a few items in there. I think basically when I first started I had...I would have just my general idea and then, I went to see which variables would kind of fit in what my interest was. It is kind of how I started. (Heidi, PhD).” (Curty, 2015, p. 69)

⁴⁹ “Perhaps some of the concerns stem from differences of opinion on the nature of the research process. Many investigators have been taught to view research as a linear process whereby the researcher begins with a literature review that reveals gaps in knowledge about a phenomenon. After reflecting on a theoretical framework to guide his or her study, the researcher chooses a population, a set of variables, and a series of relationships; selects measures; plans data collection and analysis; then gathers and analyzes the data and writes for publication [...]. At least one

The three strategies that lead to the reuse of data imply that researchers relinquish in some way their ideal or initial research goals. This raises the questions of why and when some researchers would adopt these strategies. The only attempt to reply to these questions identified to date is Hyman's study on secondary analysis (1972).

Hyman (1972) recognizes that there are several obstacles – legal, logistical, financial, technological – to secondary analysis, but for Hyman the onus is on the secondary analyst, who lacks not only analytical knowledge but also the motivation to do complete studies, with regard to the lack of utilization of vast amounts of information. In his book *Secondary analysis of sample surveys*, Hyman provides three accounts of the successful reuse of data, which are based on his empirical research on case studies of successful secondary analysis (Hyman, 1972). In the three accounts, he also attributes the successful use of secondary data to the reuser or secondary analyst, and not to the data, sharer, archive, or any kind of data infrastructure.

First, Hyman, distinguishes *two contrasted types of secondary analysts – the casual and the compulsive* (p. 75-76). The casual researcher – not necessarily novel – does not pay attention to the treatment of errors in data. In contrast, the compulsive researcher – not necessarily experienced – provides myriad details or admits limitations in her findings due to some lack of quality of the data or to the impossibility to verify thoroughly the quality of the data. The casual and the compulsive researcher are two extremes or poles, and for Hyman most of the uses of secondary data would fall in between these poles. Hyman recognizes, though, that in order to become a compulsive researcher, one might have started out being a casual one. Ideally, the better extreme is the compulsive one. This type of researcher does not need any training in quantitative methodology. However, she has to...

be concerned about the ambiguity or invalidity of [her] indicators as instruments for the measurement of particular variables. If [she] wishes to take advantage of almost unique opportunities that [old data] provide, then ultimately [original emphasis] [she] must familiarize [herself] with certain research designs, whose properties are complex and difficult to master (Hyman, 1972, p. 76)

version of this worldview posits that in "good research" the study's methodology flows from research questions, not the other way around.

In practice, however, the planning of studies involves simultaneous consideration of the state of the literature, what can be measured, what subject pools might be available for data collection, and what analytic methods exist and are within the researcher's expertise and budgets of time and money. A linear approach, in which each "step" of the research process or portion of the research protocol is completed before the next one begins, is rarely possible. Research questions may be stimulated [...] by considering ways in which existing measurement tools or data sets can be put to use." (Clarke & Cossette, 2000, p. 111-112)

Hyman's second explanation of the successful reuse of data is the researcher's *style of work* and *distinctive qualities*⁵⁰. He reaches this conclusion by studying some researchers who are "persistent practitioners of secondary analysis and who succeed in making repeated contributions to the literature" (Hyman, 1972, p. 77).

The *style of work* consists of having both different and broad interests, or "a degree of vagueness", as William James has, who is one of Hyman's study participants (p. 78). An appropriate degree of vagueness consists of not having neither too narrow a research purpose or question, nor one that is too broad. The first option would lead us to view all data as not fitting the research question, while the second one would lack specific focus, and thus would impede the identification of the right data to answer the research question. Too much broadness leads to vagueness, which is not a good approach because "direction is focused in all directions at once and their search is unguided by any purpose" (Hyman, 1972, p. 79). Broad research interests bring about serendipity in finding data, which may fit any of the aspects of the broad interests, but also sagacity is necessary to *convert imperfect data into valuable materials* (p. 83). Broadness may also later lead to narrower research goals (Borgman, 2015).⁵¹ However, initial narrow interests could potentially explain the factor *effort of identifying original contributions or unexplored issues from the data* (Curty, 2015; Zimmerman, 2008) that can affect researchers when reusing data.

Third, Hyman explains that the reason why most researchers are not successful with secondary analysis is that they lack the necessary skills in the principles and procedures of the analysis (1972). Other researchers, though having some of these skills, may not be fully trained for the rigorous and tedious work of the whole process and, thus, end up deciding to give it up. Authors in fields other than sociology also consider that knowledge and skills are required to conduct secondary analysis⁵², but also time and patience (Rew et al., 2000)⁵³. This is also aligned with Niu's findings

⁵⁰ "The complete secondary analyst success only by his wits, and he is the pure or ideal type to examine. That he has some distinctive qualities cannot be doubted. Contrast him with the many who fail! Recall that in the survey of potential users of the Berkeley Archive, over half of the initial inquiries never reached the point of acquiring data. Many secondary analyses thus seem to be aborted almost at the point of conception. Interruptions at later stages no doubt reflect a lack of technical skills. At the earliest, or preanalytic, stage, there might be a failure of nerve, an apprehensiveness about the analytic skill one will need and may not have, or insufficient stamina or resources to face practical obstacles. But there is also some failure in thought that stops so many secondary analysts at the point of conception, and, on the other hand, some fortunate turn of mind that ushers in success." (Hyman, 1972, p. 78)

⁵¹ "When first framing a problem, researchers are open to many possible sources of data. As they narrow a problem, they often narrow the scope of their data collection." (Borgman, 2015, p. 59).

⁵² Two examples in nursing: "Although the same research principles used in primary research also apply to secondary data analysis, conducting a secondary analysis requires important knowledge and skills. The nurse scientist must have a good understanding of the existence of rich data sets and how to locate, obtain, and evaluate the data (Aponte, 2010; Garmon Bibb, 2007)." (Dunn, Arslanian-Engoren, DeKoekkoek, Jadack, & Scott, 2015, p. 1299). A researcher who uses an existing data set requires knowledge of general research principles and techniques. Additionally, this researcher must understand concepts that are unique to the challenges associated specifically with analyzing an existing data set. (Doolan & Froelicher, 2009, p. 203-204).

⁵³ "Secondary data analysis using [...] data [...] involves a sound approach to conceptualizing the problem to be studied, having a theoretical framework, carefully delineating the research questions to be answered, identifying concepts and how they are operationalized [...]. Thus secondary data analysis, like other types of research [...] requires

(2009) regarding the fact that it is the reuser's stronger absorptive capacity the one that eases the understanding of data documentation, and thus a potential use of the data more than a reuser's weaker absorptive capacity.

Hyman also mentions a lack of motivation to conduct secondary analysis (1972), but does not expound on it. He mentions that some researchers persevere only to complete secondary analysis, but except for the benefits of secondary analysis that he mentions, namely practical benefits for the researcher (saving on money, time and personnel), social benefits, and benefits for theory and knowledge, he does not provide any explanation of why some researchers do persevere and others do not.

In summary, despite all the knowledge we have gained in trying to understand researchers' challenges, motivations, and facilitators for reusing data, we still do not know why some researchers decide to reuse data and some others decide not to reuse data at all, or abandon data reuse after having commenced. Also we do not know how researchers weigh up the tradeoffs of using secondary data against using primary data (Pasquetto et al., 2017), or even if they do this at all.

In addition to the complexity of comparing findings due to the heterogeneity of aspects of the data reuse process tackled by IS scholars, these studies present two main general shortcomings. On the one hand, although no one can question that the conceptual or theoretical frameworks used are suited to and appropriate for scholars' research questions and ontological positions, none of them, including Curty's use of UTAUT (2015) and Curty et al.'s theory of reasoned action (2017), is sufficient to explain the phenomenon of why and how data reuse happens⁵⁴. On the other hand, the reuse of data is a process that mediates between the need or willingness to address a research gap, or to solve a problem, and the actual filling of the knowledge gap or solution of the problem. The latter is, in turn, the means for researchers to achieve other ends, for instance, to obtain rewards for their scientific contributions or to satisfy their own knowledge curiosity. Although IS scholars recognize that data reuse is not an end in itself (Borgman, 2015), and other scholars from other social sciences and health sciences disciplines implicitly or tacitly acknowledge it. I would argue that, to date, this acknowledgment has not been adequately addressed in the methods used in studies about data reuse.

There is one main reason for this argument about methods. IS scholars have considered two general outcomes –or ends– when studying the reuse of data, namely data reuse happens or does not happen. I argue that this can be methodologically problematic to the extent that IS scholars consider these two outcomes as *ends* in themselves in their empirical studies. This is because these are not the research

that [...] the investigator has the necessary analytic skills to meet these challenges. Working with archival data requires time and patience in understanding the data set to be used, including its strengths and limitations." (Rew et al., 2000, p. 226).

⁵⁴ The fact that the reuse of data is, ultimately, a researcher's decision is acknowledged consistently in most, if not all, IS scholars' contributions. The choice of a single quotation or even a few is too reductionist.

goals of any research inquiry. While the consideration of these two *ends* as research goals can be useful for practical reasons in order to simplify the analysis of the use of secondary data, it blurs our understanding of the process of reusing data understood as a step within a researchers' decision-making process that is embedded in a larger context, namely a scientific inquiry process. The main reason is that data reuse is not a goal *per se*, but a process that mediates between the availability of secondary data and the answer to a research question, which, in turn, is a means to satisfy researchers' curiosity, make a scientific contribution, solve a problem, and obtain rewards that contribute to advancing their careers.

I suggest that it is important to consider researchers' research goals when studying the process of reusing data, because our understanding of the phenomenon may increase. In general, we can summarize a researcher's research goals (satisfy researchers' curiosity, make a scientific contribution, solve a problem, and obtain rewards) using the term "scientific contribution", since this can be understood to involve all of the various goals mentioned. However, a scientific contribution can be both an *expected research goal* and an *achieved research goal*. The former is a goal that researchers plan to achieve and that is set at the beginning of a research project or study. It precedes the process of reusing data. The latter –*achieved research goal*– is the fulfilled *expected research goal* and it follows the process of reusing data. I suggest that researchers' *expected research goal* can help to explain why some researchers decide to reuse data and to keep reusing data despite all challenges they face. Researchers' motivation and willingness to achieve their *expected research goal* of making a scientific contribution could explain why some researchers are successful when doing secondary analysis and others are not after having tried to use secondary data.

Kim and Yoon (2017) suggest that further studies are needed to understand why some researchers face challenges and keep reusing data when the challenges were not always known at the outset of the process of reusing data. These authors suggest that "[p]erhaps the needs of the scientist and the usefulness of the data are more important than any effort that might be required to reuse data" (Kim & Yoon, 2017, p. 2716). This suggestion is aligned with my conceptualization of the reuse of data as a process that mediates between the need or willingness to address a research gap, or to solve a problem, and the actual filling of the knowledge gap or solution of the problem. Expected goals could explain perseverance and motivation that some researchers have in reusing data despite all efforts and challenges they face, and that Hyman (1972) does not finally account for. Actual goals could also explain why some researchers would adopt some of the aforementioned strategies to overcome fitness, and why other researchers would not.

Drawing upon the above discussion on the studies about factors affecting data reuse, namely factors in which the user's subjective cognitive and intellectual assessment play a relevant role, I suggest that data reuse cannot be explained merely by attributes of data infrastructures or the epistemic practices

of particular research communities, as much of the existing literature suggests. A more complex model is required: one that mobilizes a broader range of elements and processes for the empirical studies of data reuse in which researchers' scientific and individual career goals play a relevant role. Furthermore, I suggest that such a model requires a theoretical framework, which considers data reuse as an outcome of a researcher's decision-making processes, and empirical causal methods, which allow us to identify the causal conditions and processes under which data reuse use happens. This can be a long process, in which several changes in decisions may occur, or new decisions taken due to changes in the conditions affecting reuse.

In conclusion, despite all the knowledge we have recently gained on challenges and facilitators affecting researchers when reusing data, both the questions why researchers decide to use secondary data despite challenges and how researchers manage to use secondary data remain largely unanswered.

Chapter 3

Conceptual framework and theoretical strategy

Theories often incorporate ideas and tools from earlier generations of social thought. There is also substantial cross-over as scholars play “arbitrage” and create new insights by linking previously disconnected theory. Theory spillover is evident when one examines the mechanisms offered by empirical researchers. Furthermore, when theories are applied to social life, they require modification or extension because have their limits. To produce a plausible cause-and-effect chain, a researcher may have to combine new and old ideas or borrow from other styles of argument. [...] Sociological theory is instead more like a toolbox or playbook of ideas that are used in practice. (Rojas, 2017)

3.1. Definition of data, primary data, secondary data and reuse of data

In order to define the concept of *data* in this dissertation, I draw upon two main concepts: the *relational framework* by Leonelli (2015, 2016) and the *data stream model* by Hilgartner and Brandt-Rauf (1994), which I complement with Reichenbach’s *context of discovery* and *context of justification* (1938).

According to Leonelli's *relational framework* (Leonelli, 2015, 2016), the function of an object as *data* will depend on who uses data, how and for which purposes. She defines data as

[...] *objects that (1) are treated as potential evidence for one or more claims about phenomena and (2) are formatted and handled in ways that enable its circulation among individuals or groups for the purpose of analysis. In the case of scientific data, these groups will most likely include at least some scientists, although this is not a necessary requirement in my framework.*²⁰ *This definition frames the notion of data as a relational category, which can be attributed to any objects as long as they fulfill the two requirements above. What counts as data depends on who uses them how, and for which purposes. Within this view, the specific format of the objects in question does not matter [...] Also, there is no intrinsically privileged type of data, as judgements on which objects best work as evidence depend on the preferences of the researchers in question, the nature of the claims under considerations, the materials [...] with which they work, and the availability of other resources of evidence.[...] A key implication of this approach is that the same objects may or may not be functioning as data, depending on which role they are made to play in scientific inquiry and for how long. This is particularly significant given the contradictions and uncertainties, evidenced in much scientific and policy literature, about how data should be defined and whether their identity changes whenever they shift format, medium or context.[...] I advocate defining data in terms of their function within specific processes of inquiry accounts, rather than in terms of intrinsic properties. Within this [relational] framework, it is meaningless to ask what objects count as data in the abstract. This question can only be answered with reference to concrete research situations, in which investigators make decisions about which research outputs could be used as evidence and which are instead useless in that regard. (Leonelli, 2016, p. 78-79).*

I agree with Leonelli that data are best understood by considering them in a contextualized inquiry situation, and not in an abstract way. In this dissertation, any object is a datum as long as there is a researcher who treats it as potential or definitive evidence of scientific claims. Leonelli's definition (2016, p. 78-79) includes not only *research data*, which are data collected, produced or used by academics or researchers in research or scholarly settings, and used as evidence of scientific claims, but data collected in other organization settings. Therefore, her definition of data includes objects such as, for instance, government statistics, business records, archival records, administrative data, hospital and health care data. In other words, it includes both research data and data collected in institutional settings other than research or scholarly contexts.

The second part of Leonelli's definition requires that "data are formatted and handled in ways that enable [their] circulation among individuals or groups for the purpose of analysis". I suggest that this part of her definition clashes with process-oriented models of scientific work, whose proponents suggest that data are not static or unchanging objects as Hilgartner and Brandt-Rauf (1994) suggest with their *data stream model*. These authors argue that *data streams* are more faithful to how data are actually managed in practice along the process of a research or scientific inquiry.

In common discourse about science, the normal course of research is often described in terms of what one might call the "produce and publish model": first, a scientist produces data (or "scientific findings"), which are conceived of as the output of scientific production; second, he or she disseminates the data through open publication (or informal communication); and third, the freely-available data serve as an input for other scientists, who evaluate it, certify it, and build on it. Although the produce and publish model is a profound oversimplification, it implicitly underlies much discussion of data-access issues, which tends to frame restrictions on access as departures from the normal course of science. In contrast, ethnographic studies of scientific laboratories suggest the need for a process-oriented model that does not treat data as well-defined, stable entities and that is oriented toward flow and continuity (e.g., Knorr-Cetina 1981,1992; Latour 1987; Latour and Woolgar 1979; Lynch 1985; see also Collins 1985). Drawing on the ethnographic literature, we argue that an alternative "data stream model" should be employed in the analysis of access practices. Data should be conceptualized not as the end-products of research or even as isolated objects, but as part of an evolving data stream. (Hilgartner & Brandt-Rauf, 1994, p. 559)

Hilgartner and Brandt-Rauf (1994) blame the produce-and-publish scientific model for biasing our understanding of data for end-products of the scientific work. In fact, access to data can be possible from the produce stage of the scientific work, not only in private sharing acts, but also in public sharing ways. The latter is the case, for instance, of sequence data of the Human Genome Project (HGP), for which Hilgartner has introduced a neologism, namely *UJAD data*, data unpublished in journals and available in databases (Hilgartner, 2017). The UJAD data are not end-products of the scientific work, but a portion of the data stream of the produce stage. Yet, be data shared from the produce stage or the publish stage, the person that shares or curates the data has to make a decision about the format and the analysis level in which data will be shared for circulation, as Leonelli suggests in the second part of her definition (2016, p. 78). In other words, even if data have not reached yet the status of outputs in the context of a research inquiry, the act of sharing them requires defined and stable forms of data. This is also applicable even in disciplines such as biology there is "a vast choice of file formats in which those data could be stored and visualized" (Leonelli, 2016, p. 76).

I suggest, then, that a better way to reconcile the data stream model with the need of a stable format to circulate data, and to view data as not only end-products is to disregard the produce-and-publish model, and to consider both Reichenbach's *context of discovery* and *context of justification* (1938). With these two contexts, Reichenbach wanted to distinguish between a scientist's way of reasoning and thinking (context of discovery) and a scientist's way of presenting her findings publicly (context of justification). Reichenbach's distinction of the two contexts has been both widely criticized and adopted by other philosophers of science (Leonelli, 2016). However, I suggest that the usefulness of the two contexts applied to data lies in viewing them as two different layers of the data stream creation and circulation. The context of discovery would be the layer in which the data stream is locally collected or created, and managed or used in an organization setting. The context of justification would be the layer in which any portion of the data stream is communicated to other persons in a private or public way.

It could be argued that in some organizational settings or in some scientific disciplines, the distinction of these two contexts, or rather layers, are not clearly distinguishable, and that the format of any portion of the data stream is the same in both contexts. I would agree with these objections. However, I suggest that in many disciplines, data streams in the context of discovery are mostly individual *raw data* or *unprocessed data*. This means that they have not yet been processed or "*combined into complex units*" (Secrist, 1920, p.16) until they are packaged in a processed and aggregated way in order to be moved into the context of justification and, thus, function as evidence of scientific claims. When any portion of the data stream is moved into the context of justification, it is "*collected, tabulated in simple or composite form, and made available for use, but which [is] removed one or more steps from the form in which [it is] reported and consequently do[es] not show [...] the treatments to which [it has] been subjected in analysis, etc.*" (Secrist, 1920, p.16).

In any case, I find Reichenbach's definition of these two contexts useful in order to understand that any portion of the data stream in the context of discovery has to be packaged in some type of format in order to be communicated and/or shared.

Hence, for the purposes of this dissertation, data are any portion⁵⁵ of the data stream, which represents phenomena in any kind of organizational setting, be scholarly or not, and which a researcher or group of researchers regard as potential or definitive evidence of scientific claims in the context of a concrete scientific inquiry, regardless of whether data have been already circulated or not⁵⁶. I have depicted this

⁵⁵ As it can be any portion of the data stream, this means that data could have been only collected, created or captured without having been analyzed or used at all, being "*raw materials, at the upstream end*" (Hilgartner & Brandt-Rauf, 1994, p. 361)

⁵⁶ Any portion of the data stream in the context of justification or in the context of discovery.

definition in Figure 1, where the outer rectangular form represents the concrete scientific inquiry in which data streams exist from a starting point (data collection or production) to an endless one, since data can be moved into the context of justification without being outputs of a produce-and-publish model as it happens with *UJAD* data (Hilgartner, 2017).

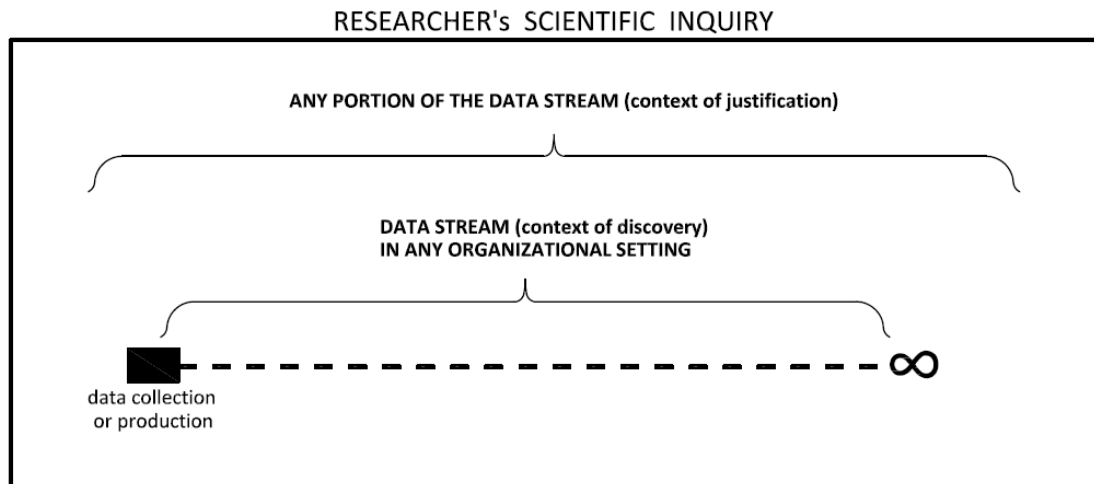


Figure 1 - Definition of data in this dissertation

The difference between *primary data* and *secondary data* does not lie in “what” primary and secondary data are, but in “when” data are primary or secondary and who uses them. Borrowing Niu's words, “data are situational” (2009, p. 5). Time plays a role in differentiating primary data and secondary data, as the latter can only *exist* in a later moment than the former (Hakim, 2013; McAllister, 2018). A different user of the data also plays a role in differentiating primary data from secondary data. So, while *primary data* and *secondary data* are the same data, *primary data* are *data* –as defined above– in a specific moment (t_1) collected or produced by a user (A), and *secondary data* are *primary data* willing to be used or actually being used in a later moment (t_2) by a different user, (researcher B) other than the user in moment (t_1). In other words, *secondary data* are in reference to their use or potential use by a secondary user or secondary analyst in time (t_2).

According to the above definition of *data*, it is worth noting that a researcher in time (t_2) will be able to use only the portion or portions of the *primary data stream* that a person A⁵⁷ decides to circulate in a private or public way. Data circulation can happen in time t_1 , in time t_2 , or in both times t_1 and t_2 . The definitions of primary data and secondary data are depicted in Figure 2, which allows me introduce the term which motivates this research, the use of secondary data or reuse of data, which are used herein interchangeably.

⁵⁷ Note that I have used the term *person* and not *researcher* purposely to insist on the fact that the collection or creation of data in t_1 does not necessarily happen in a research or scholarly setting. Primary data can be collected or produced and/or used in any type of organizational setting.

The use of secondary data is represented by the arrow, by means of which a researcher B⁵⁸ access and uses any portion of the data stream that has been packaged in any format that enables its circulation (context of justification) by person A in time t1 or in time t2⁵⁹. The portion of the primary data stream has not been necessarily used for any purpose by person A. This could question whether “reuse” of data is an appropriate term. I suggest that a more appropriate way to represent the idea that primary data have only been collected and not necessarily used by person A is “(re)use”. Yet, I will use this term without parenthesis in this dissertation for practical and legibility reasons. Also for practical reasons, I will use the terms *data* and *secondary data* interchangeably in this dissertation when referring to their use in time (t2). I will use *primary data* when I want to highlight that these data are in reference to the primary user in time (t1).

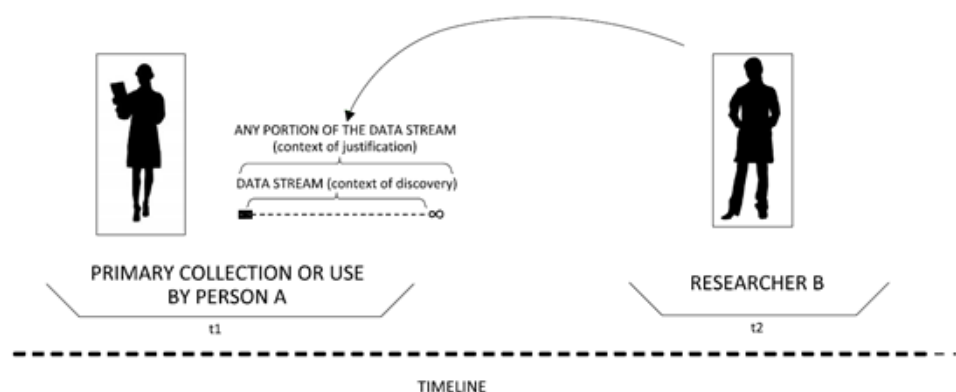


Figure 2 – Visual representation of the definition of reuse of data or use of secondary data

Furthermore, the *use of secondary data* or *reuse of data* (t2) includes any research purpose whether it is the same, similar or different from the primary purpose of collection or use of data in time (t1)⁶⁰. The main reason for considering any purpose for the reuse of data is that we can gain a better understanding on the phenomenon than if we circumscribe the meaning of *reuse of data* to the answer of only novel research questions with secondary data. After all, both replication (validation or reproduction) and the answer to novel research questions do create “new knowledge”.

⁵⁸ Note that I have used the term *researcher* purposely to insist on the fact that the reuse of the data in t2 is done by an academic or a researcher in a research setting.

⁵⁹ The circulation of the data can happen at time t1, time t2 or at both times t1 and t2.

⁶⁰ In most previous research, scholars in Information Science have considered the reuse of data only in cases when the research purpose of the secondary user (t2) is different from the research purpose of the primary user (t1), e.g., R. G. Curty, Crowston, Specht, Grant, & Dalton, 2017; Pasquetto, 2018; A. S. Zimmerman, 2003 (van de Sandt, Dallmeier-Tiessen, Lavasa, & Petras, 2019).

Therefore, and for the purposes of this research, the term use of *secondary data* or reuse of *data* refers to the use of data by someone in time (t2), who is different from the person who collected and/or used the data in time (t1) in any kind of organizational setting, for the same or different purpose than user did in time (t1). The “purpose” has to be understood in a very broad sense⁶¹, and “a person different from” means that the person in time (t2) or secondary user is unconnected or alien to the process of collection or use of data in time (t1). Last, but not least, and as part of the definition of the term, the use of secondary data is not an end in itself, but a means to achieve a goal, namely to make a scientific contribution or to create background knowledge.

For practical purposes, I will use interchangeably reuse of *data*, *data reuse*, *use of secondary data*, and *secondary analysis* in this dissertation. *User*, *primary user* and *primary analyst* are synonyms with regard to data in time (t1). Reuser, secondary user and secondary analyst are synonyms with regard to data in time (t2).

3.2. Rational choice theory, bounded rationality and procedural rationality

As I have argued above, the reuse of data is, after all, a researcher’s decision, which can change⁶² along the whole process of conducting a research study. Although IS scholars have not approached the study of the use of secondary data as a researchers’ decision, -strictly speaking-, some of them have attempted to explain that researchers decide reusing data after outweighing the costs or efforts and benefits of doing it (Curty, 2015). This explanation evokes rational choice theories, which take for granted that human beings are goal oriented, and are able to weigh all pros and cons of every potential choice in order to come up with the best decision for them (Little, 1991).

Rational choice theory, rationality theory or economic rationality, was born in the discipline of economics in the 17th century or earlier. In the 20th century, this theory has been used in sciences studying social behavior and has been widely criticized by both economists and social scientists (Morçöl, 2007b). The main reasons for this criticism are that this theory assumes that when people make decisions (1) they have all information about the different options and (2) know all the options or alternatives; (3) it is easy to evaluate these options and that people choose the best option; (4) people are goal oriented and scheming in order to achieve their goals; (5) context does not interfere in their

⁶¹ For example, data in time (t1) could be collected as part of an administrative process in a health care setting, and be used in time (t2) for answering a research question in order to increase scientific knowledge. Data in time (t1) could be collected as part of a research project in a scholarly setting, and be used in time (t2) for answering the same or different research questions than the project did in time (t1); etc.

⁶² For instance, researchers, who initially decide to use some data, may eventually end up not doing so and vice versa.

decisions; (6) decisions do not depend on other's actions; and (7) all people are equal, and thus, values and social identities do not interfere in their decisions (Little, 1991; Morçöl, 2007b; Rojas, 2017). However, despite these criticisms, rational theory is a kind of ghost or myth still used and applied in decision-making studies:

There were attempts to formulate a universal theory of decision making in the past. The assumption was that the "rational individual" could make decisions in a purely logical fashion (without the interference of values or any other distractions) and with the complete knowledge of the problems to be solved and the consequences of his or her actions. This notion, which is largely regarded as a myth today (some see it as a convenient myth, however) is the bases of the so-called "rational comprehensive model of decision making." The rational comprehensive model is an abstraction. Or, perhaps, it is a ghost, which has no material existence; it has no direct applications in any real-life situations, nor does it have any strong proponents. Some consider it a useful approximation, but not serious theoretical perspective assumes that individuals are (or can be) purely "rational" decision makers. Yet, the rational model is invoked again and again in theoretical debates on decision making. It is a ghost in the middle of the debates – a ghost that refuses to go away, despite the fact that it was criticized to death and buried [...] (Morçöl, 2007, p. 3-4).

In fact, be it a ghost or not, some social scientists outside economics and scholars of philosophy are proponents of the use –though with caution– of rational choice theories because they are useful as a starting point to understand processes usually examined by sociologists and to focus on the instrumentality of human nature (Little, 1991; Rojas, 2017).

Assumptions of rational theory from 1 to 4 (Barros, 2010) were especially criticized by the economist Herbert Simon. Simon introduced some nuances into rational choice theory, which he conceptualized in two terms: *bounded rationality* (Simon, 1955, 1957, 2000), and *procedural rationality* (Simon, 1976, 2000).

Bounded rationality is simply the idea that the choices people make are determined not only by some consistent overall goal and the properties of the external world, but also by the knowledge that decision makers do and don't have of the world, their ability or inability to evoke that knowledge when it is relevant, to work out the consequences of their actions, to conjure up possible courses of action, to cope with uncertainty (including uncertainty deriving from the possible responses of other actors), and to adjudicate among their many competing wants. Rationality is bounded because these abilities are severely limited. Consequently, rational behavior in the

real world is as much determined by the "inner environment" of people's minds, both their memory contents and their processes, as by the "outer environment" of the world on which they act, and which acts on them. (Simon, 2000, p. 25)

Procedural rationality is concerned with “the quality of the processes of decision” (Simon, 2000, p. 25). *Behavior is procedurally rational when it is the outcome of appropriate deliberation. Its procedural rationality depends on the process that generated it.* (Simon, 1976, p. 67)..

While bounded rationality is widespread and well known, procedural rationality is barely mentioned in the literature of economics and rational choice theory (Barros, 2010). However, both serve the purpose to help explain researchers’ decision-making regarding reusing data. This is especially due to the concepts *satisficing* –initially labeled *satisfactory pay-offs* (Barros, 2010, p. 463)– and *search*, which I explain hereinafter. As Barros (2010) explains, both bounded rationality and procedural rationality are two complementary ways of criticizing rational theory. Yet, only bounded rationality has gained popularity and remained among us as a school of thought dissatisfied with economic theories of decision-making or choice making (Jones, 1999).

Bounded rationality is not a way of reasoning itself, but a way of warning us about the limitations that Simon found in real life when studying decision-making in organizations (Barros, 2010). Simon realized that people have a limited or bounded capacity to make rational choice decisions, on the one hand, because of time, cognitive, and information limitations, and, on the other hand, because people, who, although being goal oriented, do not chose the best choice, but one that it is good enough. Simon does not deny that people are goal oriented and that they aim to make rational choices. What Simon asserts is that people cannot be rational most of the times especially in complex situations, and when something important is at stake because of the aforementioned limitations (Mingus, 2007).

In summary, bounded rationality is, though, a static way of looking at decisions in which what matters is only the outcome or the choice done. So, Simon complements bounded rationality with procedural rationality. With the latter, he introduces two factors or variables influencing the final choice: the process and the agent of the process (Barros, 2010). For Simon, the outcome or final choice is strongly dependent on how the decision maker searches and finds alternatives in a sequential or procedural way. The decision maker does not have all alternatives into account and then makes a choice. Instead, what happens is that when the decision maker –being in the process of searching– finds a *satisficing* or good-enough option, stops searching. At this point when he stops searching, the decision maker has made a choice, which most of times is not the goal that she had in mind initially.

Alternatives can be sequentially found out, by search processes, search being interrupted when a satisfactory alternative is found. Satisficing is, hence, the theoretical step that allows Simon to abandon the idea of rationality as a tautological reasoning over given premises, which permits rationality to operate in an open, not predetermined, space. On the other hand, satisficing forces him to inquire into the process by which such premises are built by the agent. (Barros, 2010, 463)

Simon's research in the area of cognitive science, demonstrated that, in complex situations, the choice taken, its result, strongly depended on the particular process that generated it, and not only on the objectives that oriented it. Hence, it becomes indispensable to know the process by which the choice is taken. (Barros, 2010, p. 465)

Bounded rationality and Simon's little sister procedural rationality –borrowing Barros (2010) metaphor of the unpopularized concept– are yet rooted in rational choice theories. However, there are some issues that affect human's decisions or choices that were not addressed by Simon's critiques, but require some relevant considerations when studying social action and decision making, either as a process –choosing– or as an outcome –choices–. As aforementioned, some of these issues that rational theories ignore are, among other things, that context has an effect on the process of choices; decisions depend on other's actions and broader decisions; and people are different, and thus, their values and social identities do interfere in their decisions.

Rational choice theories, including bounded rationality and procedural rationality, do not consider influences of the choice maker's context in the decision-making process or outcome (Little, 1991; Rojas, 2017). Yet, nowadays, it would be rare to find scholars who would not agree with the fact that “*autonomous action is not the property of individuals but of complexly-ordered social relations*”, borrowing López' words when referring to the dependency of actions and decisions within a broader context (López, 2004, p. 879). Certainly, some scholars studying decision-making explain, among other things, how context influences the decision-making process, content, and outcomes in organizations and how small decisions are sometimes made within broader decisions (Nutt & Wilson, 2010b)

First, decision making is a complex process. [...] Decision making is also a multilevel process, with smaller decisions typically nested within larger decisions, which may themselves be part of larger group projects (McGrath and Tschan, 2004). “Every decision involves a series of activities and choices nested in choices of broader scope,

rather than a single simple choice” (Poole and Hirokawa, 1996: 9). Moreover, it is often difficult to understand a single decision without considering larger issues and prior decisions and without grappling with relatively fuzzy boundaries of the larger issues (Tracy and Standerfer, 2003) (Poole & Van de Ven, 2010, p. 544)

Although the father of the sociology of science, the American Robert K. Merton, argued that “science is an autonomous sphere of activity, able to resist external influences; it defends and champions the principles of independence, discipline and pure rationality” (Vinck, 2010, p.7), following sociology of science studies have persuasively shown that research does not happen in a vacuum. Research takes place in organizations (universities, research institutes, laboratories, etc.) that have external influences, both explicit and implicit behavior norms in order to fulfill their missions, and where power relationships take place (Vinck, 2010). Therefore, researchers are not autonomous and independent decision-makers, and processes and outcomes of knowledge creation depend on a broader context over which sometimes the researcher has little or no control (Latour, 1987; Vinck, 2010). Hilgartner & Brandt-Rauf (1994), when referring to Knorr-Cetina’s idea of the researcher as a “socially situated reasoner”, highlights the many links to other people and organizations that researchers usually have to construct and maintain in order to obtain resources. Thus, it would be mistaken to think that decisions on data reuse, as well as decisions on data sharing⁶³, which are nested in broader decisions, are not context-contingent.

During the course of a research process, a significant number of random factors are at play, whether they relate to the research itself (non-materialization of the expected results), local conditions (instrument malfunctions or unavailability, interference with the work of another scientist) or external resources (failings or a change of strategy on the part of an associate, redefinition of funding priorities) (Vinck, 2010)

Nonetheless, despite sociology of science studies having shown that feelings, beliefs, intuitions, etc. affect research and knowledge creation, researchers’ personal circumstances affecting research careers and *vice versa* are conspicuous by their absence. Researchers live embedded in different but interrelated worlds, the personal and the professional one. Both actions and decisions in one world may affect actions and decisions in the other world, at least in terms of setting career goals, including acquiring tenured positions, research career stages, maternity or paternity leaves, or a partner’s professional career. Thus, I suggest that researchers bear in mind personal actions and decisions when setting their professional goals.

⁶³ “For researchers, the inclination toward data sharing is context-dependent. Variations in institutional support, the available technological infrastructure, and interactions with other researchers are all factors that affect researchers’ desire and ability to make their data available to others [15,11].” (Tenopir et al., 2015, p. 3)

Furthermore, and contrary to what rational choice theories defend, people do not always make explicit comparisons of alternatives and of the value that each alternative has, and people rarely choose the best or most self-beneficial alternative. Rather, people usually respond to automatic personal intuitions, beliefs, personal values or feelings, but also to norms or relationships embedded in larger structures such as institutional and organizational norms, power relationships, social categories, inequality, gender, race, etc. as Rojas (2017) explains. The marginalization of arguments about how larger structures silence explicit comparisons of alternatives is rooted in one of the assumptions of rational choice theories since these theories focus on the individual decision-maker (Rojas, 2017). Thus, for example, it may happen, that a researcher belonging to a discipline that relies mainly –if not exclusively– on secondary data, i.e., computational biology, do not compare consciously the alternative of collecting primary data with the alternative of using secondary data because the epistemic norms of her discipline prevent her from doing it. These epistemic norms are embedded and ingrained in the researcher, and thus I suggest casting doubt on the existence of a reasoning mental exercise weighting up the efforts of using primary data against the efforts of using secondary data. Other similar example of decision embedded in epistemic norms could be the one by a researcher belonging to a discipline where secondary data are hardly accepted as evidence of scientific claims. Thus, this researcher would design her study with own collected primary data in a way that she can timely reap the benefits of her study, that is, make a scientific contribution and obtain rewards for it, despite having secondary data available for her study. In this complex situation, it may be better to produce primary data regardless of efficiency considerations of using secondary data. These two examples of decisions may be conditioned not only by epistemic considerations but also by expectations and norms of research groups, organizations or scientific communities, and by resource considerations, etc.

These two potential examples of decisions could be explained by Townley's *institutional rationality* (2011), which she defines as the “rationality [that] recognizes that the rationality of action is informed by the institutionally grounded, historically evolved, value sphere in which [the action] takes place” (Townley, 2011, p. 113). Institutional rationality would explain why a researcher in computational biology would not consider the option of collecting primary data at all. While her choice, only based on secondary data, is something that could be viewed as *irrational* in other disciplines, it is *rational* in her own discipline. This institutional rationality could be also applied to the researcher that decides collecting primary data, when she has the opportunity to use secondary data to answer her research question.

The fact that people do not always make explicit comparisons to make a decision, and that decisions are embedded in larger contextual situations leads also to the problematic question of whether decision-making really takes place in real life or people do simply act. Tsoukas (2010) explains that the problem lies in whether we consider *the language of the actor* or *the language of the observer*.

When participants or actors talk retrospectively about their actions, they adopt the language of the observer and refer to their actions as decision-makings. However, some scholars argue that participants just act, but do not make decisions. Decisions are conceptual inventions that are believed to always precede action.

'Decision' is observers' construct rather than actors' experienced reality: 'decisions often do not exist; they are merely constructs in the eyes of the observer', note Langley et al. (1995: 265). [...] As Chia (1994: 794) remarks: 'Understanding decision-making as an explanatory principle involves a recognition that it is the product of a post-hoc rationalization process in which the cause/effect relationship established has been abstracted, reified and chronologically reversed. "Decision-making" is a conceptual invention but one which has been reified and chronologically inverted so as to appear as "event" precede action.' (Tsoukas, 2010, p. 381-382)

Karl Weick (1995) shares this latter view and contends that decisions do not exist as a process of choosing, but as an act of interpretation of action⁶⁴. However, we should take Weick's point of view with caution because most of his work is focused on choices that people have made in dramatic and urgent situations, for example, The Man Gulch Disaster (Weick, 1993). I suggest that both cases happen in real life: sometimes human beings act without making prior explicit choices, and sometimes human beings consider alternatives before acting, depending on the varying circumstances, and on the time available to act or make a decision.

Values, beliefs, feelings, intuitions, desires, etc. are not either part of rational choice theories. They are not usually considered in organization studies either in general, nor in science studies in particular because they are considered the *non-rational* side of the human beings (Townley, 2011). Merton, for example, firmly proclaimed that rationality and emotional neutrality is some of the norms to be met by scientists (Vinck, 2010). However, some outstanding research in the sociology of science, for example, *Science in Action* by Latour (1987), gives us examples of how feelings and "gut feelings" (p. 181) play a crucial –if not decisive– role in researchers' choices. I suggest that researchers' values, beliefs, etc. play an important role in the two personal traits that researchers must have to be successful secondary analysts, namely motivation and persistence (Hyman, 1972).

⁶⁴ "[W]henver people are said to make a decision, what really happens is that they are working retrospectively. When one feels compelled to declare that a decision has been made, the gist of that feeling is that there is some outcome at hand that must have been occasioned by some earlier choice. Decision making consists of locating, articulating, and ratifying that earlier choice, bringing it forward to the present, and claiming it as the decision that has just been made. The decision actually has already been set in motion before people declare that it has been made. The recent history is viewed in retrospect, with tentative outcomes in hand, to see what decision could account for that outcome. That plausible decision is the decision people announce. What is crucial about this is that a decision is an act of interpretation rather than an act of choice." (Weick, 1995, p. 184-185)

In summary, I suggest that studies of researchers' decision-making should take into consideration all these issues – context, goals, values, broader decisions, feelings – because they can contribute to shedding light on our understanding of researchers' decisions, including when making decisions about research resources such as primary data and secondary data. In following my own suggestion, below I present a heuristic model for understanding and analyzing scientists' decisions and behavior when working.

3.3. A model of the scientific actor's behavior and decision-making: the *bounded individual horizon (BIH) model*

Drawing upon some studies in the sociology of science and science policy studies, we know that science evaluation systems may influence researchers when making decisions involving variables including research topics, methods, resource choices, research collaborations, types of publications, organizational issues and publication language. For instance, Hammarfelt & De Rijcke (2015) studied the effect of research evaluation systems on researchers' publication practices, disciplinary norms, and individual working routines. Laudel (2002) studied the relationship between research collaborations and rewards. Butler (2003) found that, after the introduction of a new criterion for distributing funds to universities, researchers increased their journal publication productivity in SCI (Science Citation Index). Moore, Newman, Sloane, Steely, & Corp, (2002) studied how productivity changes after research assessment exercises⁶⁵. The main underlying reason of the evaluation systems' influence on researchers' decisions is the potential rewards that they can obtain depending on the results of the evaluation. These rewards can take any form, for instance, resources, funding, prestige or prizes, for example.

Rewards, in their broadest meaning and in any form, usually depend on the individual researcher's performance, which is mainly assessed quantitatively by their outputs, namely scientific contributions. The issue of rewards, which is essential for the development of researchers' career, has been overlooked in previous studies as a potential cause of the use of secondary data. Researchers, apart from their will to advance knowledge, to satisfy their curiosity and to contribute to their scientific community, do ultimately want and need rewards. However, I suggest that researchers, like all human beings, have limited time to achieve their goals, hence they do not maximize their goals but accept options that are good enough or *satisficing*, which is a better option than optimizing (Simon, 2000). All their goals, for example, satisfying their scientific curiosity and obtaining rewards in the short or long term, depend on their capability to make a scientific contribution. Failure to make scientific

⁶⁵ For more examples, see de Rijcke, Wouters, Rushforth, Franssen, & Hammarfelt (2016). These authors summarize some of existing literature that shows how assessment systems have an effect on the production of knowledge.

contributions means their research career could be at stake. In many scientific disciplines, and most of the time, researchers need data in order to make scientific contributions, although sometimes data are used only for creating background knowledge or proving background context to a new study (Pasquetto, 2018; Wallis, Wynholds, Borgman, Sands, & Traweek, 2012). Thus, researchers' decisions regarding resources, namely primary data and/or secondary data, could be considered a nested decision within a larger decision, which consists of making a scientific contribution with the ultimate goal of achieving rewards. This type of goal-oriented decision or action is what Weber called "instrumental action" differentiating it from other types of action.

[S]ome actions are pursued for their own sake (value-rational action); others may satisfy an emotional need (affectual action); and yet others may be done out of habit (traditional action). [... Others] consider the costs and benefits. Weber called this behavior "instrumental action" in that it is goal directed and the goal is not merely habit, affect, or perceived intrinsic moral value. (Rojas, 2017)

However, I suggest that Weber's several types of actions are not mutually exclusive. Rather, they can coexist simultaneously since researchers can pursue one unique goal based on all or only some of these types of actions. In case I may be misunderstood, I am not arguing that researchers are selfish. Rather, I argue that rewards are the only way for them to achieve further milestones in their research careers, and thus to develop them.

Based on Hyman's explanation on the style of work needed to be successful secondary analysts (1972, p. 77-94), I suggest that research career milestones and, thus, scientific contributions, may be the main causes that keep researchers motivated and persistent in completing their studies with secondary data (Hyman, 1972). This suggestion is underpinned by causation in decision-making process theories, in which X (cause) does not lead to Y (effect). Conversely, it is Y (effect) that triggers and impels X (cause).

In Mohr's (1982: 59) terminology, process theories incorporate a 'pull-type causality: X [the precursor] does not imply Y [the outcome], but rather Y implies X'. The purpose or form that is to be realized is what drives the process. (Poole & Van de Ven, 2010, p. 548-549)

From the four types⁶⁶ of decision-making process theories that Poole and Van de Ven (2010) propose, I suggest that *teleological theory* is the one which best represents researchers' actions and decisions, since it is the goal of a scientific contribution that "puts the process [of using secondary data] in motion" (p. 551). These authors categorize this decision-making theory under a *constructive* cyclic

⁶⁶ Evolutionary, dialectical, life cycle and teleology (Poole & Van de Ven, 2010, p. 550-557)

mode of change in which the goal is formulated or envisioned at the outset and the sequence of actions surfaces from the cycle. Actions or goals may be modified along the cyclic sequence depending on what the decision-maker finds as she acts.

A teleological [in italics in the original text] process views decision making as a cycle of goal formulation, implementation, evaluation, and modification of actions or goals based on what was learned or intended by the entity. This sequence emerges through the purposeful enactment or social construction of an envisioned end state among decision makers. [...] In a teleological theory setting a goal in response to a perceived problem or opportunity puts the process in motion. The unit is assumed to be purposeful and adaptive; by itself or in interaction with others, it constructs an envisioned end state, takes action to reach it, and monitors its progress. Thus, teleological theories view development as a repetitive sequence of goal formulation, implementation, evaluation, and modification of goals based on what was learned or intended by the unit. Teleological processes are goal driven, and hence the developmental path followed by the decision-making unit is not predetermined, but is generated by activities necessary to get to a decision. Since there are many ways to get to a decision, multiple paths are possible and there is no present sequence of stages or steps. While a number of teleological theories define steps or stages, there are multiple paths through these steps and the path is determined by exigencies that arise during the process as problems to be solved by the developing unit. (Poole & Van de Ven, 2010, p. 551-552)

I argue that, in any case, the completion of a research study cannot be longer or overtake researcher's self-assigned deadline for obtaining her expected rewards. Otherwise, the researcher may lose her motivation for completing the study, and for making the scientific contribution. I suggest that a scientific contribution and its potential rewards determine ultimately decisions on a study to be completed, and thus on the resources needed to complete it.

As I have expounded above regarding the criticisms of rational choice theory, researchers, when envisioning scientific contributions, make decisions that are nested in larger decisions related to their career goals and influenced by institutional and organizational norms, and power relationships. Their decisions are also influenced by their own personal values, feelings, and personal and familiar situations.

I call this model of the scientific actor's way of working and decision-making the *bounded individual horizon (BIH) model*. In this model, researchers self-allocate a goal –a scientific contribution or a career milestone –, which they expect to achieve within a limited period and with a limited amount of

available material⁶⁷ and cognitive⁶⁸ resources by keeping in mind both their personal and professional situations, their values, beliefs, and feelings, their discipline's epistemic norms, and the reward system they belongs to. This model is bounded for two reasons. On the one hand, researchers have a bounded capacity to self-allocate the best scientific goal. Thus, they may only self-allocate one that is good enough or that *satisfices* them (Simon, 1955, 1957, 2000). On the other hand, researchers have a bounded capacity to foresee or calculate the actual efforts and resources that they will need in order to make a scientific contribution or achieve a career milestone within the limited period.

⁶⁷ Funding, human resources, equipment, data, etc.

⁶⁸ Her own skills and knowledge

Chapter 4

Methodology and methods

"Social phenomena are complex." As social scientists we often make this claim. Sometimes we offer it as justification for the slow rate of social scientific progress. According to our collective folklore there are many, many variables—too many to specify—affecting the phenomena that interest us. Consequently, our explanations are often inadequate. This folklore implies that social phenomena are inordinately complicated and that it is surprising that anyone knows anything about social life.

Yet this depiction of social life does not fit well with experience. We sense that there is a great deal of order to social phenomena—that there is method to the madness. In fact, it is our strong sense that social phenomena are highly ordered that keeps us going. What is frustrating is the gulf that exists between this sense that the complexities of social phenomena can be unraveled and the frequent failures of our attempts to do so. The complaint that social phenomena are complex is not so much an excuse as it is an expression of this frustration.

This sense of order-in-complexity is very strong in comparative social science because it is not difficult to make sense of an individual case (say, a general strike) or to draw a few rough parallels across a range of cases (a number of general strikes separated in time and space). The challenge comes in trying to make sense of the diversity across cases in a way that unites similarities and differences in a single,

coherent framework. In other words, it is often impossible to summarize in a theoretically or substantively meaningful way the order that seems apparent across diverse cases.

The problem of identifying order-in-complexity has two general forms. One is the identification of types of cases—the problem of constructing useful empirical typologies. [...]

The other characteristic form of the problem of order-in-complexity concerns the difficulty involved in assessing causal complexity, especially multiple conjunctural causation. When an outcome results from several different combinations of conditions, it is not easy to identify the decisive causal combinations across a range of cases, especially when the patterns are confounded. (Ragin, 1987, p. 19-20)

I *guess* that the design of this thesis has been mainly conducted from a *critical realism* perspective, which is a branch of philosophy that differentiates between a real world and an observable world. This inquiry paradigm takes for granted that social phenomena exist outside our minds, and that we – researchers– can search and find a causal explanation of events as well as try to find evidence that the causal explanation is present in each entity or event (Hartwig, 2007; Mathew B Miles & Huberman, 1994; Sayer, 2000). However, I say “I guess” because although ontologically it is relatively easy to place ourselves in one the research paradigms (Guba & Lincoln, 2005), it is not so easy to do so methodologically. Miles and Huberman (1994) argue that methods are a “continuum between “relativism” and “post-positivism”” in which it is nearly practically impossible to situate ourselves in a fixed point, and that it makes no sense to use or boost the infertile discussion of qualitative-quantitative methods. Instead, we should distinguish between analytical and systemic analysis, in citing Salomon, or do variable-oriented studies or case-oriented ones, respectively, in borrowing Ragin’s distinction.

In a deeper sense, as Salomon (1991) points out, the issue is not the quantitative-qualitative at all, but whether we are taking an "analytic" approach to understanding a few controlled variables, or a "systemic" approach to understanding the interaction of variables in a complex environment." (M. Miles & Huberman, 1994, p. 41).

Therefore, I prefer saying that I use a systemic approach in trying to understand and to explain causally the phenomenon of the use of secondary data. I use qualitative instruments to collect the data and qualitative methods to analyze the data since a “realistic understanding of causality is compatible with the key characteristics of qualitative research” (Maxwell, 2004, p. 3).

4.1. Research questions and an approach to answer them

The formulation of research questions may precede, follow, or happen concurrently with the development of a conceptual framework. They also may be formulated at the outset or later on and may be refined or reformulated during the course of fieldwork. (Miles, Mubermann, & Saldaña, 2014, p. 25)

Two research questions guide the empirical part of this dissertation:

1. Why do researchers decide to reuse and keep reusing data despite the challenges they face?
2. How do researchers manage to reuse data despite the challenges they face?

As I have suggested above, on the one hand, the reuse of data cannot be explained merely by the epistemic practices of particular research communities or by attributes of data repositories or infrastructures. On the other hand, a causal approach would allow us to have a deeper understanding, and an explanation of why and how data reuse happens. Therefore, causes are prioritized in this study of the use of secondary data. This forces us to distinguish between causes and factors. In this dissertation, a *cause* is a necessary condition –be it sufficient or not– that produces an effect, and a *factor* is a facilitating condition that influences the effect. In other words, if we eliminate a cause, it will eliminate the effect, and if we eliminate a factor, the effect will not disappear (Meltzoff, 1998). This distinction between causes or necessary conditions and factors or facilitating conditions may be the basis of a useful heuristic method for developing a more contextualized and empirically grounded model for studying data reuse. However, causality is not necessarily explained only by knowing the necessary conditions for something to happen. I suggest that, in order to understand researchers' decision making with secondary data, a different view of causality is needed, one that aims, not only to identify which conditions are necessary for something to happen, but to track how the different conditions and elements are related in such a way that the effect happens.

This allows me to introduce a controversial philosophical topic in science in general, and in social sciences in particular, which is causality, or rather, the explanation of causality. Explanation of causality in social life is one of the main goals of sociology, and of all social sciences in general. Satisfactory explanations of social-life causality are “plausibl[e] connect[i]ons [of] different

observations of the social world into a logical chain of cause and effects” (Rojas, 2017). However, sometimes, we cannot provide satisfactory explanations because of how we view and explain social life. We are unable to identify the logical chain from a cause (X) to an effect (Y) maybe because our tendency⁶⁹ to see things simpler than they are, but also maybe because our research methods are conditioned by our ontological view of social events.

There are three main ontological causal views of social events. One is *inductive regularity* (IR), also known as *general linear reality* (GLR), which has materialized methodologically in the *general linear model* (GLM) that has been adopted as a standard method⁷⁰ in social sciences disciplines to explain social life (Abbott, 2001). The GLM limits our sound understanding of social life by treating and explaining complex social issues as simple ones reducing causality to the regular association of X and Y. This ontology views causality as patterns of regular association in a probabilistic way. Opposite to the regular and probabilistic view is the ontological mechanistic and deterministic view of causality. In this view, causation is explained by the theoretical process(es) by which X produces Y (Beach & Pedersen, 2013) as a causal mechanism (CM). A third ontological causal view of social events is the *necessary and sufficient condition* (NSC) one. This view presupposes the idea that causes are necessary conditions for an event to happen and that some set of conditions is sufficient for an event to happen (Little, 1991). For Little (1991) these three ontological views complement each other, with the mechanistic and deterministic view being the most relevant one in their relationship⁷¹.

I take the mechanistic and deterministic view to answer the research questions that guide this dissertation. In doing so, I use *mechanisms* because they are an appropriate tool for providing satisfactory explanations of the forces by which and how X produces Y (Beach & Pedersen, 2013),

⁶⁹ Feltovich et al. contend that most human beings have a cognitive tendency to think and learn that is biased to impede a sound understanding of the world. They call it the *reductive bias*, which “is a tendency for people to treat and interpret complex circumstances and topics as simpler than they really are, leading to misconception, as well as to error and to limitation in knowledge use due to inertness” (Feltovich, Spiro, & Coulson, 1997, p. 128). Klein et al. make a summary of the dimensions of the *reductive bias* (Klein, Phillips, Rall, & Peluso, 2007, p. 121), which I reproduce here: *We define continuous processes as discrete steps; We treat dynamic processes as static; We treat simultaneous processes as sequential; We treat complex systems as simple and direct causal mechanisms; We separate processes that interact; We treat conditional relationships as universals; We treat heterogeneous components as homogeneous; We treat irregular cases as regular ones; We treat nonlinear functional relationships as linear; We attend to surface elements rather than deep ones; We converge on single interpretations rather than multiple interpretations.* (p. 121)

⁷⁰ “Many sociologists treat the world as if social causality actually obeyed the rules of linear transformations. They do this by assuming, in the theories that open their empirical articles, that the social world consists of fixed entities with variables attributes; that these attributes have only one causal meaning at a time, that this causal meaning does not depend on other attributes, on the past sequence of attributes, or on the context of other entities. So distinguished a writer as Blalock has written “These regression equations are the “laws” of a science. To say this is to reify an entailed mathematics into a representation of reality.” (Abbott, 2001, p. 59)

⁷¹ “What are the relations among these conceptions of causation? I will hold that the causal mechanism view is the most fundamental. The fact of a correlation between types of events is evidence of one or more causal mechanisms connecting their appearance. This may be a direct causal mechanism— C directly produces E— or it may be indirect— C and E are both the result of a mechanism deriving from some third condition A. Likewise, the fact that C is either a necessary or sufficient condition for E is the result of a causal mechanism linking C and E, and a central task of a causal explanation is to discern that causal mechanism and the laws on which it depends.” (Little, 1991)

and because they can illuminate some discoveries in an already studied phenomenon (Machamer, Darden, & Craver, 2000). This ontological view of causality might also help in reducing the complexity of factors that affect the process of reusing data, and thus in understanding this phenomenon.

Mechanisms or *causal mechanisms* go back to the end of 60's and beginning of 70's of last century, including the work of Merton in social theory and Whitley in the sociology of science (Gläser, 2012), but have been also used in non-social scientific fields, for example, molecular biology and neurobiology (Machamer et al., 2000). Contestation about the meaning and usefulness of mechanisms, also called *social mechanisms* by some authors in social sciences, arise not only from the many different understandings of the concept, but also from some philosophical constraints of a mechanism with regard to causality (Gerring, 2010; Hedström & Ylikoski, 2010; Machamer et al., 2000; Mayntz, 2004).

There are myriad definitions of a mechanism by scholars in the field of sociology, for example, Beach & Pedersen, 2013; Elster, 1989; Hedström & Swedberg, 1998; Hedström & Ylikoski, 2010; Little, 1991; Maxwell, 2004; Mayntz, 2004; Rojas, 2017; Sayer, 2000, 2010; etc.

In theoretical terms, we can say that mechanisms are theoretical models, which can explain phenomena at a middle level, and thus let us build middle-range theories in order to concentrate on commensurable aspects of social life. The need for middle-range theories is that “grand theories” cannot explain the whole complexity of the social realms (Beach & Pedersen, 2013; Gläser, 2012; Martin, 2012). In other words, a mechanism is a pragmatic intellectual practice that can compensate the incapacity of a grand theory to capture the complexity of the social world (Rojas, 2017). In practical or empirical terms, a mechanism is “a sequence of causally linked events that occur repeatedly in reality if certain conditions are given and link specified initial conditions to a specific outcome” (Gläser, 2012) or “the building block[...] to construct explanations of actual events” (Elder-Vass, 2010, p. 169).

4.2. Researchers' decisions based on the BIH model when reusing data: the data-reuse mechanism

Based on the *bounded individual horizon* (BIH) model of the scientific actor's ways of working, I suggest that we have to consider researchers' final goal, namely a scientific contribution or a career milestone in order to understand researchers' decisions when using secondary data. I suggest doing this by theorizing a decision-making mechanism, namely the *data-reuse mechanism*, as a plausible causal explanation of why and how the reuse of secondary data –used as evidence of scientific claims– happens. Mechanisms have been used previously to theorize on rational choice decision-making, e.g., Oneal (1988). Mechanisms are usually theorized from an inductive approach in the empirical work.

However, it is also possible to identify plausible mechanisms from existing theory (Beach & Pedersen, 2013). The *data-reuse mechanism* is a plausible causal explanation that I have theorized mainly from the *bounded individual horizon* (BIH) model, and partly from findings of previous empirical work included in the literature review. The concept of *satisficing* (Simon, 1955, 1957, 2000) plays a significant role in this mechanism.

In order to disclose researchers' decisions about the use of secondary data, I use Sayer's structure of causal explanation (Sayer, 2010) because he provides a comprehensive and easy guide to its use in practice (Easton, 2010). For Sayer (2010) a causal explanation has the following elements and structure. An *object* belongs to a *structure* by virtue of *internal or necessary conditions*. This very object has both *causal powers* and *liabilities*. Causal powers or active powers are "capacities to behave in particular ways, and causal liabilities or passive powers, that is, specific susceptibilities to certain kinds of change" (Sayer, 2000, p. 11). Both causal powers and liabilities act not by virtue of merely existing, but by the activation under some *conditions*, which are not necessarily inert, but can be other objects having their own causal powers and liabilities (Sayer, 2010). Some authors have suggested that this can lead us to an explanatory infinite regress. However, this is not necessarily the case, because "for a mechanism to be explanatory it is not required that the entities, properties, and activities that it appeals to are themselves explained" (Hedström & Ylikoski, 2010, p. 52). Last, but not least, the *events* are the outcomes of the mechanism, that is, what we investigate. Below in Figure 3, I have reproduced exactly Sayer's visual representation of the structures of a causal explanation (Sayer, 2010, p. 74).

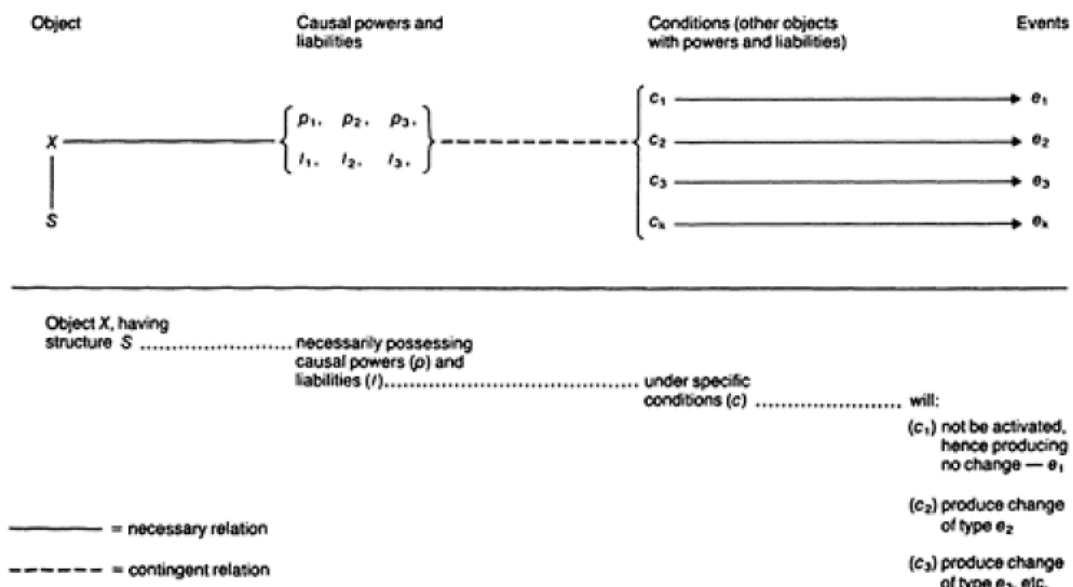


Figure 3 - Source: Figure 7 – The structures of causal explanation (Sayer, 2010, p. 74)

In both his account and his representation (Figure 3) of a mechanism or structure of causal explanation, Sayer (2010) distinguishes between internal or necessary relations and external or contingent relations –contingent meaning here accidental or fortuitous. This distinction is essential for identifying the causal powers and liabilities of objects operating within a mechanism, as well as for realizing that conditions are independent of the objects. In an external or contingent relation, objects can exist without the other, while in an internal or necessary relation objects cannot exist without the other⁷². Therefore, the relationship between objects and causal powers and liabilities is internal, while the relationship between objects and their conditions are external.

There is a relevant aspect of mechanisms, among other aspects⁷³, that should be considered for understanding how mechanisms function and which types of effects they produce. This aspect is their temporal dimension, which is two-fold according to Beach & Pedersen (2013). When citing Pierson, these authors suggest that when theorizing a mechanism, we should keep in mind, “the length of the time within which the mechanism is theorized to be acting and the time horizon of the outcome” (Beach & Pedersen, 2013, p. 55). These authors differentiate between short and long time horizons of the mechanism producing the outcome, and short and long time horizons of the outcome. For instance, an outcome can become only apparent after a mechanism’s long time of period (for instance, institutional change). Short-term mechanisms may have outcomes that may have a bigger impact or outcome in a cumulative way (for instance, climate change). However, it is not easy to deduct from their text what they mean by *a short* or *long period*. Ambiguity aside about Beach & Pedersen (2013) consider a short or long period of a mechanism or of the effect of a mechanism, I argue that we have to consider both the temporal dimensions of the data-reuse mechanism for both theoretical and methodological reasons.

I have theorized the data-reuse mechanism in such a way that, in order to function, the researcher has to have necessarily the expectation of achieving a career milestone, for instance a PhD degree or a tenured position, or making a scientific contribution in a *satisficing* period of time. Based on Beach & Pedersen (2013), I suggest that the data-reuse mechanism works in relatively short time horizons most of the times when the researcher’s expectation is a scientific contribution. I am not considering here

⁷² “Another useful distinction can be made between external, or contingent relations and internal or necessary relations. The relation between yourself and a lump of earth is external in the sense that either object can exist without the other. It is neither necessary nor impossible that they stand in any particular relation; in other words, it is contingent. (Note that this sense of contingent is quite different from that common in everyday uses where ‘contingent upon’ means ‘dependent upon’.) Although a relation may be contingent it may still have significant effects; thus people may break up lumps of earth or be buried beneath them—but the nature of each object does not necessarily depend on its standing in such a relation. By contrast, the relation between a master and a slave is internal or necessary, in that what the object is is dependent on its relation to the other; a person cannot be a slave without a master and vice versa. Another example is the relation of landlord and tenant; the existence of one necessarily presupposes the other.” (Sayer, 2010, p. 60)

⁷³ For example, the type of theoretical explanation, the analytical level of the mechanism, and the extent of the scope conditions of the mechanism (Beach & Pedersen, 2013, p. 52-56).

the time that a scientific contribution, e.g., article, book, etc., takes to be published, since it sometimes may take years after acceptance by journals in some disciplines. When the researcher's expectation is a career milestone, I suggest that the time horizon in which the data-reuse mechanism acts is a long one. In long-time horizon mechanisms, it is more probable that researchers may change their decisions with regard to the use of secondary data more often, as I have discussed when situating *bounded rationality* (Simon, 1955, 1957), and *procedural rationality* (Simon, 1976) when explaining the *bounded individual horizon* (BIH) model of the scientific actor. However, several scientific contributions may have also a long-term impact, since advancement in a research career depends sometimes on an incremental research activity, which implies several, if not many, cumulative scientific contributions and maybe a few previous career milestones. For example, a career milestone such as a PhD degree can be obtained after a relatively long period (between 3 or 6 years) depending on the discipline, while other career milestones are the cumulative effect or result of several data-reuse mechanisms, for example, a tenured professorship position.

Sayer (2010) also suggests that the time horizon of a mechanism has to be considered in a relevant way since both conditions (in external relation with objects) and causal powers and liabilities (in internal relation with objects) do change over time. Several changes in conditions and causal powers and liabilities are more probable to happen in long-time horizon mechanisms, in which their effect or outcome takes time to be evident (Beach & Pedersen, 2013). While the possibility of change over time may be obvious for the contingent conditions, it may be not so obvious for an object's causal powers and liabilities. However, the nature of an object can change, and thus its causal powers and liabilities will do consequently⁷⁴.

So, change over time of both causal powers and liabilities, and conditions have to be considered when testing and theorizing a mechanism, since the operation of the same mechanism can have different effects or results depending on changes of the conditions, and thus on researchers' decisions. On the contrary, under the same conditions different mechanisms can operate to lead to the same results (Easton, 2010; Hedström & Ylikoski, 2010; Sayer, 2010). Some authors, when referring to the conditions under which a mechanism functions, use the term *context*, defined "as the scope conditions that are necessary for a given mechanism to function" (Beach & Pedersen, 2013, p. 54).

In disciplines where scientific contributions depend on data, the event we should be interested in is secondary data used as evidence of scientific claims, which ultimately will lead to a public recognition of the contribution in the form of published article, book, PhD dissertation, for example. Thus, methodologically, testing the data-reuse mechanism will require that a scientific contribution is done with secondary data. However, in order to track changes in objects' causal power and liabilities and

⁷⁴ [E]ngines lose their power as they wear out, a child's cognitive powers increase as it grows. Therefore, in positing the existence of causal powers I am not invoking fixed, eternal essences. (Sayer, 2010, p. 71-72)

conditions during the time horizon of the data-reuse mechanism, I suggest that a diachronic data collection is also appropriate until, at least, an actual, and not only expected contribution is made.

In using Sayer's terminology (2010), a *data-reuse mechanism* is the mechanism that allows an individual (*object*) to have the ability to calculate a decision (*causal powers and liabilities*) with regard to secondary data (*object*) in the context of all her structural relationships (*structure*), and under some specific conditions (*conditions*) in order to make a scientific contribution (*events*). In other words, the *data-reuse mechanism* is a researcher's decision to make a scientific contribution, which the researcher expects to achieve with secondary data within a limited period and with her own skills and knowledge by keeping in mind both her personal and professional situations, and both the discipline's epistemic norms and the reward system she belongs to. The *data-reuse mechanism* captures the process of calculating a decision that ultimately allows the use of secondary data as evidence of scientific claims.

The mechanism or calculation of decision is bounded, because the researcher has not only limited time and resources, but also insofar as she has a bounded capacity to foresee the actual efforts and resources she will need for making a scientific contribution along the self-allocated period. Thus, she may not choose the best choice –or make the best decision– but one that *satisfices* her. Furthermore, her causal powers and liabilities, and the conditions may change along the process of pursuing the expected milestone or contribution, which will probably affect sequentially her next decision(s) or the way(s) that she searches for alternative options until she finds a satisficing one –procedural rationality (Simon, 1976, 2000). These subsequent decisions throughout the process of trying to make the scientific contribution may affect any part of the causal explanation: the contribution itself, the time in which she wants to achieve it, the resources she will use, or all of these three.

I have hypothesized five initial conditions for a data-reuse mechanism to function: the researcher knows that secondary data exist (condition C1), data have to be accessed or obtained by the researcher (condition C2), secondary data are a satisficing option for the researcher (condition C3), the idea of collecting particular primary data is not a satisficing option (condition C4), and an expected scientific contribution exists and the researcher finds its potential rewards satisficing (condition C5). However, changes in these conditions, namely in condition C4, may still lead to the use of secondary data as evidence of scientific claims as I expound later in this section.

The main object of the data-reuse mechanism is a researcher, but there are other objects in the mechanism. As Sayer (2000, 2010) explains, conditions are nothing else but objects with their own causal powers and liabilities acting on and thus activating the main object's causal powers and liabilities. Therefore, conditions C1, C2, and C3 refer to the object *secondary data*, while condition C4 refers to the object *the idea of collecting particular primary data*, and condition C5 refers to the object *an expected scientific contribution*. From these three objects, only *secondary data* are tangible

(even if they are digital data. After all, they need physical support to exist (Leonelli, 2016), while *primary data* and an *expected scientific contribution* exist initially only in researchers' minds. Primary data have not been generated and the scientific contribution has been done yet. However, this is not a problem since, as Sayer (2010)⁷⁵ explains, non-physical objects can be certainly the causes of certain events.

The main *event* under scrutiny is the use of secondary data used as evidence of scientific claims. My choice for the outcome –or event in Sayer's terminology (2010)– of data being used as evidence of scientific claims as opposed to data merely being used for the creation of background knowledge, is underpinned by the fact that only scientific claims⁷⁶ provide rewards to researchers, which, as I have argued in Chapter 3, trigger and drive researchers' decisions and actions with regard to resources, namely data.

Hereinafter, I provide details of all parts of the theorized data-reuse mechanism. Conditions C1, C2, and C3 refer to the object *secondary data*, so I have listed their causal powers and liabilities only under condition C1.

The researcher's structure and causal powers and liabilities

In general, we can say that a researcher belongs to both organizational and institutional structures. Organizations can be any research, academic or scholarly site. Institutional structures are both epistemic norms of the discipline the researcher belongs to, and science rewards norms. A researcher is a person, and thus she can belong concurrently to other structures. Keeping in mind that “[p]owers are thus the capacity to do or become, [and] liabilities the capacity to suffer or be affected” (Hartwig, 2007, p. 57), a researcher necessarily possesses the causal powers and liabilities to:

⁷⁵ “Now it might reasonably be objected that many of my examples in this discussion have been of physical causes, with the consequence that the applicability of causal analysis to the study of society might still be in doubt. In particular, one special type of social phenomenon whose causal status is widely doubted is that of ideas, beliefs and reasons. While it might be accepted that people have the causal power to reason and form ideas, the suggestion that reasons can be causes—that is, be the things which produce certain changes—is more difficult to accept. Reasons are very different from the material things in which we more readily recognize causal powers, and their enabling conditions are poorly understood. As was seen in Chapter 1, whereas the natural scientist has only the meanings of scientific concepts to interpret, the student of society has also to understand the intrinsic meanings of social practice. Reasons can also be evaluated as good or bad, false, inconsistent, etc., but it would make no sense to evaluate a physical cause in this way, although we might evaluate its results for our own interests.

Yet while reasons are certainly different in these respects from physical causes, it doesn't follow from this that they cannot be the causes of certain events. Indeed, why should we want to evaluate reasons if they could not be causes? If repugnant beliefs never did anyone any harm—because they never caused anyone to do anything—there would be little point in wasting our breath criticizing them. And why should anyone bother to argue (reason) that reasons cannot be causes if such arguments could never cause people to change their minds? One may grant that we know little about how beliefs (e.g. my beliefs in realism), intentions (my intention to write about it) and actions (my writing) are connected, but there are few things in life that we do which don't presuppose that reasons can be causes; indeed, in general, communicative interaction presupposes material results”. (Sayer, 2010, p. 74-75)

⁷⁶ Taking for granted that scientific claims are publicly communicated.

- Make (*satisficing*) decisions
- Take action
- Learn
- Identify knowledge gaps
- Make scientific contributions
- Act according to personal values
- Set up and pursue goals (whatever the motivation or reason is) and obtain resources (e.g., data, funding, etc.) to achieve them
- Interpret primary and secondary data
- Know epistemic practices of her discipline
- Know the limitations of the use of secondary data
- Know the limitations of the collection of primary data
- Be influenced by norms (e.g., epistemic norms, institutional norms, etc.)
- React to unexpected circumstances

Condition C1 – The researcher knows that secondary data exist

The researcher has to know that the *some particular secondary data* exist, no matter how she knows –whether by serendipity or purposely– (from experience, from searching, from asking colleagues, from the literature, etc.). This condition may seem too obvious and, thus, superfluous, but we cannot afford to disregard the obvious because the consequences could be deceptive. Every insignificant part of the black box has to be clearly identified and independently acknowledged from, but interrelated linked to, other parts of the box.

The existing particular secondary data possess the causal powers and liabilities to:

- Have many interpretations
- Be evidence of scientific claims
- Create knowledge
- Change their state (e.g., from an unprocessed state to a processed state)
- Become obsolete in several ways (e.g., conceptually, technologically, etc.)
- Change their availability (released data, stewarded data, proprietary data)

Condition C2 – Secondary data are obtained

Secondary data have to be obtained or accessed by the researcher. The fact that data are publicly released, published or shared on an “open” repository or in an archive is not a necessary condition for the actual or empirical realization of the data-reuse mechanism. The reuse of secondary data has existed for a long time even though data were not “openly” available. The researcher may or may not know, *a priori*, that data can be obtained, and thus some efforts in finding out might be necessary before realizing that she can or cannot obtain the data.

I have identified three plausible categories regarding researcher’s knowledge (condition C1) about data availability and accessibility, and the effort for obtaining the data:

- a) The researcher knows that data are available for reuse and are publicly released or published. I will term this data *publicly released data* or simply *released data*⁷⁷.
- a) Data are available for reuse, and this is known by the secondary user, but are not publicly released or published (the data are available for others to reuse them, but there may be some type of walls, e.g., payment walls, confidentiality walls, technical walls, etc., or conditions on the reuse). I will term this option *stewarded data* from now on.
- b) Data have not been publicly released, and the availability of the data for being reuse is uncertain. I will term this option *proprietary data* from now on.

Option a) does not require any effort to find out whether data can be obtained because they are available under no conditions of access, while options b) and c) do require some effort.

Condition C3 – Particular secondary data are an initial satisficing option

At an early stage of the process, the researcher perceives secondary data as a *satisficing* option for two purposes, namely for making a scientific contribution (C3-SC) or for creating background knowledge (C3-BK). Since *satisficing* is the result of the researcher’s subjective assessment, it is difficult to provide the parameters or criteria that a researcher would use to judge that some particular secondary data are *satisficing* for her. However, and based on the literature review, I hypothesize that the fitness of the data with the research question, the quality of the data, and researcher’s skills to interpret the data are relevant criteria for the researcher to consider particular data *satisficing*.

⁷⁷ I prefer this term to *Open Data*, as there is no one universal consensus on the meaning on *Open Data*, and the host of definitions and conceptualizations may clash with my category of *publicly released data*.

In the case of C3-Scientific contribution (C3-SC), the researcher perceives the option of using *some particular secondary data satisficing* in so far as she thinks she can make a scientific contribution with secondary data, alone or together with primary data. In this case, this perception implies necessarily that she perceives the idea of using *secondary data in general satisficing* for making a scientific contribution in so far as she can obtain her expected rewards taking into account the epistemic norms of her discipline.

In the case of C3-Background knowledge (C3-BK), the researcher perceives the option of using *some particular secondary data satisficing* in so far as she thinks she can answer a research question with these data only for creating background knowledge, e.g., generate or validate hypotheses with no intention to publish them. In this case, this perception does not necessarily imply that the idea of using secondary data is accepted in her discipline as evidence of scientific claims.

Either condition, C3-SC or C3-BK, has to be met. However, if condition C3-SC is met, it implies that C3-BK is also met. Conversely, condition C3-BK does not imply that condition C3-SC is met.

Condition C4 – The idea of collecting particular primary data is not an initial satisficing option

This condition (C4) does not mean or imply that the researcher perceives the use of *primary data in general* as a non-satisficing option. I cast doubt on the possibility that a research discipline does not accept primary data as evidence of scientific claims, and thus be not *epistemically satisficing*, even though in some disciplines knowledge is advanced mainly and usually with secondary data, e.g., computational biology.

This condition refers to the fact that at an early stage of the process, the researcher does not perceive *the idea of collecting particular primary data satisficing*. The researcher thinks that the efforts in collecting and analyzing her own primary data in order to complete her study do not compensate the rewards that she could obtain within the period she needs to obtain them. In other words, the necessary efforts cannot be carried out within the period that she has self-assigned for making the scientific contribution, and thus for obtaining rewards.

The idea of collecting particular primary data possesses the causal powers and liabilities to:

- Become into real action, that is, into the actual collection of particular primary data

Condition C5 – An expected scientific contribution exists and the researcher finds its potential rewards initially satisficing

This condition is two-fold: a researcher expects making a specific scientific contribution and she finds that the rewards she will obtain for the contribution are satisficing. One without the other makes no sense. As long as the researchers perceives the scientific contribution's rewards satisficing, she will deploy all necessary cognitive and material resources in order to achieve it.

This condition possesses the causal powers and liabilities to:

- Keep the researcher motivated to make a scientific contribution
- Become non-satisficing and, thus, be disregarded
- Become into an actual scientific contribution or achieved career milestone

Potential events of the *data-reuse mechanism*

When the five conditions of the theorized data-reuse mechanism are initially met, and provided that the researcher's causal power and liabilities and structure are, at least, the ones I have hypothesized at the beginning and at the end of the data-reuse process, I suggest that the mechanism may have three potential outcomes:

Outcome 1) Use of secondary data does not happen at all after having tried or considered the option

Outcome 2) Use of secondary data happens but reuse is not shared with the research community and the data do not end up being evidence of scientific claims. Thus, secondary data end up serving as widening the researcher's background knowledge and triggering new research hypotheses.

Outcome 3) Use of secondary data happens and only secondary data are used as evidence of scientific claims.

Potential outcomes are not necessarily the ones initially decided by the researcher. Outcomes can be different from the initial decision (Vidaillet, 2009) based on conditions C3, C4, and C5. For example, a researcher may initially decide to use secondary data for creating own background knowledge

(outcome 2), but may decide eventually to use secondary data as evidence of scientific claims (outcome 3), or to stop conducting their study (outcome 1), or vice versa.

These three outcomes are based on opposite poles regarding the use of primary and secondary data (conditions C3 – particular secondary data are an initial satisficing option, and C4 - the idea of collecting primary data is not satisficing). However, I acknowledge, though, that the combination of these two conditions C3 and C4 is more fuzzy or blurry in real life in some scientific disciplines. The option of choosing between only collecting primary data or only using secondary data in order to make a scientific contribution is not so straightforward and, thus, there can be more outcomes than the three identified above in which secondary data can be used as evidence of scientific claims. When condition C4 (the idea of collecting primary data is not satisficing) is not met, I have also hypothesized three⁷⁸ potential outcomes:

Outcome a) Use of primary data happen and primary data are used as evidence of scientific claims. Use of secondary data does not happen.

Outcome b) Use of primary data happen and primary data are used as evidence of scientific claims. Use of secondary data happens, but secondary data are not used as evidence of scientific claims. Instead, secondary data are used for the creation of background knowledge, thus, they do not appear in the scientific publication or contribution.

Outcome c) Use of primary data and secondary data happen, and primary data are presented as evidence of scientific claims. Secondary data can be presented in two ways: to support the main scientific claim done with primary data or as evidence of scientific claims in combination with primary data. These two options of outcome c) might be difficult to distinguish in a straightforward way.

Following Sayer's (2010) structure of causal explanation, Figure 4 depicts⁷⁹ the theorized data-reuse mechanism and all potential hypothesized outcomes of the data-reuse mechanism (also in Table 1 on page 58). A full size of Figure 4 is in annex 23.

⁷⁸ A fourth outcome d) in which the researcher would decide to use secondary data as the main evidence of scientific claims while considering primary data also as a satisficing option for making a scientific contribution, although possible, it is highly unlikely to happen.

⁷⁹ “*Conceptual frameworks are best done graphically rather than in text. Having to get the entire framework on a single page obliges you to specify the bins that hold the discrete phenomena, map likely interrelationships, divide variables that are conceptually or functionally distinct, and work with all of the information at once.*” (Miles, Mubermann, & Saldaña, 2014, p. 25)

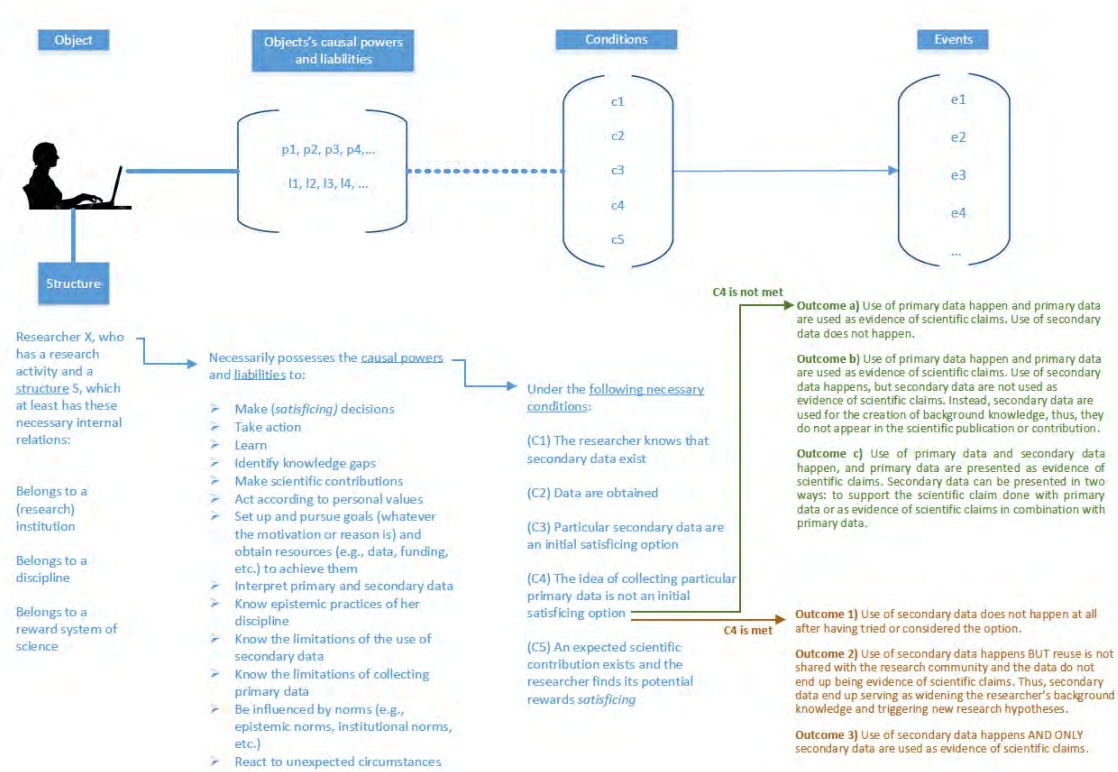


Figure 4 - The data-reuse mechanism and its structure and potential events

Table 1 - Two possible initial combinations A and B of conditions of the data-reuse mechanism, and their respective hypothesized outcomes

Potential initial combinations of conditions C3, C4, and C5 (conditions C1 and C2 are always met in both combination A and B)	Potential hypothesized events or outcomes
<p>COMBINATION A</p> <p>The idea of collecting primary data is a satisficing option.</p> <p>Particular secondary data exist and the researcher knows that they exist and secondary data are also a satisficing option for making a scientific contribution.</p> <p>An expected scientific contribution or career milestone exists and the researcher finds its potential rewards satisficing.</p>	<p>Three⁸⁰ potential outcomes or events:</p> <p>Outcome a) Use of primary data happen and primary data are used as evidence of scientific claims. Use of secondary data does not happen.</p> <p>Outcome b) Use of primary data happen and primary data are used as evidence of scientific claims. Use of secondary data happens, but secondary data are not used as evidence of scientific claims. Instead, secondary data are used for the creation of background knowledge, thus, they do not appear in the scientific publication or contribution.</p> <p>Outcome c) Use of primary data and secondary data happen, and primary data are presented as evidence of scientific claims. Secondary data can be presented in two ways: to support the scientific claim done with primary data or as evidence of scientific claims in combination with primary data.</p>
<p>COMBINATION B</p> <p>The idea of collecting primary data is not a satisficing option.</p> <p>Particular secondary data exist and the researcher knows that they exist and are a satisficing option for making a scientific contribution.</p> <p>An expected scientific contribution or career milestone exists and the researcher finds its potential rewards satisficing.</p>	<p>Three potential outcomes or events:</p> <p>Outcome 1) Use of secondary data does not happen at all after having tried or considered the option.</p> <p>Outcome 2) Use of secondary data happens, but reuse is not shared with the research community and the data do not end up being evidence of scientific claims. Thus, secondary data end up serving as widening the researcher's background knowledge and triggering new research hypotheses.</p> <p>Outcome 3) Use of secondary data happens and only secondary data are used as evidence of scientific claims.</p>

⁸⁰ A fourth outcome in which the researcher would decide using secondary data as the main evidence of scientific claims having primary data as a *satisficing* option, although possible, it is highly unlikely to happen.

4.3. Justification of a multi-case study approach

Sayer's account of mechanisms and process-tracing methods are suited for case study analysis, according to Easton (2010) and Beach & Pedersen (2013), respectively. However, there are other reasons why the case study approach is a proper choice for answering the research questions, which I explain in the next paragraphs. Indeed, a few empirical studies on data reuse have used the case study approach previously, e.g., Daniels, 2014; Hyman, 1972; Pasquetto, 2018; Zimmerman, 2003. However, my research approach is more similar to the one conducted by the sociologist Herbert H. Hyman at the early 70s of last century in several aspects which I detail along this Chapter 4. Methodology and methods.

There are many definitions of a case study, and some of them –either in their entirety or in some of their parts– are wrong, misleading, and problematic, (Flyvbjerg, 2006, 2013; Gerring, 2004). Thus, I have decided to choose two simple definitions because I find them the least misleading ones keeping in mind that a “[c]ase study is not a methodological choice but a choice of what is to be studied” (Stake, 2005, p. 443).

[A] *case study [is] an intensive study of a single unit for the purposes of understanding a larger class of (similar) units [and it can be] observed at a single point in time or over some delimited period of time.* (Gerring, 2004, p. 342)

[...] *we can define a case [case in italic] as a phenomenon of some sort occurring in a bounded context. The case is, in effect, your unit of analysis.* (Miles, Matthew B; Huberman, 1994, p. 25)

These two definitions highlight two relevant aspects of my empirical research –the *context* and the *delimited period of time*, which I expound in more detail hereinafter. Apart from these two relevant aspects, generally speaking, a case study research approach is useful when we have “how” and “why” research questions as it is the case in this dissertation; when we do not have any control over the behavior of participants involved in the study; and when we need to inquiry about a phenomenon that is contemporary (Yin, 2003). Also, case studies are useful when we need to analyze and understand the contextual conditions of the phenomenon (Baxter & Jack, 2008; Connaway & Powell, 2010; Yin, 2003), and when entering a new field (Eisenhardt, 1989).

There are two main reasons why to use a case study approach in this study. On the one hand, I have defined *data* based on the concept *data streams* by Hilgartner & Brandt-Rauf (1994) and on the

relational framework by Leonelli (2015, 2016). In my definition, data count only as data in specific research enquiry processes⁸¹. In addition, in my proposed theoretical model, I have hypothesized that conditions (initial conditions C1, C2, C3, C4, C5, and objects' causal powers) might change their value along the process of reusing data, thus the aspect of the delimited time has to be considered in the design of methods as explained hereinbefore. Thus, these two main issues require studying process of data reuse diachronically as part of a concrete research enquiry, project or question, which will be the main boundary for collecting the data, although not for eligibility criteria of cases as explained hereinafter. On the other hand, the two research questions require an in-depth and fine-grained analysis in order to open the *black-box* of the process of reusing data for identifying the necessary conditions under which data reuse happens, and the researchers' decision making process.

However, a case study approach does not guarantee by itself that the research questions can be answered. Data collecting methods and instruments have to be properly suited to the research questions within the case study approach because a case study admits all kinds of methods and instruments, not only qualitative ones (Stake, 2005, p. 443). Thus, a proper choice of data collection criteria is needed within the case study approach. In this study, my choice of methods and instruments is underpinned by an underlying causal and systemic approach based on the theorized data-reuse mechanism.

I propose answering the two research questions that guide this research by means of an *instrumental case study* (Stake, 2005) in which we have five initial conditions C1, C2, C3, C4, C5, and we know X (the goal of making a scientific contribution), and Y (outcome #3 or c – data are reused as evidence of scientific claims).

I use the term instrumental case study if a particular case is examined mainly to provide insight into an issue or to redraw a generalization. The case is of secondary interest, it plays a supportive role, and if facilitates our understanding of something else. [...] Here the choice of case is made to advance understanding of that other interest. (Stake, 2005, p. 445)

One case study, and within-case inference, would be enough for “analyz[ing] whether a theorized causal mechanism exists in an individual case” (Beach & Pedersen, 2013, p. 69). However, I suggest that an instrumental *multiple case study* (Stake, 2005) is more appropriate because, on the one hand, the five conditions (C1, C2, C3, C4, C5) can have different values and change their value along the process of reusing data. Also, the mechanism' objects' causal powers and liabilities can have different

⁸¹ [...] I advocate defining data in terms of their function within specific processes of inquiry accounts, rather than in terms of intrinsic properties. Within this [relational] framework, it is meaningless to ask what objects count as data in the abstract. This question can only be answered with reference to concrete research situations, in which investigators make decisions about which research outputs could be used as evidence and which are instead useless in that regard. (Leonelli, 2016, p. 79)

initial values and change along the process (Easton, 2010; Sayer, 2000, 2010). On the other hand, we can have two different causes X (a scientific contribution or the creation of background knowledge), which can also change along the process of reusing data, and different outcomes Y (outcomes 1, 2, 3, a, b, or c).

When there is even less interest in one particular case, a number of cases maybe studied jointly in order to investigate a phenomenon, [...]. I call this multiple case study or collective case study. It is instrumental study extended to several cases. Individual cases in the collection may or may not be known in advance to manifest some common characteristic. They may be similar or dissimilar, with redundancy and variety each important. They are chosen because it is believed that understanding them will lead to better understanding, and perhaps better theorizing, about a still larger collection of cases. (Stake, 2005, p. 445-446)

A multiple case study or cross-case study lets us make causal, breadth and boundedness inferences, have representativeness, find causal effects with probabilistic causal relationships, and confirm theory if the choice of the cases is adequate to the research questions (Gerring, 2004). In other words:

[...] multiple cases offer the researchers an even deeper understanding of processes and outcomes of cases, the change to test (not just develop) hypotheses, and a good picture of locally grounded causality. (Miles, Matthew B; Huberman, 1994, p. 26)

However, in order to make cross-case inferences from multiple case studies, other types of inferences are needed, namely comparative methods (Beach & Pedersen, 2013).

For Miles and Huberman, the case is the unit of analysis and it can be anything, from an individual to a nation, but also an event, a process, a place, and so on. Whatever the unit of analysis is, boundaries between the case study's focus (or heart) and its context are most of the times –if not all– blurry. Boundaries of case studies should include the natural setting and time context of each case study; concepts from the theoretical framework, from personal or professional experience and background, from current relevant societal concerns, etc.; and from sampling decisions (Baxter & Jack, 2008). Sampling decisions can be based on the former –the context and concepts–, but also on other criteria, for example, available resources. Sampling decisions are crucial to avoid becoming overloaded with data and to make sure the research is feasible or reasonable in scope, but a sampling exercise also determines the breadth and depth in which the phenomenon is studied and not only the sample that is studied (Baxter & Jack, 2008). Furthermore, criteria for sampling may change along the process of the

study, because it is not always possible to find the case studies that one decides at the outset of the study, so sampling decisions have to be changed and thus define the case study further along the process of the research (Matthew B Miles & Huberman, 1996).

In this dissertation, the focus of each case study is an individual researcher and her decision making of data reuse. The boundaries of each case study are each individual researcher's context (structure, and causal powers and liabilities) together with sampling decisions based on the events (outcomes 1, 2, 3, a, b, c), and on the conditions (C1, C2, C3, C4 and C5). However, when entering the field, we may know very little or nothing about many of the causal powers of the mechanism and, an empirical study should not be constrained by initial and rigid sampling decisions. Therefore, researchers should remain flexible and open in adapting the sampling decisions. Abbott (2004) warns us that sampling decisions are always a starting point, which evolves along the study, and may not provide what initially they were aimed to provide. He argues that research proposals' ideal empirical objects⁸² have nothing to do with what finally is used as an empirical object.

[...] *in the social sciences we [...] often don't see ahead of time exactly what the problem is, much less do we have an idea of the solution. We often come at an issue with only a gut feeling that there is something interesting about it. We often don't know even what an answer ought to look like. Indeed, figuring out what the puzzle really is and what the answer ought to look like often happen in parallel with finding the answer itself. [...] original research proposals usually turn out to have just been hunting licenses, most often licenses to hunt animals very different from the ones that have ended up in [publications].* (Abbott, 2004, p. 83)

In any case sampling choices are necessary and in qualitative studies “tend to be purposive, rather than random”, contend Miles and Huberman when citing Kuzel and Morse (1994, p. 27), and each type of sampling has a different purpose. Drawing upon the different sampling strategies presented in *Typology of sampling strategies in qualitative inquiry* by Miles and Huberman (1994, p. 28), initially⁸³, my sampling strategy was both *maximum variation* because its purpose is to document

⁸² In fact, originally, my research proposal consisted on studying *The collective negotiated process of reusing data*. In my original research proposal, the unit of analysis or case study was a collective or group of researchers, while the unit of data collection was each of the individuals belonging to the research group. However, I had to abandon my original research goal because I found serious difficulties in finding "collectives of researchers" in willing to participate. The problem was not so much to find individual researchers (not easy either, though), but to find a group whose 80% or 90% of the members were willing to participate.

⁸³ Later on, and once I entered the field, I added more sampling strategies. *Snowball or chain* that identifies cases of interests from participants, *opportunistic* because it lets you take advantage of some interesting unexpected variability, and *extreme or deviant case or polar types* (Pettigrew, 1990, p. 275), which gave me the opportunity to learn from intensive cases of data reuse.

variations and identify common patterns, and *criterion* because all cases had to meet a criterion, which is useful for quality assurance. The maximum variation is useful because it requires comparative methods, which “[...] work[...] best when the entities to be compared are different enough to present interesting contrasts, yet similar enough for the variations to be disciplined” (Jasanoff, 2005, p. 29).

So, ideally, the best sampling criteria would have been the maximum variation with the five initial conditions (C1, C2, C3, C4, C5), with the causal powers and liabilities of data and of researchers, and with the six different outcomes (1, 2, 3, a, b, c). However, finding cases that met any possible combination of these sampling criteria was difficult within the time constraints in which I had to conduct the empirical work. So, I decided to base the sampling criteria on some of the initial conditions, and I searched for case studies with different and unknown outcomes in order to compare the causal forces of the mechanism for each of the outcomes. We can learn more about the causal forces of the mechanism when we are open to have different outcomes. “When observations are selected on the basis of a particular value of the dependent variable, nothing whatsoever can be learned about the causes of dependent variables without taking into account other instances when the dependent variable takes on other values (King, Keohane, & Verba, 1994, p. 129). Therefore, I decided to disregard the outcomes (1, 2, 3, a, b, c) as a compulsory sampling criterion, and thus to include ongoing case studies in which the event or outcome was unknown. Other main reason for this choice is the advantage of studying decision-making processes in an ongoing or prospective way, since it minimizes observant’s and participant’s biases⁸⁴ in the reconstruction of a process (Nutt & Wilson, 2010a; Poole & Van de Ven, 2010). Therefore, I have purposely searched for both ongoing and finished research projects using secondary data as case studies.

So, the final main *instrumental* sampling criteria (Stake, 2006) was to choose case studies, which met necessarily condition C3 – *Particular secondary data an initial satisficing option*, and the main *criterion* sampling was that all cases met the definition of *reuse of data*. During the recruiting process, I tried to formulate all these three conditions together in such a way that potential participants would identify in a rather easy way whether their situation would meet the criteria. Researchers should have been reused recently or being reusing data that they did not collect themselves, and that the reuse should be tied to a specific research question⁸⁵ or project.

⁸⁴ Bias may come from the observer. It can be tempting to fall into the trap of the trope of the precursor, purposely or unconsciously, trying to match forcedly the known effect (Y) with the known cause (X). In other words, proving a retrospective inaccurate or wrong account of the events, and, thus of the causal forces of the mechanism. The other drawback may come from the participants when reporting their decisions, as they can be reified and chronologically inverted before actions (Weick, 1995).

⁸⁵ The fact that reuse is tied to a specific research question with different potential outcomes (data reused as evidence of scientific claims or reuse of data does not happen) was also addressed by Hyman, although in a more tacit way. In the recruiting letter that he sent to researchers who reused data from the Archives of the Michigan Inter-University Consortium or the Archives of the Roper Public Opinion Research Center. However, for Hyman there are only two potential outcomes: data reuse happens and data are used as evidence of scientific claims, or does not happen. Here I reproduce his exact words in his letter. I have underlined the words where he tacitly refers to the same sampling conditions that I use. My comments are in bold and between brackets immediately after:

However, in order to add some variability to the cases, I added a sampling criteria based on C2 – *Secondary data are obtained*. The goal was to find *polar* case studies based on the three plausible options regarding data availability and accessibility: *released data*, *stewarded data*, and *proprietary data*.

Furthermore, and in order to be able to make disciplined comparative inferences (Jasanoff, 2005), I searched for cases in which the two material objects –in Sayer’s terminology– (the researcher and the secondary data) would not change. Therefore, I searched for case studies with the same researcher reusing different data and for case studies with the same data being reused by different researchers. However, due to the prospective and diachronic characteristic of the data collection process, the two material objects’ causal powers and liabilities could change along the process.

Regarding the empirical field, I chose health disciplines in a very broad way. I made this choice after reviewing previous studies on data reuse by scholars in information science at the beginning of 2015. There were mainly two reasons for choosing health disciplines. First, I saw the opportunity to make a potential empirical contribution to understand why and how data reuse happens by including a nearly⁸⁶ unexplored empirical field. In 2015⁸⁷, I could not find any study on data reuse in health disciplines conducted by IS scholars. Second, the research institute, which has hosted me as a PhD student (INGENIO, CSIC-UPV)⁸⁸, conducts part of its research in health sciences, so I saw the opportunity to use the already established relationships that INGENIO had with health researchers and clinicians to carry out this study. However, the choice of a discipline is an instrumental criterion, and thus any discipline would serve to answer the research questions of this study.

If your use culminated in some publication. [Hyman assumes that data can end up being evidence of scientific claims] I would greatly appreciate knowing the exact citation, and if it is available in article or report form, 1-2 reprints would be helpful to me. By examining such writings myself, I can attempt to abstract some general conclusions. But if you were to take the additional time to write a brief chronicle or letter describing your experience in that secondary analysis, it would be even more helpful.

It may be that your work with the data did not result in any publication. [Hyman assumes that data can end up not being reused] Then, the only way in which such experiences can become useful to the larger profession is by your own account, and I hope very much, if you are among this category of users, that you will write me a brief statement (Hyman, 1972, p. 338)

It was also addressed by Zimmerman My interviews with ecologists focused on papers they published in an issue of Ecology or Ecological Applications in 1999, 2000, or 2001. The term case refers to each of the instances of data reuse by one of the thirteen ecologists. (Zimmerman, 2003, p. 138).

⁸⁶ One of Piwowar’s research goals was to find out the impact of publicly shared raw gene expression microarray datasets. Despite the title of her dissertation and her further work on “data sharing”, she was delving somehow into the “reuse of data”, since impact of data sharing was measured upon citation as an indicator of reuse. However, her research goals were related to understand the process and the conditions under which data reuse happens. (see Piwowar, 2010)

⁸⁷ While finishing my analyses and drafting some parts of this thesis, Irene Pasquetto defended her thesis about data reuse in a biomedical fields in 2018 (see Pasquetto, 2018).

⁸⁸ <http://www.ingenio.upv.es>

4.4. The search process of case studies

Health science disciplines, as an empirical field is too broad to look for cases of data reuse through the literature. Therefore, I first used automatic searching was in order to identify cases of data reuse through the literature in health sciences as other scholars have done before, e.g., Yoon (2014b). August 18, 2016 I emailed Michael Boutet, a Replacement Health Sciences Research Liaison Librarian in order to get some help in the search. I shared with Michael Boutet my early tentative sampling expectation of cases reusing data. He searched in the Ovid MEDLINE(R) In-Process & Other Non-Indexed Citations and Ovid MEDLINE(R)⁸⁹. We soon realized that it was very difficult to find cases of data reuse in the literature in an automated way due to, among other things, lack of standard citation practices (Heather A Piwowar, Street, & Suite, 2011). Later on, I realized of other disadvantages when looking for cases of data reuse through the literature in an automated way. Data would have always been used as evidence of scientific claims, either alone or supporting claims done with primary data, and thus I would not have been able to find no-data-reuse cases or cases in which data were reused merely for the creation of background knowledge. In addition, published studies imply that they are finished, so the search in the literature would have impeded me finding ongoing cases of data reuse.

One solution was to use the word-of-mouth at OHRI (Ottawa Hospital Research Institute) in order to find case studies with different outcomes and ongoing case studies. In March 2017, after obtaining ethics clearance from both OHRI and the University of Ottawa, I decided to email the directors of the five research programs at OHRI, namely the Cancer Therapeutics Program⁹⁰, the Chronic Disease Program⁹¹, the Clinical Epidemiology Program⁹², the Neuroscience Program⁹³, and the Regenerative Medicine Program⁹⁴. In some cases, directors of these programs also belonged to health departments at the University of Ottawa (i.e., Department of Medicine, Department of Biochemistry, Microbiology and Immunology), so I also asked them to help me in recruiting participants in their OHRI research programs and in their departments. From the five programs, I got three responses, but only two responses were positive with the possibility of finding cases of data reuse, and came from the Cancer Therapeutics Program (CTP) and the Clinical Epidemiology Program (CEP). The director of the Cancer Therapeutics Program, Michael McBurney, after a face-to-face preliminary meeting, introduced me to David Cook of the Varderhyden Lab as a potential participant. The director of the CEP program, Dean A. Fergusson, suggested me cases, which reuse individual participant data in meta-analysis studies (IPD MA), and he recommended me to contact the principal investigators of an

⁸⁹ <https://www.ovid.com/product-details.901.html>

⁹⁰ http://www.ohri.ca/Programs/cancer_therapeutics/default.asp

⁹¹ <http://www.ohri.ca/Programs/ChronicDisease/default.asp>

⁹² http://www.ohri.ca/Programs/clinical_epidemiology/default.asp

⁹³ http://www.ohri.ca/Programs/clinical_epidemiology/default.asp

⁹⁴ http://www.ohri.ca/Programs/Regenerative_Medicine/default.asp

IPD MA in renal transplant and thrombosis. The recruitment of an IPD NMA study that accepted to participate in this research was through mouth-of-word thanks to Jordi Pardo Pardo at the Centre for Practice Changing Research, which is located at the OHRI General Campus. The third IPD NMA study is a snowballing contact from my participant of the second IPD MA case study.

With the Varderhyden Lab and the IPD MA studies, I had case studies of *released data* and *proprietary data* respectively.

For finding case studies of reuse of *stewarded data*, the best option was to identify a data repository. I stumbled upon the BORN Ontario repository unexpectedly and without any effort. It was during my presentation (“Use of secondary data: new knowledge from old data” borrowing part of Zimmerman’s article’s title (Zimmerman, 2008)) at the Clinical Epidemiology Program Debates Series at OHRI in November 2, 2016. One of the attendees, Sandra Dunn, Knowledge Translation Specialist of the BORN Ontario data repository made two very interesting comments regarding one of the conditions for reusing data, namely the fitness of the research question with the data. I contacted her a few days after my presentation to discuss about her enlightening comments about fitness during my presentation and to ask her if it would be possible if BORN Ontario could help me to identify cases of reuse of BORN Ontario data. The direction of BORN Ontario was very kind and emailed researchers using or having used BORN data with my recruiting message after having ethics clearance and having signed both OHRI and the University of Ottawa the Data Sharing Agreement.

BORN Ontario data was not the only data repository where I searched case studies. I also contacted the three principal investigators of the Canadian Longitudinal Study on Aging/Étude longitudinale canadienne sur le vieillissement (CLSA-ÉLCV)⁹⁵ to ask them for authorization in contacting researchers or research groups who had used data from CLSA-ÉLCV. One of the PIs, Christina Wolfson, PhD, answered me on behalf of the other two PIs. They gave me their authorization to contact researchers of projects approved⁹⁶ by CLSA-ÉLCV. When I looked at the projects on the CLSA-ÉLCV’s web site in February 2017, there were only projects approved in 2014, 2015 and 2016 because the CLSA-ÉLCV, at that time, was a very recent⁹⁷ national data collecting effort. I contacted the only eight projects whose PI or data applicant was located in Ottawa or nearby⁹⁸. Only two researchers answered my message. One researcher apologized for declining participation in my study because her project was at a very early stage. The other respondent doubt about being the right person for my

⁹⁵ <https://www.clsa-elcv.ca>

⁹⁶ The list of projects approved to use data from CLSA-ÉLCV are publicly listed on this site <https://www.clsa-elcv.ca/approved-projects>

⁹⁷ <https://www.clsa-elcv.ca/about-us/history>

⁹⁸ This was both a strategic and a pragmatic decision. It was strategic because I had designed the collection of the data in two interviews, and I preferred face-to-face interviews to build trust with the interviewee. It was pragmatic because despite my Erasmus Mundus NOVA DOMUS scholarship, I could not have been able to afford travelling across Canada for the interviews.

study, so we scheduled a 15-meeting face-to-face meeting in his office to discuss it. When the meeting ended, it was clear that he was a right person to participate in my research, but he declined participating. Therefore, I disregarded completely the idea of incorporating CLSA-ÉLCV data reuse case studies in my thesis, and I communicated my recruiting experience to Christina Wolfson as she requested me when we talked the first time.

However, after collecting and doing a preliminary analysis of data from these participants in Canada, and once I was back in Spain, I decided to add more variability to my cases of reuse of *released data* since the participant of the three case studies in Canada was the same researcher. So, I decided to look for at least one more case study in which the reuse of data was of *released data*. I used again the word-of-mouth strategy by asking Enrique Lanuza, PhD, who is a professor in biology at the *Universitat de València*. He suggested me a name and it worked out. Therefore, I could add one more case study (#4) to the nine case studies that I already had.

4.5. Description of data collecting methods and instruments

The case study researcher faces a strategic decision in deciding how much and how long the complexities of the case should be studied. Not everything about the case can be understood - so how much needs to be? Each researcher has choices to make. (Stake, 2005, p. 448)

I completely agree with Stake that a researcher has to make choices, and thus I have made the choice to use several instruments to collect empirical data for this study. My decisions for choosing them, although are justified by some literature mainly about methods, are also conditioned by the resources and time I had to conduct this study, and by the conception of data reuse as a means to obtain a goal, and not as an end.

I use the interview, author's publications –that sometimes precede the reuse of data–, visual representations, documents such as research protocols or data sharing agreements, and written messages between participants and myself. Each of these instruments complement the others, but the interview is the main and central one. Rather than listing all good qualities of each of these instruments, which I am sure the reader is familiar with, I prefer highlighting the main reason why I have chosen them and how they complement other instruments.

The interview

“If you do interviews with scientists and show that scientific method rests upon social negotiation, Ashmore will come and interview you and show you that your method also relies upon social negotiation” (Pinch, 1990)

Interviews have been both criticized and praised as an instrument of data collection. One of the criticisms alludes to our inability to access the “true” social life, thus in interviews we only have “modes of articulation”, instead of “observable realities” (Baker & Edwards, 2012; Silverman, 2004). While proponents of *observation* argue that observation avoids dependence on participants’ own interpretative exercise, Gilbert and Mulkay (1984) argue, on the one hand, that scientists’ acts do not reveal the underlying and implicit meanings of those acts. For example, if a researcher is performing an experiment, how can we know if the researcher wants to test a hypothesis or if she is checking the machine or instrument? The only way to clear up this question is to ask the researcher doing the experiment. On the other hand, social actions also have multiple meanings⁹⁹. Therefore, observations are not an infallible tool in order to conduct social studies of science, such as researchers conducting their studies and analysis. Neither are they always useful. Whether observation is an appropriate method will depend on the research questions that we have, and on what is possible¹⁰⁰ to observe.

The problem with the interview is not in the interview itself as an instrument of data collection. Instead, “[t]he error is that we mistake the socially shaped account for the authentic voice of truth” (Baker & Edwards, 2012, p. 12). I argue that interviews are the most suitable¹⁰¹ instruments for this research, despite the perils of interviewees’ potential (un)intentioned distorted accounts or own interpretive meanings. However, I have tried to counteract participants’ biases in the reconstruction of the process by including ongoing case studies, and thus be able to carry out the reconstruction myself.

I have chosen a semi-structured¹⁰² interview script (annex 1) in order to obtain a balance of depth, specificity and range in participants’ answers for the first interview. There is a second interview with

⁹⁹ For instance, does a given set of activities constitute an experiment, an attempt indirectly to raise more research funds, an effort to secure professional credibility, a bid for more students; or can it be any or all of these, depending on the context in which the actor is talking or writing about his actions? If the latter is the case, and we suggest that it is, then “the meaning” of his action is variable and context-dependent. It will be quite impossible to establish the nature of the action unequivocally by being present at and directly observing the original laboratory experiment. (Gilbert, 1984, p. 9)

¹⁰⁰ There is no scientific or commonsense reason to observe researchers working in front of a computer during 6 or 8 hours a day, waiting for them to make any action related to the project or study in which she would use secondary data.

¹⁰¹ According to Silverman, it is possible to gather information about participants’ social world through in-depth interviewing. The solution lies, partly, in to understand and acknowledge how, where, and why the stories we have about social life are produced (Silverman, 2004). After all, all representations, such as interviews, are perfect for something, whatever “something” is (Becker, 2007).

¹⁰² There is no such thing as a completely structured interview or a completely unstructured interview (Parker, 2005)

an ad-hoc script for each case study. Despite participants' freedom in choosing their own words in open-ended questions, the interview establishes an asymmetrical power relation between the interviewer and the interviewee, for example, the interviewee rarely asks questions, and does not introduce new topics or directions in the interview. Conversely, the interviewer imposes the topic agenda. So, the interview could be also described as a guided asymmetrical conversation (Brinkmann, 2018).

Authors' publications and their complementary role

Author's publications, –given for granted that they exist and are related to the case study under scrutiny– play an important role in three ways. First, publications can provide details about the phenomenon that may not surface in the interview. Second, they may provide useful information in order to prepare an *ad hoc*¹⁰³ interview script, and, thus, to be able to dig deeper into some issues of the reuse of secondary data in a more efficient way since they are known in advance by the interviewer. Third, publications let us contrast interviewees' black-boxed account of the process of using secondary data. Differences in these accounts might be useful when analyzing the empirical data and come to conclusions.

Hyman recognizes the complementary role that authors' publications play in analyzing cases of secondary use of data. He wrote in the recruiting letter of his study:

If your use culminated in some publication, I would greatly appreciate knowing the exact citation, and if it is available in article or report form, 1-2 reprints would be helpful to me. By examining such writings myself, I can attempt to abstract some general conclusions. But if you were to take the additional time to write a brief chronicle or letter describing your experience in that secondary analysis, it would be even more helpful (Hyman, 1972, p. 338).

However, while authors' publications let us analyze the original reports that include the use of secondary data, authors can omit details about the process of decision-making, and the challenges they find and how they overcome them in their publications. This may be explained by the fact that the scientific discourse follows standards in the context of justification (Reichenbach, 1938) that, in turn, force scientists to omit some details and to distort the sequence in which things happened (Gilbert & Mulkay, 1984). Therefore, authors' publications alone cannot provide enough information in order to

¹⁰³ This has also proved to be useful in building trust with the interviewee and showing the interviewee that I was really interested in her use of secondary data.

answer the two research questions that guide this study, but are a useful instrument for collecting data that complements the interview.

When possible, and mainly in order to carry out *informed interviews* (Laudel & Gläser, 2007), and to make sound data analysis, I have tried to learn as much as possible about each researcher's discipline's epistemic practices and methodologies. Apart from their publications related to the reuse of data, I have gathered other data sources, including, research proposals and protocols, researchers' online research profiles and their scientific production and data repositories' online websites.

Visual representation of each participant's data reuse process

Visual maps or representations are widely used in social sciences research (M. B. Miles, 2014; Dodge & Kitchin, 2003; Jørgensen, 2012; Leydesdorff & Rafols, 2012; Porter & Rafols, 2009). More specifically, visual representations are used in the process of gathering data from interviewees (Gläser & Laudel, 2015; Huvila, 2009), and are significantly useful for studying decision making processes (Van de Ven & Poole, 1990).

I have chosen a visual representation or workflow diagram of the process of using secondary as one of the complementary instruments of the interview, and it has a four-fold utility. It has served for collecting data, analyzing data, validating findings with participants, and presenting results. Thus, first, I used the workflow diagram of the process of reusing data as a tool for identifying information gaps and connections among the different events, stages and conditions of participants' processes of using secondary data. Second, the workflow diagram served as a validating tool with participants of my interpretation of their own accounts of the process, as I explain in section 4.6. Data analysis methods. Third, it served me to carry out within-case and cross-case data analyses. Fourth, but not least, I use the workflow diagrams –with participants' corrections–, for presenting results of this study.

As a word of caution, the visualization represents the workflow or process of using secondary data in each of the case studies, and not the researcher's decision-making process nor much less the causal forces of the data-reuse mechanism. The workflow diagrams of the different processes of using secondary data show *only* the empirical events, stages, and some conditions of the process temporally connected between a start and an end that the interviewee narrates. Yet, they do not represent nor explain by themselves the underlying decision-making process or the causal forces of the data-reuse mechanism. Tracing visual representations or workflow diagrams of events, stages, and conditions in a process does not equate to studying causal mechanisms (Beach & Pedersen, 2013).

Follow-up messages

For ongoing¹⁰⁴ case studies, I have followed up with participants about the progress of the research project and the embedded process of using secondary data by email. I have also used these email messages as data sources for answering the research questions.

4.6. Data analysis methods

There are two different stages of analysis in this study, for which I have used different types of inferences, namely induction, deduction, retroduction and abduction, which, in turn, have required different analytical methods. Nevertheless, in general, I have used these types of inferences all along the whole study, which makes it difficult to match a type of analysis with a specific stage.

At a first stage, there is an analytical process for theorizing the *bounded individual horizon* (BIH) model and the data-reuse mechanism, for which I have mainly used both abduction and the retroduction. Abduction analysis consists of inferring to the best explanation¹⁰⁵ or, simply, by guessing (Reichertz, 2014). For this type of inference, I have based my analysis on literature about data reuse from different disciplines, on literature from both science policy and sociology of science studies, and on vignettes by a myriad of researchers, who I have encountered mainly in the last five years. Retroduction is the type of inference, which allows us to track and understand the linked and continuous process between a cause and its effect. It is a “mode of inference in which events are explained by postulating (and identifying) mechanisms which are capable of producing them” (Sayer, 2010, p. 72).

In order to present the empirical analysis of case studies from both abduction and retroduction methods, I have used Sayer’s structure of a causal mechanism (Sayer, 2010).

However, I suggest that Sayer’s account of mechanisms does not fully address “an adequate *ontic* account of mechanisms” (Machamer et al., 2000, p. 4) since his account of mechanistic explanations is more situated within the view that entities have capacities to do things than within the view that activities are reified. Yet, both views together are suggested to be necessary for a proper ontic account of mechanisms (Machamer et al., 2000)¹⁰⁶. Thus, I have also used *process-tracing* methods (Beach &

¹⁰⁴ They were ongoing at the moment of case studies selection

¹⁰⁵ Source: The Stanford Encyclopedia of Philosophy [<https://plato.stanford.edu/entries/abduction/>]

¹⁰⁶ *Entities and a specific subset of their properties determine the activities in which they are able to engage. Conversely, activities determine what types of entities (and what properties of those entities) are capable of being the basis for such acts. Put another way, entities having certain kinds of properties are necessary for the possibility of*

Pedersen, 2013; Bril-Mascarenhas, Maillet, & Mayaux, 2017) to find out how the data-reuse mechanism works in section 6.5. Process-tracing is a research method that allows to gain a better understanding of the causal forces that produce an outcome, and can be used for both theory-building and theory-testing purposes (Beach, 2017). Both Sayer's structure of mechanisms and process-tracing methods are suited for case study analysis, according to Easton (2010) and Beach & Pedersen (2013), respectively.

Furthermore, Sayer's structure of a causal explanation, and its visual representation, present one drawback. Time and, thus, the change in conditions and in the object's causal powers and liabilities, as well as the sequence of conditions –which may affect the values of the conditions and the outcomes or events (Abbott, 2001)– cannot be easily represented. Abbott's account of social events is in line with Simon's procedural rationality (Herbert A Simon, 1976) since according to Simon, both how the process evolves and the choice maker's decisions may affect the outcome. Thus, once more, I suggest that *process-tracing* (Beach & Pedersen, 2013; Bril-Mascarenhas, Maillet, & Mayaux, 2017) can be a solution for compensating the drawback of Sayer's static structure of a causal explanation.

At a second stage, which aims to test the theorized data-reuse mechanism and build theory, I have mainly used both deductive and inductive approaches¹⁰⁷ when conducting a within-case analysis for getting into the details for answering the two research questions that guide this research. For Flick, the inductive logic or emergent meanings come from the individual being studied, while the deductive logic or theoretical meanings come from the researcher who conducts the study (Flick, 2006). Yet, I have also used a comparative cross-case analysis in order to answer research question #2.

In addition, for presenting the findings and discussing them, I have also considered Ragin's three caveats about causality in social life, which I suggest that can be applied to mechanisms. First, there is, typically, no one single mechanism for an outcome. Second, we may find multiple and conjunctural mechanisms for an outcome, and third, a mechanism may have opposite effects depending on the conditions under which the mechanism operates (Ragin, 1987).

acting in certain specific ways, and certain kinds of activities are only possible when there are entities having certain kinds of properties. Entities and activities are correlatives. They are interdependent. An ontically adequate description of a mechanism includes both. (Machamer et al., 2000, p. 6)

¹⁰⁷ *Deduction helps to identify the phenomenon of interest, suggests what mechanism may be at play and provide links with previous research and literature. Induction provides event data to be explained and tests the explanations* (Easton, 2010, 124)

Within-case analysis

As explained hereinabove, decision-making is a problematic concept to study because it is not directly observable and, for some scholars, it only exists on the eye of observer, who can bias actors for providing an account of their actions as if the latter were decisions or as if a decision preceded an action (Tsoukas, 2010). Therefore, although the unit of data collection is the researcher and the unit of analysis is her decision-making with regard to the reuse of research data (Neuendorf, 2002)¹⁰⁸, the unit of observation is the events that the researcher makes happen (Abbott, 2001). For analyzing researchers' events, I have carried out *thematic analysis* (Braun & Clarke, 2006) with a predominant deductive inference from the early conception of this research.

I have carried out the deductive inference based on the theoretical concepts¹⁰⁹ that I have also used for theorizing the *bounded individual horizon (BIH) model*, i.e., bounded rationality, procedural rationality, satisficing, constructive cyclic mode of change, uncertainty, and search. Concepts from the literature on data reuse, science policy studies, and sociology of science studies, the methodological underlying assumptions of process theory (the sequence of conditions or events can affect the outcome, and values of conditions change over time) have also been part of my initial semantic¹¹⁰ coding ontology¹¹¹. I have modified this ontology in an inductive way by adding, disregarding and merging new codes and themes as I was doing the analysis.

For each of the theoretical concepts and the causal elements and conditions of the data-reuse mechanism, from which I have created the themes and codes of the ontology, I have looked for *occurrences* or *actual happenings*¹¹² (Abbott, 2001, p. 8), namely researchers' actions (unit of observation), which I have later translated into researchers' decisions (unit of data analysis). For capturing researchers' actions I have followed Poole and Van de Ven's key steps (2010). Therefore, I have identified actions, represented them, characterized action sequences, found temporal ordering and dependencies among actions, and finally I have fit all the findings to the theoretical concepts and to the causal forces of the data-reuse mechanism.

The data analysis process has followed a very similar pattern to the data collection process, although the data analyzed in each of the stage is different in each case study depending on the availability of

¹⁰⁸ “[t]he unit of data collection is the element on which each variable is measured. The unit of analysis is the element on which data are analyzed and for which findings are reported” (Neuendorf, 2002, p. 13)

¹⁰⁹ “In theory-guided qualitative research, it is important to prepare for the data analysis by deriving categories from the same theoretical framework that already has guided data collection” (Gläser & Laudel, 2013, p. 23).

¹¹⁰ Semantic codes are terms or sentences that represent what the participant has said or written without looking for underlying meanings or ideas of participants' words (Braun & Clarke, 2006).

¹¹¹ With NVivo Plus vs11.

¹¹² In the empirical data extracted from the data collecting instruments, i.e., participants' interviews and publications, my follow-up messages with them, the visual representation of researchers' data reuse processes, etc.

data sources at each of the stages. For instance, in case studies where the outcome is known –the reuse of secondary data was used as evidence of scientific claims–, I have analyzed researcher’s publications in the first stage. In cases where the outcome is unknown, I have analyzed the researcher’s publications –when possible– in the fifth or sixth stage. Since both data collection and data analysis have been carried out nearly simultaneously, both processes are explained together in the following section 4.7. A diachronic process of data collection and data analysis.

Cross-case analysis

In order to make cross-case inferences to a broader population of cases where data reuse happens, process-tracing methods cannot be used. We need other type of analysis, which lets us make these inferences (Beach & Pedersen, 2013). So, I have used a comparative analysis based on the principles of necessity and sufficiency of Ragin’s seminal work (e.g., Ragin, 1987, 1999). A necessary-or-sufficient-condition analysis seems to suit this study because it has a small number of cases, a small number of conditions, very concrete research questions, and it is theory driven (Kane, Lewis, Williams, & Kahwati, 2014).

4.7. A diachronic process of data collection and data analysis

This section includes a general overview of the process of collecting and analyzing data together for all ten case studies since these two processes happen simultaneously and are intertwined. Details of data collection instruments for each case study are provided in Chapter 5. General overview of cases and data sources collected for each case in from of a visual representation.

Both data collection and data analysis processes have been different in all the ten case studies mainly due to the characteristics of participants’ projects and their availability for participating in this study. Yet, there are some data-collection and data-analysis aspects, which are common to all of them, and some aspects, which are common to only some of them. Commonalities and differences are mainly because there are two main groups of case studies. On the one hand, case studies of ongoing research projects (OCS) –where we do not know the event or outcome (Y)–, and, on the other hand, case studies of finished research projects (FCS)–where we know the event or outcome (Y). So, occasionally, I may refer to OCS or FCS.

The most relevant aspect of both data collection and data analysis methods used in this study is their diachronic aspect in order to, on the one hand, counteract participants’ biases in the reconstruction of her decision-making process when telling the process of reusing the secondary data (Poole & Van de

Ven, 2010; Tsoukas, 2010). On the other hand, in order to track changes in conditions, and sequence of events over time (Abbott, 2001; Pettigrew, 1990; Salda. a, 2003). I collected and analyzed data at different stages over time, even for case studies where the reuse of secondary had concluded before the first and second interviews. Due to the diachronic data collection and data analysis process that I have used, my relationship with participants of some OCS lasts¹¹³ up to two years, more or less. I started collecting and analyzing data at the beginning of 2017, and at the end of 2019 or beginning of 2020, I still had contact with a few participants to trace their process and validate findings with them.

I have divided both the data collection and data analysis processes in five main stages (six stages in case study #4), although the amount of interactions with participants is very varied. Sometimes I could have some control on the time span between stages. Sometimes I could not. Needless to say is that these stages happened “when possible”. However, I will omit this comment to avoid repetition. Details of what was and was not possible at each stage –when relevant for findings and conclusions– are in Chapter 6. This applies also to the instruments gathered in each of the stages. Although I tried to gather and analyze all instruments in a very similar way, I was not successful in all case studies.

Stage #1

At this stage, I gathered as much information as possible about the participant, her research project, discipline, and the secondary data to answer the question(s) of her research. I did this in several ways. I requested this information to the participant; I searched or browsed the information online (e.g., LinkedIn, institutional web pages, data repositories’ websites, etc.); I read publications related to the project and to the secondary data, and publications that were research outputs from the reuse of secondary data¹¹⁴; I read about methods used by participants (e.g., about IPD MA and its differences with IPD NMA). In some cases, it was possible to have a face-to-face meeting with the potential participant to check whether the case study met the hereinabove sampling criteria, and to explain her my PhD research in more detail, and conditions of participation.

Stage #2

At stage #2, the first interview with the participant took place (there is a second interview at Stage #4). At the outset of the meeting, I explained the participant about my research and I got her consent regarding her participation in this study and for voice-recording the interviews. Second, I started

¹¹³ This does not mean that I have been in contact with participants along that time –which could be intrusive or annoying for them despite having their authorization to do it after deadline granted by ethics boards–.

¹¹⁴ In FCS this may only happen at this stage, but in OCS this happens in stage #5 or #6 (case study #4).

interviewing the participant with the interview script (annex 1) using it as a *loose*¹¹⁵ guide, strictly speaking. All participants authorized to voice-record the interviews. At the end of our meeting, I reminded the participant about the next step: I would analyze the conversation, draw her process of reusing data, and emailed her for scheduling the next interview.

The interview at this stage, and sometimes the preliminary meeting at stage #1 served me to realize if I had to interview someone else related to the reuse process. For example, in case study #4 (*released data*) I knew that it would be convenient to interview someone else during the preliminary face-to-face meeting that I had with the PI (principal investigator). However, in case study #6 (*stewarded data*, BORN Ontario) it was during the first face-to-face interview with the PI that I realized that I should interview the other person involved in the project. This does not mean that the PIs of cases #4 and #6 were not the “right person” (Baker & Edwards, 2012), but that the person, who actually played around with the data, should be also interviewed.

After interviewing the participant, I took notes regarding anything that I thought was interesting for the analysis. For example, I wrote down how comfortable I was, how the interviewee felt, how candid I thought she was, if I had to look for some extra information to understand something I did not fully understand, whether I was a “good listener”, whether I was able to create an atmosphere of trust, etc.

Stage #3

This stage was mainly an analysis stage. First, I transcribed some of the interviews. Most of them was professionally transcribed, yet reviewed carefully by myself. Second, I did a preliminary¹¹⁶ but thorough analysis of the first interviews’ transcriptions together with my notes during and after the interview. In case study #4 (*released data*), I also analyzed the preliminary face-to-face meeting with the PI because he gave me valuable information about his research project and the reuse of secondary data. The analysis allowed me to identify information gaps about the data-reuse process itself and about the participant’s decision-making process, to draw¹¹⁷ a visual representation of the data-reuse

¹¹⁵ I never read the interview script word-for-word as I was supposed to do according to the ethics protocol of the Ottawa Health Science Network Research Ethics Board/ Conseil d’éthique de la recherche du réseau de science de la santé d’Ottawa (OHSN REB). The main reason is that, fortunately before making any interview, I realized that my interview script or guided asymmetrical conversation with participants (Brinkmann, 2018) was *too much guided*. I realized how much contaminated I was from the literature I had read about reusing data, information or knowledge, in which, at least within a data reuse process, there are several stages, i.e., data search and discovery, access, selection, preparation and analysis (Faniel et al., 2012; Rolland & Lee, 2013; Zimmerman, 2007, 2008). The interview script was too biased toward a linear process of data-searching, data-analyzing and finally data-using. So, I completely changed the order of questions and how I made them. For example, instead of asking questions grouped in the different stages I biasedly conceived, I asked the participant to tell the process of how reuse of the secondary data in her research project happened. Had I acted unethically according to OHSN REB or the University of Ottawa Ethics Board, it is my sole responsibility.

¹¹⁶ I have written “preliminary” because in order to write the findings, I conducted two more analyses of the interviews’ transcripts.

¹¹⁷ I used Microsoft Visio for drawing the data-reuse process.

process, and to prepare the interview script for the second interview (stage #4). The analysis, the drawing of the visual representation of the participant's process of reusing data, and the preparation of ad-hoc questions for the second interview took me approximately from 6 to 10 hours, not included the time of the transcription. The visual representation includes two main types of data. On the one hand, a simplified account of the data-reuse steps and some of the conditions under they happen. On the other hand, some participants' quotes, which I found interesting at least at that time. These quotes are in quotation marks and, most of the times, shadowed in light blue color.

Apart from questions about information gaps in the participant's account of the data-reuse process, I tried systematically to include in the second interview a hypothetical situation where one or two theorized parts or causal powers of the data-reuse mechanism would be different for the participant at the time of making a decision. This hypothetical situation was especially useful for the FCS where I could not follow the participant's data-reuse process and thus "to "follow the action", that is, the sequence of actions or events that lead to the decision from the beginning to end" (Poole & Van de Ven, 2010, p. 559).

Also, during the interview at stage #2 some participants mentioned some documents, e.g., research protocols, publications, etc., which were part of their research projects. I could have access to some of these documents, which I also analyzed for preparing the following interview.

Stage #4

At this stage, a second face-to-face interview took place. In most case studies, it was shorter than the first one, but in a couple of case studies, it was longer than the first one. I tried to leave a time span of at least 10 or 15 days between the first interview and this second one because the analysis and preparation of the second interview required some time, and because of OCS since some participants' decisions with regard to the reuse of secondary data could have changed during that time span. I also tried to conduct no interview from other case study during the time span of the two interviews of one concrete case study because of my concern of mixing up information of the two cases. This strategy of concluding the data collection in a case study by case study way gave me also the opportunity to improve the interview method and strategy in the following cases in both first and second interviews.

The main three goals of this second interview were, first, to fill in the information and details gaps that I needed to identify the sequence of the participant's decision-making; second, to lay out an hypothetical situation –for which I thought that the participant would have taken a different step or path in her decision-making process–; and third, but not least, to validate the visual representation of the process of reusing data.

For the hypothetical situation, I always tried to lay out a situation plausible for the participant. It was not an easy task. In some cases, I laid out a different situation related to the participant's structure, her

causal powers and liabilities, or the secondary data's causal powers and liabilities. In other cases, I laid out a hypothetical change in the conditions (C1, C2, C3, C4, and C5) under which the theorized data-reuse mechanism works.

For the validation of the visual representation, I took two color-printed copies of it to the interview place. I showed the diagram that I depicted to participants and asked them to do two things. First, to rectify any incorrectness and add some details with a red pen that I took with me, and, second, to comment aloud their corrections. In general and surprisingly, all participants liked the drawing. They saw a quite accurate representation of their own processes of using secondary data. In fact, one of participants (case study #9 of the group of *proprietary data*) asked me the drawing for herself because she thought it was also useful for herself.

At this stage, for the *stewarded data-reuse* case studies (cases #5, #6, and #7), I decided interviewing staff of BORN Ontario data and staff at ICES data (for case #5) in order to have a better picture of the data's causal powers and liabilities.

Stage #5

Stage #5 is both an analysis and follow-up stage. It starts the day after the second face-to-face interview –or first interview in cases where a second interview was not possible–. The follow-up with participants consisted mainly in asking them how their research projects or questions with secondary data were evolving, and the reasons for the way they were evolving. I requested them this information by email, or paying attention to their publications on their online institutional profiles, or in research networks such as ResearchGate. In one case study, the analysis of the secondary data had not concluded at the time of writing this dissertation at the end of 2019 and beginning of 2020. In other cases, although the analysis of secondary data was concluded and sent for publication, I read and analyzed such publications since they are both evidence of the outcome 2 or outcome 3 (see Figure 4).

Analysis and re-analysis of interviews and other documents (their emails with clarification and updates, their publications, research protocols, for example) has lasted until the time of writing the findings and discussion sections of this dissertation.

This stage also includes a validation exercise with all participants about their decision making process when using secondary data. In January and February 2020, I emailed participants to share with them the narrative of my findings regarding their decision making process, together with their validated workflow diagram of their process of reusing data, and the data collection instruments and dates. Except for one case study (#6), they all answered me agreeing with the narrative, and confirming or providing some small details of the process of reusing data.

4.8. Ethics protocols and data sharing agreements

I have followed both ethic protocols and other administrative norms of both the Ottawa Health Research Institute/L'Hôpital d'Ottawa Institut de Recherche and the Université d'Ottawa/University of Ottawa. As part of the ethics requirements in Canada, I had to take an online course on research ethics, namely the Tri-Council Policy Statement: Ethical conduct for research involving humans (TCPS 2: CORE)¹¹⁸. A certificate of completion was issued July 31, 2016.

At the Ottawa Health Research Institute / L'Hôpital d'Ottawa Institut de Recherche (Ottawa, Canada) this research was approved by the Ottawa Health Science Network Research Ethics Board/ Conseil d'éthique de la recherche du réseau de science de la santé d'Ottawa¹¹⁹ with protocol number 20160949-01H. At the Université d'Ottawa / University of Ottawa this research was approved by the Bureau d'éthique et d'intégrité de la recherche / Office of Research Ethics and Integrity¹²⁰ with protocol number A01-17-0.

Furthermore, a data sharing agreement between the Ottawa Health Research Institute/L'Hôpital d'Ottawa Institut de Recherche and the Université d'Ottawa/University of Ottawa was signed as a pre-condition for collecting data.

4.9. A small exercise on reflexivity

Yo soy yo y mi circunstancia, [...]

José Ortega y Gasset, in *Meditaciones del Quijote* (1914)

“Reflexivity is the process of reflecting critically on the self as a researcher” (Guba & Lincoln, 2005, p. 210), and to recognize that we, as social scientists, are beleaguered by the preconstructed and, thus, that we borrow problems and concepts from the social world (Bourdieu & Wacquant, 1992). Thus, I have to admit that...

¹¹⁸ <https://tcps2core.ca/welcome>

¹¹⁹ <http://www.ohri.ca/ohsn-reb/>

¹²⁰ <https://research.uottawa.ca/ethics/reb>

... I am an information scientist playing a sociologist, being inspired by historians' empirical work in archives, and influenced by studies on science policy.

... I am embedded in the Open Science movement.

... I belong to the "datafication society"¹²¹ or to the dataverse, as Bowker calls it (Bowker, 2013)

... I live within a woman-empowering discourse and social movement.

... I believe that human beings can do more with motivation, efforts and willingness than with only means, though a minimum amount of the latter is still necessary.

... I have been thinking of information scientists, STS scholars, and science sociologists as the potential audience interested in the results of this dissertation.

... I use visual representations for nearly everything quotidian in my life, if not all, when I want to understand or explain something, be simple or complex.

... when I analyze something, I analyze it in terms of relations and processes, and not in terms of static objects or concepts.

... I have been working with WOP (work, organizational, and personnel) psychologists for six years.

... I firmly believe that we can understand social phenomena better if we draw on different disciplines' theories or concepts. Thus, I am prone to drawing on different disciplines when answering research questions.

The above reflexive lines are a small exercise of how by background, the way I view the world, and the socio political moment of science that I live in, may have influenced my choice of the topic, the methods, conclusions, my narrative, and my focus or perspective to approach the topic. The goal of this exercise is not to search for objectivity or to apologize for a biased research. Nothing further than this. With this exercise, I presuppose that the reader will understand better the rationale of this

¹²¹ Had I done my PhD in the 70s or 90s, my research topic would have been about "information" (Bell, 1976) or "knowledge" (UNESCO, 2005) respectively.

dissertation. However, I prefer that the reader concludes how *I and my circumstances* have affected this research. If I conclude it, I might bias or cloud the readers' own conclusions.

Two caveats are also necessary on this dissertation. On the one hand, all processes in this study –the review of the literature, the formulation of the research questions, the collection and analysis of the data, the choice and development of the theory, etc.– have not followed a straightforward plan despite I am presenting them in a linear, logical, sequential way in order to fulfill with scientific discourse standards. In reality, conducting this research has been quite the opposite as Abbott explains for any research project to happen (Abbott, 2004)¹²². Moreover, my initial research proposal could not be tightly planned or followed since one of the characteristics of “qualitative” research is its emergent design (Creswell, 2014). Therefore, I have recast data, methods, and theory constantly and in a looping way until the very last moment of writing up this thesis.

On the other hand, I have omitted or black-boxed some small details and I have practiced *retrospective falsification* (Vinck, 2010, p. 154) in a couple of aspects in this dissertation. Had I not done this, I might have compromised some people or institutions. However, neither the black-boxed details nor the retrospective falsification affects the answer to the research questions that guide this dissertation.

¹²² “Most teaching on methods assumes that the [researcher] will start a research project with a general question, then narrow that to a focused question, which will dictate the kind of data needed, which will in turn support an analysis designed to answer the focused question. Nothing could be further from reality. Most research projects—from first-year undergraduate papers to midcareer multiyear, multi-investigator projects start out as general interests in an area tied up with hazy notions about some possible data, a preference for this or that kind of method, and as often as not a preference for certain kinds of results. Most research projects advance on all of these fronts at once, the data getting better as the question gets more focused, the methods more firmly decided, and the results more precise.” (Abbott, 2004, p. 83-84)

Chapter 5

General overview of cases and data sources collected for each case

This chapter provides a general overview, details and collection dates of the data sources that I have used in the ten case studies.

The general overview is presented in a table and shows how each case study meets the sampling criteria at the time of selection (grey shadowed table cells). It also includes characteristics of the cases not purposely sampled (light blue shadowed table cells). There are three tables, since the ten cases are grouped into the three above mentioned categories based on availability and accessibility of the data: *released data*, *stewarded data*, and *proprietary data*. Detailed information about participants, data and conditions related to the data-reuse mechanism of each case study is presented in Chapter 6., which includes the empirical analysis.

Regarding details and collection dates of data sources of the ten case studies, they are included in ten independent visual representations, one for each case study. The upper half section of the visual representation includes the five or six stages of the process of collecting data. At each stage, there are different images or symbols that represent each of the instruments or data sources that I have used for collecting data in each of the case studies. Table 2 includes the keys of these images and symbols. The













lower half section of the visual representation includes a time line with the most important dates when I interacted with each participant¹²³.

During the validation process of findings regarding the decision-making process, some participants authorized me to use their original names. All interviews were conducted in English except the ones with the two participants of case study #4, which were conducted in Valencian¹²⁴.

¹²³ It also includes the total number of messages I exchanged with participants regarding their research question or project with secondary data.

¹²⁴ One of the official languages in Spain. For reporting findings, I have not translated participants' quotes. I will translate them upon request and whether I have the time and resources to do it.

Table 2 - Key of the images and symbols used in the visual representation of data collection instruments and dates

 <p>Reading of publications, research protocols, ethics protocols, participants' online profiles, etc.</p>	 <p>No reading of publications, research protocols, ethics protocols, participants' online profiles, etc.</p>
 <p>Exchange of email messages with participant takes place</p>	 <p>Exchange of email messages with participant does not take place</p>
 <p>Checking of the data repository online</p>	 <p>No checking of the data repository online</p>
 <p>Visual representation of the process of reusing data is validated by the participant(s)</p>	 <p>Visual representation of the process of reusing data is not validated by the participant(s)</p>
 <p>Face-to-face interview takes place</p>	 <p>Face-to-face interview does not take place</p>
 <p>Interview on Skype takes place</p>	 <p>Time line. It includes the dates of the most important interactions with participants</p>

5.1. Case studies reusing *released data*

Released data—or publicly released data—are data that are publicly released or published, and available for being reused with, in principle, no restriction other than the polite request of citing the original publication related to the data set, the data set accession number or both, and sometimes the data repository.

Under this category, there are four case studies. Three of them (#1, #2, #3) are located in Canada, and one (#4) in Spain. All of them belong to the field of cell biology applied to cancer. Participant of cases in Canada (David Cook) belongs to a research laboratory (web lab) in molecular biology applied to ovarian cancer. In the case study in Spain there are two participants. Joan Climent does mainly molecular biology in a web lab, while Jaume Forés does computational biology (dry lab), both applied mainly to breast cancer.

The participant of case studies #1, #2 and #3 is the same, and his reuse of data is from different data repository (GTEx, GEO, and TCGA) in each of the cases. Participant of case #4 reuses data from GEO and TCGA data repositories. These data repositories store gene expression profiles from curated data sets, and genomic, epigenomic, transcriptomic, and proteomic data.

In order to get familiar with data reuse in the field of molecular biology, I read some articles, related to data both sharing and reuse, for instance, Hilgartner, 1995; Hilgartner, 1998; Hilgartner 2017; Hilgartner & Brandt-Rauf, 1994; Ioannidis et al., 2009; Kahlem & Birney, 2006; Kaye, Heeney, Hawkins, de Vries, & Boddington, 2009.

Figure 5 shows visually how the four case studies of *released data* are related to each other. Green arrows relates each participant with the data repository that they have reused. With the three case studies in Canada, I can compare decisions on data reuse and test the data-reuse mechanism making sure that the researcher's causal powers and liabilities are the same in principle¹²⁵. With case study #2, #3, and #4, I can compare decisions on data reuse and test the data-reuse mechanism making sure that the data's causal powers and liabilities are the same in principle.

¹²⁵ Sayer reminds us that causal powers and liabilities are not eternal and can change over time (Sayer, 2010)

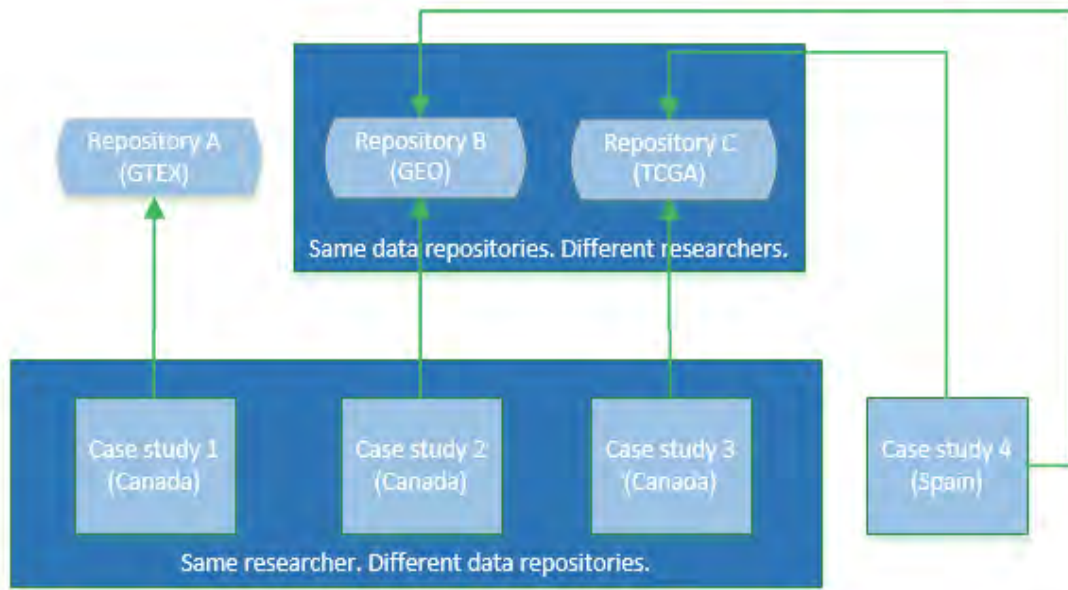


Figure 5 - Four case studies of reuse of “released data”. Same researcher in case studies #1, #2, #3 reuses data from three different repositories (A, B, C). Researcher in case study #4 reuses data from repositories B and C.

Table 3 shows the variability of the case studies #1, #2, #3 and #4 reusing *released data*. Grey shadowed cells refer to the variability based on the initial sampling criteria. Light blue shadowed cells refer to characteristics of the case studies, which were not sampled purposely.

Table 3 - Variability of case studies reusing "released data"

	Case 1	Case 2	Case 3	Case 4
C2 – Secondary data are obtained	Released data	Released data	Released data	Released data
C3 - Particular secondary data an initial satisficing option	Yes	Yes	Yes	Yes
C5 – An expected scientific contribution or career milestone exists	Yes	Yes	Yes	Yes
Is the outcome or event (Y) known or unknown?	Known	Unknown	Known	Unknown
Material object: researcher	David Cook (real name)	Same as case #1	Same as case #1	Jaume Forés and Joan Climent (real names)
Material object: data repository	GTEX	GEO Profiles	TCGA	GEO Profiles and TCGA
Fulfills the definition of reuse of secondary data	Yes	Yes	Yes	Yes
Health discipline	Yes	Yes	Yes	Yes
Ongoing or finished	Finished	Ongoing	Finished	Ongoing
Name of project or research question	<i>Chromatin regulators: jacks of all states</i>	<i>Transcriptional and epigenetic determinants of the epithelial-to-mesenchymal transition in ovarian cancer</i>	<i>Undetermined name (related to cancer)</i>	<i>Gene and biological relationship between autism spectrum and cancer; the role of TRIM29</i>
Is the research question of the secondary user different from the research question, which motivated the collection of the data?	Yes	Yes	Yes	Yes
Discipline - field	Molecular biology (applied to research on ovarian cancer)	Same as case #1	Same as case #1	Molecular-computational biology (applied to research on breast cancer)
Country	Canada	Canada	Canada	Spain

5.1.1. Collected empirical data and collection dates in case study #1 (GTEX data repository)

Figure 6 shows that I conducted three interviews with my participant. See annex 2 for a full-sized image of Figure 6. Interview at stage #1 was a short preliminary face-to-face meeting to check conditions for eligibility in this study, and to explain conditions of participation. In this preliminary interview, conditions for eligibility of the next reported case studies #2 and #3 were also checked since the participant is the same.

Except for the GTEX portal, I did not analyzed any other data source for this case study since there was none.

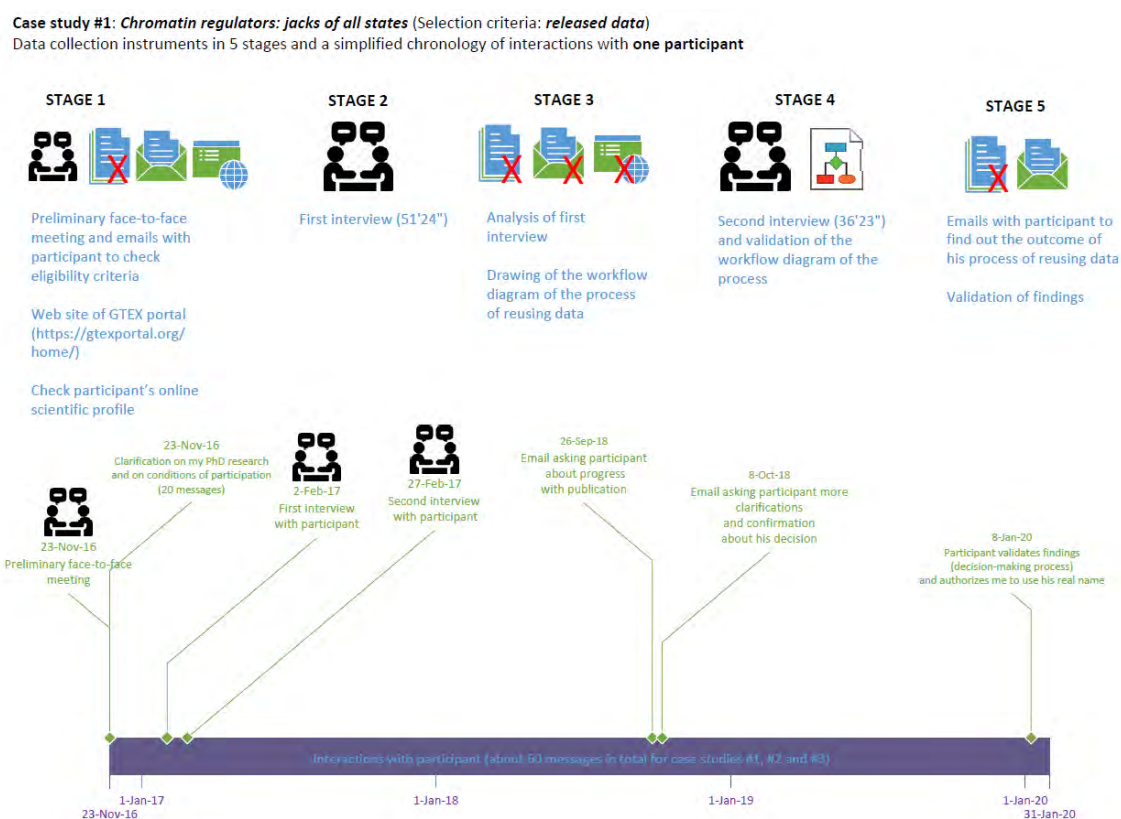


Figure 6 - Data collection instruments and dates. Case study #1

5.1.2. Collected empirical data and collection dates in case study #2 (GEO Profiles repository)

Participant of this case study is the same as of case study #1. In this case study, there are two relevant documents, which were analyzed. One document was the participant's abstract of his PhD research proposal, which I analyzed at stage #3. In his proposal, he did not mention any secondary data or any data repository. The other document is a scientific publication related to the reuse of data from the GEO Profiles repository, namely (Cook & Vanderhyden, 2019).

Figure 7 shows the data collection sources, stages and dates in a timeline of three years. See annex 3 for a full-sized image of Figure 7.

Case study #2: Transcriptional and epigenetic determinants of the epithelial-to-mesenchymal transition in ovarian cancer (Selection criteria: *released data*)
Data collection instruments in 5 stages and a simplified chronology of interactions with **one participant**

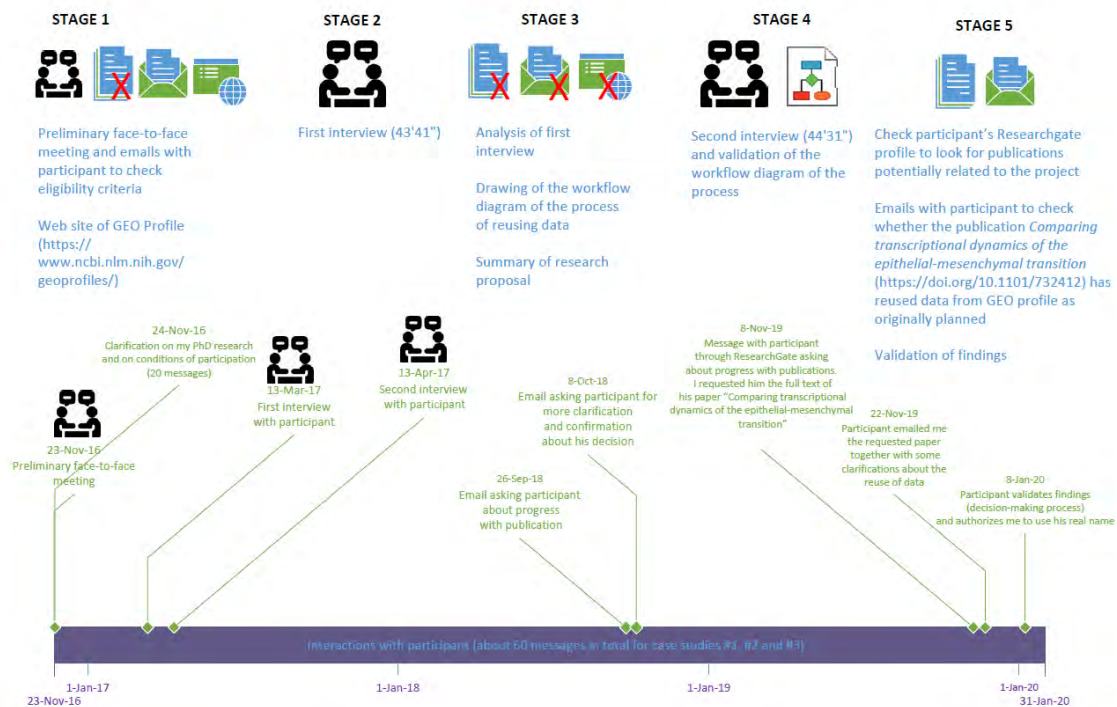


Figure 7 - Data collection instruments and dates. Case study #2

5.1.3. Collected empirical data and collection dates in case study #3 (TCGA data repository)

This case study is initially a case of no reuse of data after trying to. However, my participant finally reused the data after some time as explained in section 6.1.3. We only talked about this case study at stage #2. I did not draw any workflow diagram of the process of reusing the data, and thus there was not validation of it by the participant. I analyzed no documents, i.e., research proposal or scientific publication, because the reuse of data from the TCGA portal were aimed to create background knowledge and there were no related documents to the reuse.

At stage #5 as shown in Figure 8, I emailed my findings of the decision-making process of this case study with my participant and he agreed with my account of the process. See a full size of Figure 8 in annex 4.

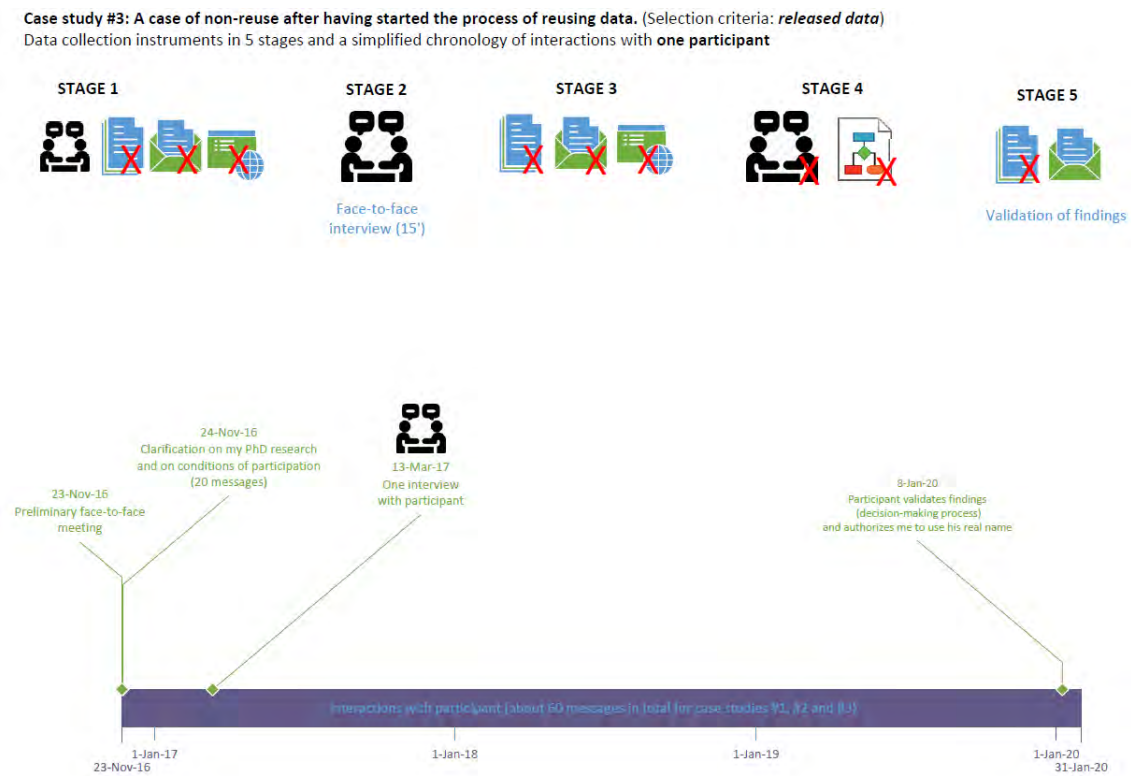


Figure 8 - Data collection instruments and dates. Case study #3

5.1.4. Collected empirical data and collection dates in case study #4 (GEO Profiles and TCGA repositories)

As shown in Figure 9 (full size in annex 5), this case study includes six data collection and analysis stages, unlike the rest of case studies, which include five stages. The main reason is that I interviewed two participants and I wanted to validate the process of reusing data with both of them. Participant A refers was the principal investigator and participant B was to the actual secondary analyst of the data, who was a PhD candidate at the time of the interviews.

There was also a preliminary interview at stage #1 to check eligibility criteria. At stage #6, both participants validated my findings about their decision-making process, and authorized me to use their real names in this dissertation.

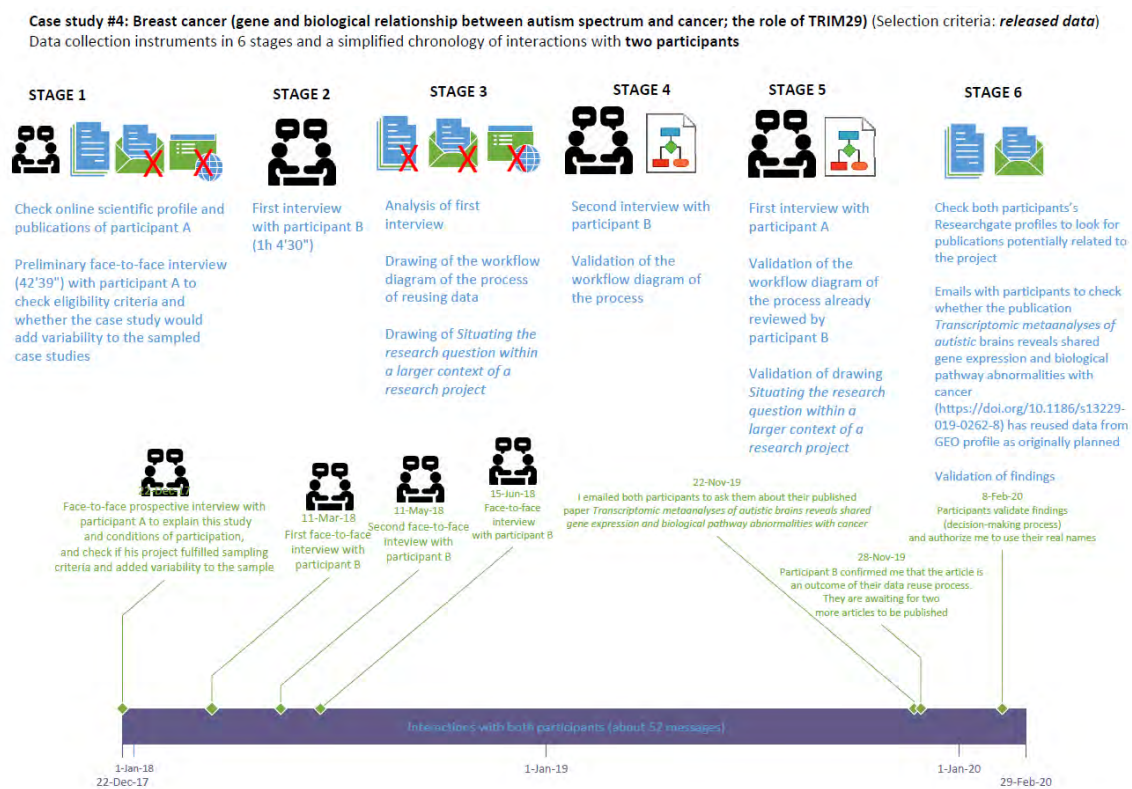


Figure 9 - Data collection instruments and dates. Case study #4

5.2. Case studies reusing *stewarded data*

Stewarded data refer to data that are available for reuse, but are not publicly released or published. Data are available for others to reuse them, but there may be some type of walls, e.g., payment walls, confidentiality walls, technical walls, in order to reuse them, or conditions on the reuse.

Under this category, there are three case studies (#5, #6, and #7). All of them are located in Canada and belong to the field of epidemiology exploring clinical issues related to birth and pregnancy health with data from BORN Ontario data repository (Better Outcomes Registry & Network / Registre et Réseau des Bons Résultats dès la naissance).

The Better Outcomes Registry & Network (BORN) is Ontario's prescribed maternal, newborn and child registry with the role of facilitating quality care for families across the province. BORN collects, interprets, shares and rigorously protects high-quality data essential to making Ontario the safest place in the world to have a baby.

BORN is funded by the Ontario Ministry of Health and Long Term Care, administered by the Children's Hospital of Eastern Ontario (CHEO) and active in every region of the province. As a prescribed registry under Ontario's Personal Health Information Protection Act, BORN safeguards data while making information available to facilitate and improve the provision of healthcare. To ensure all personal information is protected in accordance with privacy legislation and data-system standards, BORN is overseen by the Information Privacy Commissioner of Ontario. (Source: BORN Ontario's web site¹²⁶)

A serendipitous interesting variability across these three case studies is that participants of the three case studies have a different relationship with BORN Ontario data repository (BORN from now on). For example, researcher of case study #5, Deshayne Fell, is BORN staff. Researcher of case study #6, Mary Smith, is an external user of these data, and researcher of case study #7, Sarah Wilson, is a BORN *agent*.

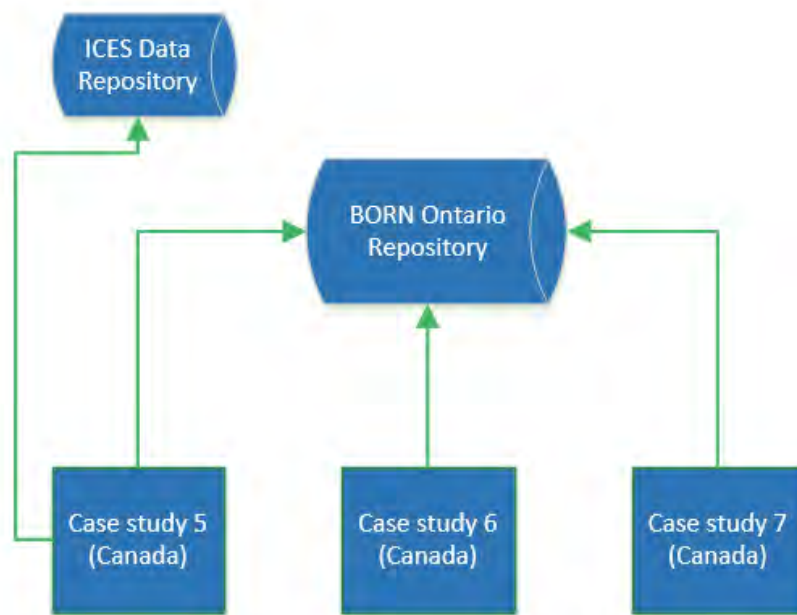
Deshayne Fell (case study #5) is an interesting case study where we have the same researcher's structure and causal powers and liabilities for different secondary data repositories' structures and causal powers and liabilities –BORN Ontario data and ICES data–. ICES, the Institute for Clinical Evaluative Sciences is a *not-for-profit research institute encompassing a community of research, data and clinical experts, and a secure and accessible array of Ontario's health-related data* (Source: ICES's web site)¹²⁷.

¹²⁶ <https://www.bornontario.ca/en/about-born/about-born.aspx>

¹²⁷ <https://www.ices.on.ca/About-ICES/Mission-vision-and-values>

Also, Sarah Wilson in case study #7, provided other interesting variation within these three cases of reuse of stewarded data. Unlike Deshayne Fell and Mary Smith, Sarah Wilson is not an academic researcher, but a government researcher. She conducts research in a public health office.

Apart from interviewing each of the researchers in the three case studies, I also interviewed one employee at BORN (Sandra Dunn) and one employee at ICES (John Davidson). Figure 10 shows the relationship of the three cases with BORN Ontario data, and one of the cases (#5) with ICES data. BORN Ontario data have the same causal powers and liabilities in the three case studies.



Same data repository. Different researchers.

Figure 10 - Three case studies of reuse of "stewarded data". Three different researchers (case studies #5, #6, #7) reuse data from the same repository (BORN Ontario). One researcher (case study #5) reuses also data from ICES repository

Table 4 shows the variability of the case studies of *stewarded data*. Grey shadowed cells refer to the variability based on the initial sampling criteria. Light blue shadowed cells refer to characteristics of the case studies, which were not sampled purposely.

Table 4 - Variability of case studies reusing "stewarded data"

	Case 5	Case 6	Case 7
C2 – Secondary data are obtained	Stewarded data	Stewarded data	Stewarded data
C3 - Particular secondary data an initial satisficing option	Yes	Yes	Yes
C5 – An expected scientific contribution or career milestone exists	Yes	Yes	Yes
Is the outcome or event (Y) known or unknown?	Known	Known	Unknown
Material object: researcher	Deshayne Fell (real name)	Mary Smith (pseudonym)	Sarah Wilson (pseudonym)
Material object: data repository	BORN Ontario ICES	BORN Ontario	BORN Ontario
Fulfills the definition of reuse of secondary data	Yes	Yes	Yes
Health discipline	Yes	Yes	Yes
Ongoing or finished	Finished	Finished	Ongoing
Researcher's relation with BORN Ontario	BORN Employee	BORN Agent	External user of BORN
Name of project or research question	<i>Influenza illness and influenza vaccination during pregnancy and risk of preterm birth and fetal death</i>	<i>The effect of maternal obesity on stillbirth and neonatal death</i>	<i>A research study in epidemiology related to maternal and neonatal health</i>
Is the research question of the secondary user different from the research question, which motivated the collection of the data?	It does not apply	It does not apply	It does not apply
Discipline - field	(Perinatal) Epidemiology	Epidemiology	Epidemiology
Country	Canada	Canada	Canada

5.2.1. Collected empirical data and collection dates in case study #5 (BORN Ontario data & ICES data)

The reuse of both BORN data and ICES data had already happened when I interviewed participant in this case study. So, as Figure 11 shows, at stage #1 I analyzed the scientific manuscript *The relationship between 2009 pandemic H1N1 influenza during pregnancy and perinatal outcomes in Ontario*, which was part of my participant’s PhD dissertation (Fell, 2015) and was finally published in 2018 (Fell et al., 2018).

Furthermore, I interviewed BORN Ontario staff and ICES staff at stage 4 in order to understand better my participant’s process of reusing the data and her decision-making process.

See the full-sized Figure 11 in annex 6.

Case study #5: The relationship between 2009 pandemic H1N1 influenza during pregnancy and perinatal outcomes in Ontario, manuscript 2 of PhD thesis titled *Influenza illness and influenza vaccination during pregnancy and risk of preterm birth and fetal death, December 2015* (Selection criteria: *stewarded data*)
Data collection instruments in 5 stages and a simplified chronology of interactions with **one participant**

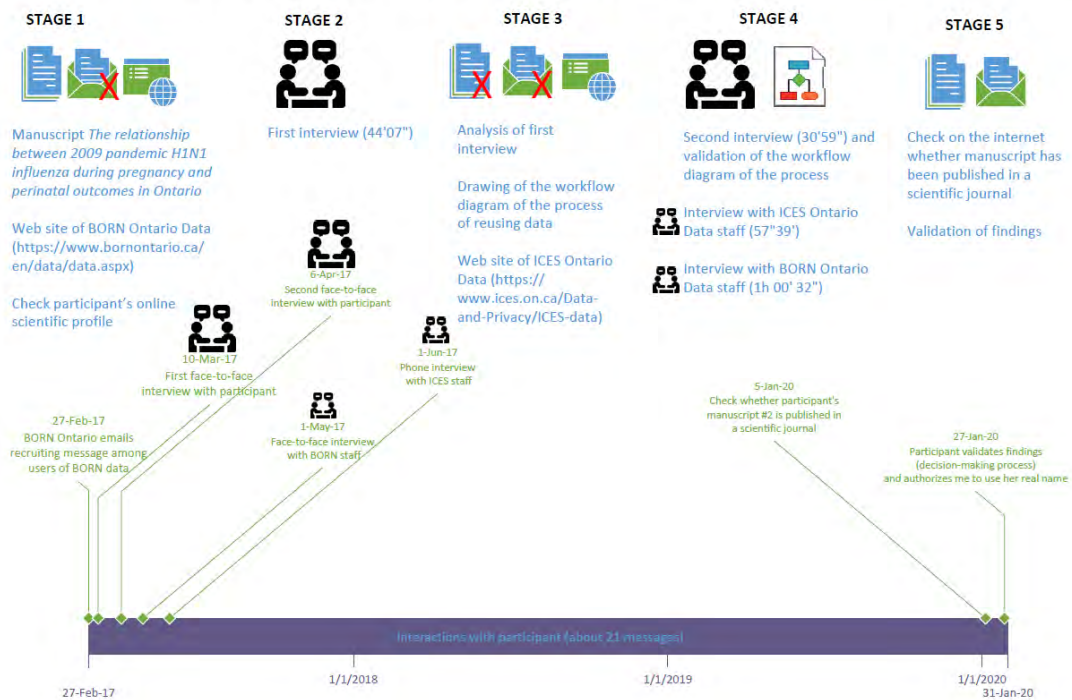


Figure 11 - Data collection instruments and dates. Case study #5

5.2.2. Collected empirical data and collection dates in case study #6 (BORN Ontario data)

In this case study, and as Figure 12 shows, there is no data collection or analysis at stage #4 since I could not interview my participant for a second time. At stage #5, apart from trying to find publications related to the case study, I emailed my participant to ask her about the outcome of her process of reusing BORN data, and I shared with her my findings. Her response was unclear.

My interview with BORN Ontario staff mentioned in the previous case study also served me for the analysis of this case study, although Figure 12 does not include this interview.

See the full-sized Figure 12 in annex 7.

Case study #6: The effect of maternal obesity on stillbirth and neonatal death (Selection criteria: *stewarded data*)
Data collection instruments in 5 stages and a simplified chronology of interactions with **1 participant**

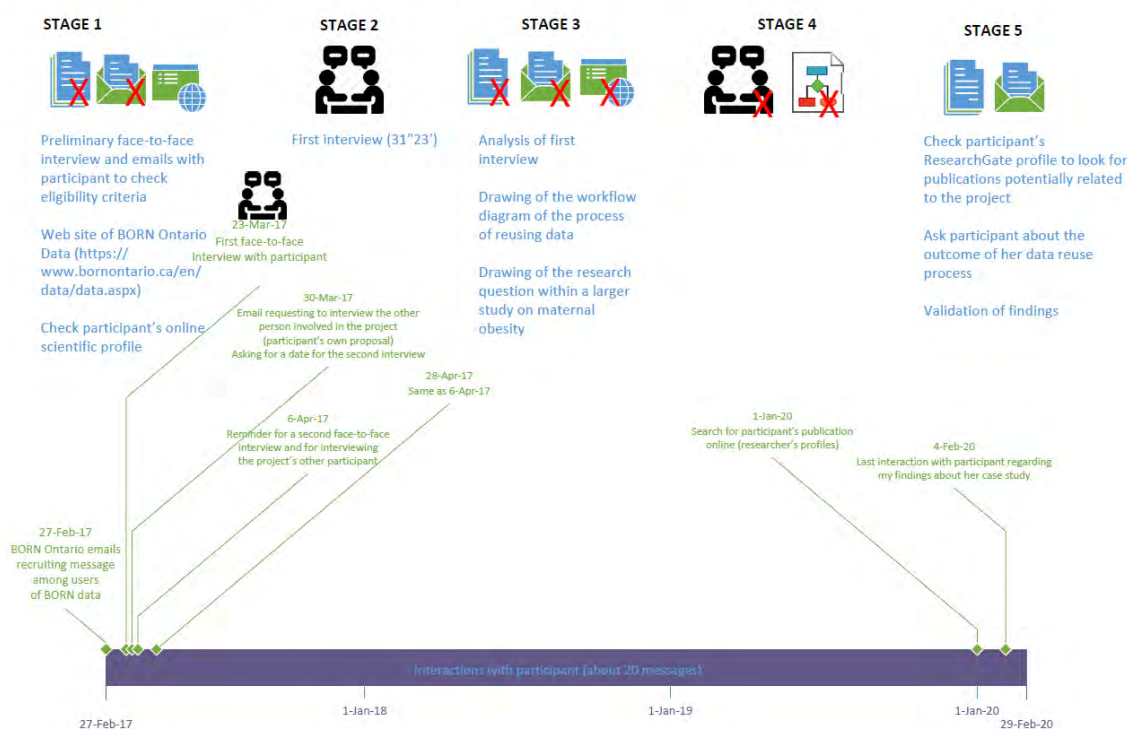


Figure 12 - Data collection instruments and dates. Case study #6

5.2.3. Collected empirical data and collection dates in case study #7 (BORN Ontario data)

In this case study, there is data collection and data analysis at all stages (Figure 13 or annex 8). At stage #3, I analyzed my participant’s research protocol and a presentation she made at the BORN Annual Conference 2017. There was a preliminary face-to-face meeting to check the eligibility of this case study and to explain conditions of participation to the researcher.

Case study #7: A research study in epidemiology related to maternal and neonatal health (Selection criteria: *stewarded data*)
Data collection instruments in 5 stages and a simplified chronology of interactions with 1 participant

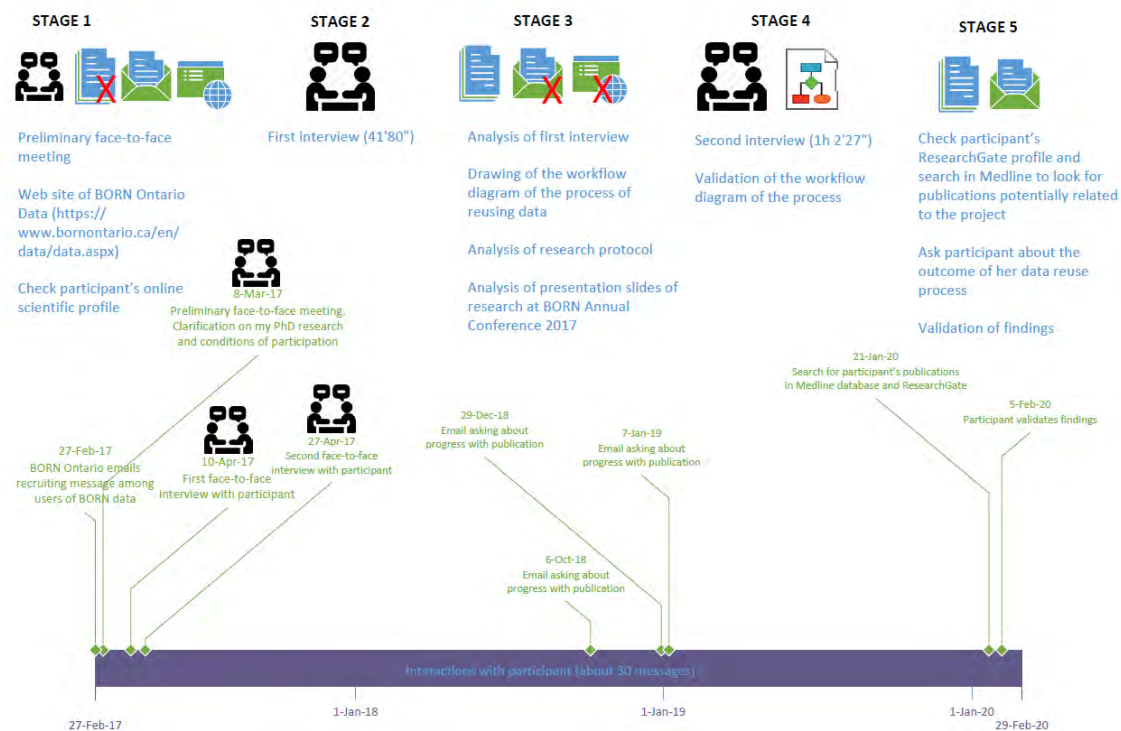
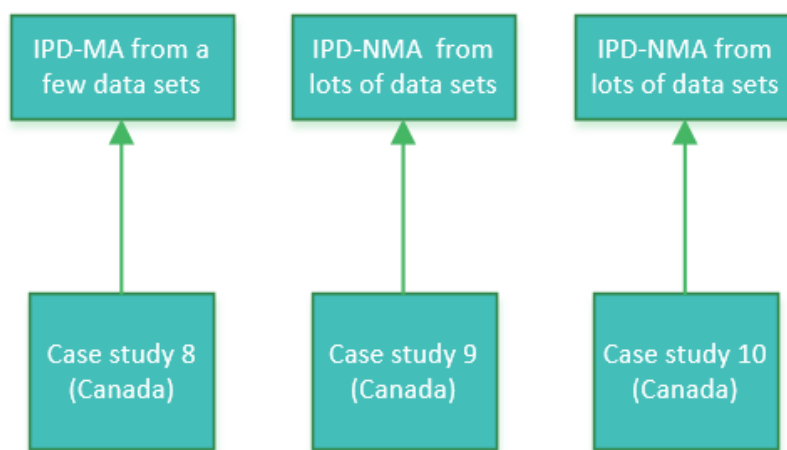


Figure 13 - Data collection instruments and dates. Case study #7

5.3. Case studies reusing proprietary data

Proprietary data are data that have not been publicly released, and their availability for being reused is uncertain. This means that data, after being requested, cannot be obtained from the data owners for any reason, be technical, ethical, legal, etc. Data owners are not necessarily the person or organization that collected the data, but the public or private organization, which funds the collection of the data.

Under this category, there are three case studies (#8, #9, and #10). All of them are located in Canada and belong to different fields. However, they all have in common that they follow a standardized methods protocol to conduct the studies. Case study #8 is an individual patient or participant data meta-analysis (IPD MA) study, and case studies #9 and #10 are individual participant data network meta-analysis (IPD NMA). IPD meta-analyses started to be conducted in the early 1990s. The main difference with meta-analyses of secondary data obtained from publications is that IPD meta-analyses need the “raw” individual data of the original studies on a specific topic (Stewart & Michael, 1995). They usually follow a statement (PRISMA-IPD, Preferred Reporting Items for Systematic reviews and Meta-Analyses) as a good practice (Simmonds, Stewart, & Stewart, 2015). IPD NMA are a type of IPD MA that have the ability to compare all treatments for a specific health problem (Brignardello-Petersen, Rochwerg, & Guyatt, 2014). An IPD meta-analysis, be network or not, can be considered an extreme case of data reuse within the category of *proprietary data* since the analysis needs a few or many “raw” data sets. IPD MA’s are very challenging to conduct because of the time and resources needed to contact original authors and request them the original individual data, among other things (Riet, Bachmann, Kessels, & Khan, 2013; van Walraven, 2010). Failure to obtain all data sets or some of the data sets, the IPD MA cannot be carried out, e.g., Jaspers & Degraeuwe, 2014. Figure 14 shows that there is no relationship between the case studies or the data sets, except for the fact that they use the PRISMA-IPD in their methodology. Researchers’ and data’s causal powers and liabilities of these three case studies are different.



Many different data sets. Different researchers.

Figure 14 – Three cases studies of reuse of “proprietary data”. Three different researchers, rather a research team, (case studies #8, #9, #10) reuse individual participant data (IPD) from different data sets in three different health problems

Table 5 shows the variability of the case studies of *proprietary data*. Grey shadowed cells refer to the variability based on the initial sampling criteria. Light blue shadowed cells refer to characteristics of the case studies, which were not sampled purposely.

Table 5 - Variability of case studies reusing "proprietary data"

	Case 8	Case 9	Case 10
C2 – Secondary data are obtained	Proprietary data	Proprietary data	Proprietary data
C3 - Particular secondary data an initial satisficing option	Yes	Yes	Yes
C5 – An expected scientific contribution or career milestone exists	Yes	Yes	Yes
Is the outcome or event (Y) known or unknown?	Known	Unknown	Unknown
Material object: researcher	Nicole Langlois (real name)	Claire Johnson (pseudonym)	Areti Angeliki Veroniki (real name)
Material object: data repository	Data sets from 8 randomized clinical trials (RCT)	Data sets from 67 RCTs	Data sets from 108 RCTs
Fulfills the definition of reuse of secondary data	Yes	Yes	Yes
Health discipline	Yes	Yes	Yes
Ongoing or finished	Finished	Ongoing	Ongoing
Name of project or research question	<i>Low-molecular-weight heparin and recurrent placenta-mediated pregnancy complications: a meta-analysis of individual patient data from randomised controlled trials</i>	<i>Tentative title: Mass deworming to improve developmental health and wellbeing of children in low-income and middle-income countries</i>	<i>Tentative title: Comparative safety and effectiveness of cognitive enhancers for Alzheimer's dementia: a systematic review and individual patient data network meta-analysis</i>
Is the research question of the secondary user different from the research question, which motivated the collection of the data?	No (IPD MA)	No (IPD NMA)	No (IPD NMA)
Discipline - field	Clinical in nature but it can fall into epidemiology	Epidemiology. Global health.	Clinical/health in nature but it can fall into epidemiology
Country	Canada	Canada	Canada

5.3.1. Collected empirical data and collection dates in case study #8 (IPD MA)

At stage #1 before the first interview with participant, I analyzed three articles related to the process of reusing individual participant data (IPD) from other studies, namely Rodger et al., 2014, 2016, 2015. This analysis let me have a short but very productive first interview at the second stage. At stage #3, I analyzed the eight publications of the original studies that used primary IPD in order to understand the process of the reuse of data better, namely de Vries, van Pampus, Hague, Bezemer, & Joosten, 2012; Gris et al., 2010, 2011; Kaandorp et al., 2010; Martinelli et al., 2012; Rey et al., 2009; Rodger, Hague, et al., 2014; Visser et al., 2011.

There was a short time span between the first interview and the second interview with my participant since the reuse of data as evidence of scientific claims had already happened. Thus, at stage #5, I only asked my participant to validate my findings of the decision-making process as shown in Figure 15 (or annex 9).

Case study #8: Low-molecular-weight heparin and recurrent placenta-mediated pregnancy complications: a meta-analysis of individual patient data from randomised controlled trials ([http://dx.doi.org/10.1016/S0140-6736\(16\)31139-4](http://dx.doi.org/10.1016/S0140-6736(16)31139-4)) (Selection criteria: *proprietary data*)

Data collection instruments in 5 stages and a simplified chronology of interactions with 1 participant

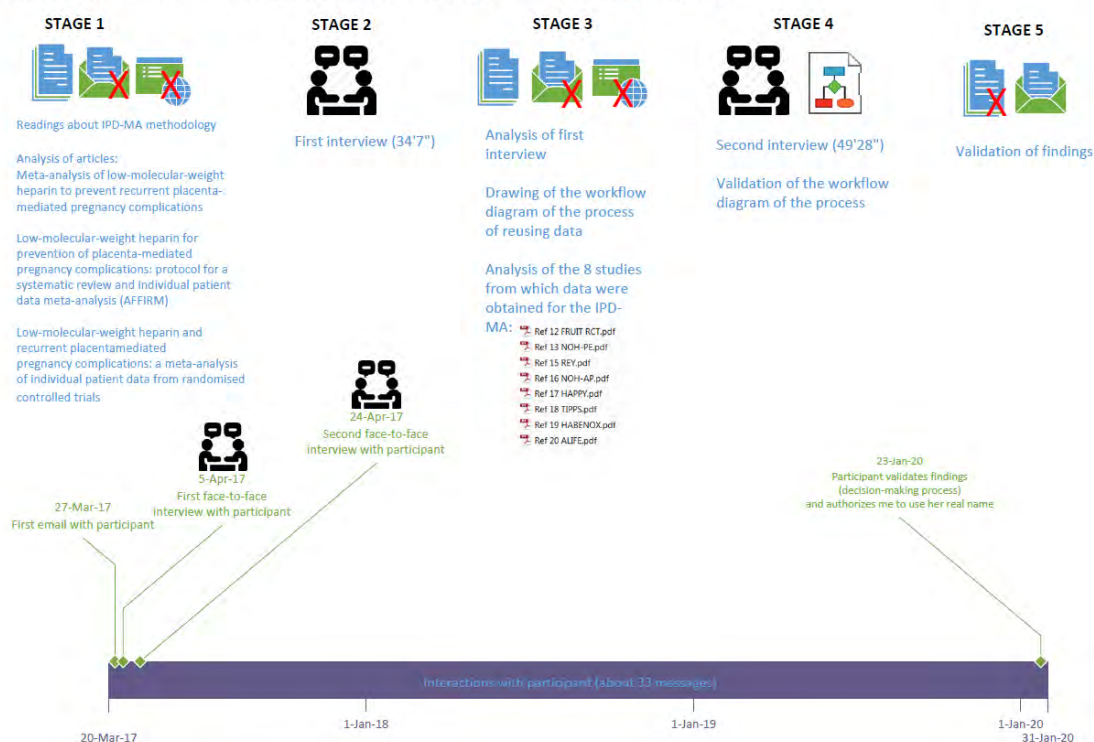


Figure 15 - Data collection instruments and dates. Case study #8

5.3.2. Collected empirical data and collection dates in case study #9 (IPD NMA)

In this case study, there was data collection and data analysis at all stages (Figure 16 or annex 10). At stage #3, I analyzed my participant's study on the same health topic but with aggregated data, and both the Data Transfer Agreement and the Terms of Reference of the IPD study.

At stage #5, I analyzed the publication of the scientific contribution with the IPD (Welch et al., 2019) to identify differences between the expected reuse and the actual one. I also asked my participant to validate findings about the decision-making process.

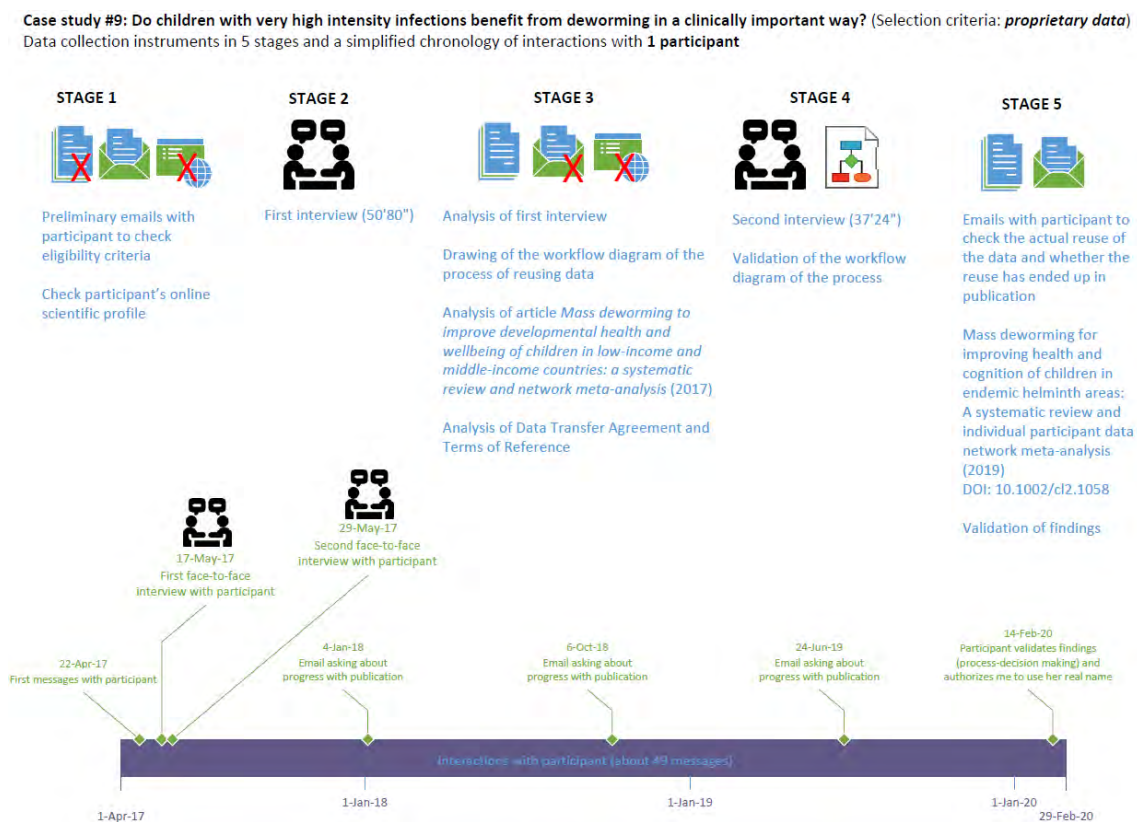


Figure 16 - Data collection instruments and dates. Case study #9

5.3.3. Collected empirical data and collection dates in case study #10 (IPD NMA)

In this case study I made an analysis of publications related to the health issue (Alzheimer’s dementia), i.e., Veroniki, Straus, Ashoor, Hamid, et al., 2016; Veroniki, Straus, Ashoor, Stewart, et al., 2016 at stage #1. Both interviews with my participant were on Skype with a long time span between the two because of the long period needed for accessing the data. The first interview happened while she was in Canada, and the second one happened when she was in Greece.

At stage #5 I could only analyze a publication related to the process of reusing data (i.e., Veroniki et al., 2019) since analysis of the IPD NMA were not finished in February 2020.

See annex 11 for a full size of Figure 17.

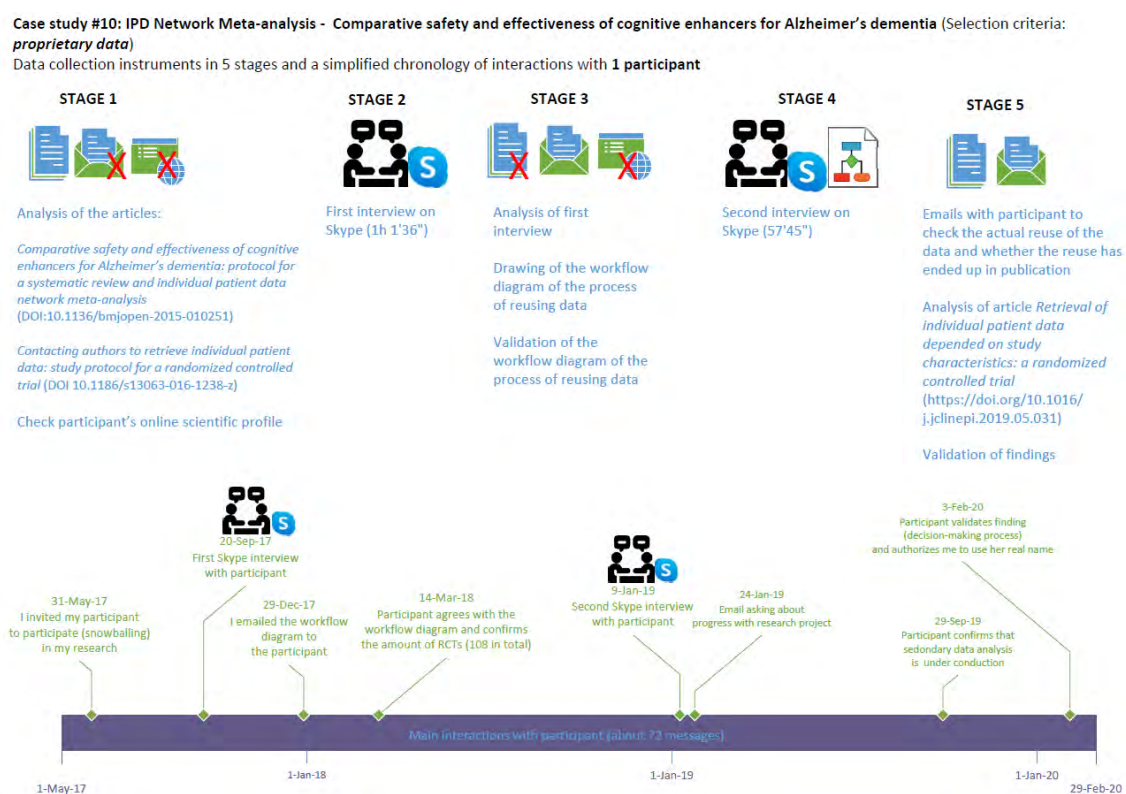


Figure 17 - Data collection instruments and dates. Case study #10

Chapter 6

Empirical analysis: testing the data-reuse mechanism

The case studies in this book are drawn mainly from the published literature, some dating from the era before archives. The old examples remind the reader that no one is barred from secondary analysis for lack of an accessible archive. They also reveal that, although those who proceed without recourse to the archive may be taxed heavily in some ways and have less lavish resources at their disposal, their difficulties of formulation and analysis, ironically, are sometimes reduced. (Hyman, 1972, p. 32-33)

Analysis of the ten case studies are presented grouped together according to the ease of accessibility of the secondary data (*released data, stewarded data, and proprietary data*). I also present a comparative overview of the ten cases in section 6.4. of this chapter.

As explained above, mechanisms have different time horizons. The data-reuse mechanism may have short and long time horizons, and thus, I have collected data at different stages, when possible, for

tracking or following researchers' decision-making until the theorized outcomes –or *events* in Sayer's terminology– have happened. Therefore, I present the results of the analysis in three sections for each case study:

- 1) *Conditions at the outset of the decision-making process* (or time 1)¹²⁸. In this first section, I present findings of objects' initial structure, objects' initial causal powers and liabilities, and the initial values of conditions (C1, C2, C3, C4, and C5) in relationship with the decision-making process¹²⁹ of using secondary data.
- 2) *Conditions at a later time of the decision-making process* (or time 2)¹³⁰. In this second section, I present findings regarding both the final actual events (one of the six potential outcomes -a, b, c or 1, 2, 3-), and later changes occurred along the data-reuse mechanism's time horizon –if any– of the objects' structure, objects' causal powers and liabilities, and conditions (C1, C2, C3, C4, and C5).
- 3) *The decision-making process*. In this last section, I provide a general account of the process, and I relate some of its events and conditions with the theoretical concepts that I have used to theorize the data-reuse mechanism¹³¹.

Throughout the three sections, I highlight some researchers' specific causal powers and liabilities and some secondary data's causal powers and liabilities throughout own researchers' quotes or my own prose. I do this by underlining parts of participants' own words, or by using footnotes. I have also included a reduced size¹³² of workflow diagram (WD) of the data reuse process reviewed by the researcher¹³³. In case studies where I have anonymized participants, I do not mention details of their structure, publication or research topic details since these participants could be identified by means of this information. I have included some of the theoretical concepts in square parentheses in-between my narrative or participants' verbatim words. For instance, [satisficing], [bounded rationality], [procedural rationality], etc.

¹²⁸ For findings presented in this section, I mainly collected data during stages #1, #2 (first interview), #3, and #4 (second interview).

¹²⁹ I will use *decision-making* and *decision(s)* most of times, if not all. However, for some authors, decisions do not always happen. Instead, actions happen with no apparent prior decision.

¹³⁰ For findings presented in this section, I mainly collected data during stages #5 and #6 (for case study #4).

¹³¹ I do this by adding the theoretical concepts between square brackets immediately after the empirical data, which I think is related to the concept.

¹³² Full-sized workflow diagrams are included in this dissertation as annexes.

¹³³ In all case studies, except in case study #6

Information about data sources and data collection dates for each case study is included in Chapter 5. I have tried to use as many participants' quotes¹³⁴ as possible instead of trying to rephrase or reformulate participants' own words.

6.1. Case studies reusing *released data*

Released data or *publicly released data* are data that are available for reuse and are publicly released or published. The four case studies under this category belong to molecular biology. Current data sharing practices by molecular biologists and other surrounding disciplines have evolved from *proprietary data* or *closed data* to *released data* or *Open Data* (Brown, 2003). This has allowed the creation of a new “discipline”, i.e., computational biology, bioinformatics or *dry lab* work, while basic bench biological work is known as *wet lab* (for understanding this evolution, see, for instance, García-Sancho, 2012b, 2012a; Heeney, Hawkins, De Vries, Boddington, & Kaye, 2010; Kaye, Heeney, Hawkins, de Vries, & Boddington, 2009; Kohler, 1994; Strasser, 2010; Strasser & De Chadarevian, 2011).

The difference between the *wet lab* and the *dry lab* is mainly epistemological, and thus methodological. They represent different styles of doing science and acknowledging causality for solving the same biological problems or questions. Some authors have coined the collaboration between researchers with these two different of doing science the “moist” zone (Kahlem & Birney, 2006; Penders, Horstman, & Vos, 2008).

The participants in these four *released-data* case studies feel they “are biologists” (or *wet lab researchers*). However, they also conduct computational analyses with secondary data, and thus, can be also considered computational biologists, bioinformatics scientists or *dry lab researchers*. The Vanderhyden Lab exemplifies the “moist” zone as explained to me by email by one of my participants, David Cook.

David: *I do think that the work we've done in our lab could fit within these boundaries nicely. The only asterisk here is regarding your point (2): as a whole, our entire research group does use data that has been generated by others, however, **because our lab is primarily a "wet lab" group (ie. performing experiments at the bench), most of this data mining goes through me.** What often happens is that one of my lab-mates is interested in whether a certain data repository has information regarding a*

¹³⁴ Participants' quotes are presented indented and in italics, and only in italics when they are in footnotes. For my own words, I have also used italics after “I:”

specific question of these (eg. "Does 'The Cancer Genome Atlas' show that patients with ovarian cancer often have mutations in the BRCA1 gene"), and I go through the repositories, mine out the relevant data, and get back to them with their answer. So we all benefit from these datasets, [...]

6.1.1. Case study #1 (GTEx data repository)

Conditions at the outset of the decision-making process (Time 1)

David Cook is a PhD student at the Vanderhyden Lab of The Ottawa Hospital Research Institute (OHRI), whose dissertation is supervised by Barbara Vanderhyden, PhD, director of the laboratory. The lab is specialized in ovarian cancer research. David pursues a PhD degree, while contributing with his research to the lab's research program. His research belongs to the molecular biology or cell biology discipline, but we could also say that he belongs to computational biology, as he tries to answer research questions related to ovarian cancer from both a *wet lab* (basic bench research) and a *dry lab* (computerized research) perspective. He is the only one in the lab, who has this causal power, at least with analyzing huge amounts of data with R. For his work at the lab, and thus for his PhD, he has been awarded with two funding, the Ontario Graduate Scholarship, and, later, the CIHR Frederick Banting and Charles Best Doctoral Award.

He is subject to the reward system of science for academic researchers, and he knows the rules very well. Regarding David's causal powers and liabilities, he possesses all the ones I have hypothesized in the data-reuse mechanism. There is enough evidence in the data that I have collected during interviews, although I have only highlighted some of them.

(C1) The researcher knows that secondary data exist

David knows that GTEx (Genotype-Tissue Expression)¹³⁵ data exist. In fact, as the validated-by-participant workflow diagram shows (Figure 18; full size in annex 12), David knew about the GTEx data before he envisioned his hypothesis or research question to be answered with them.

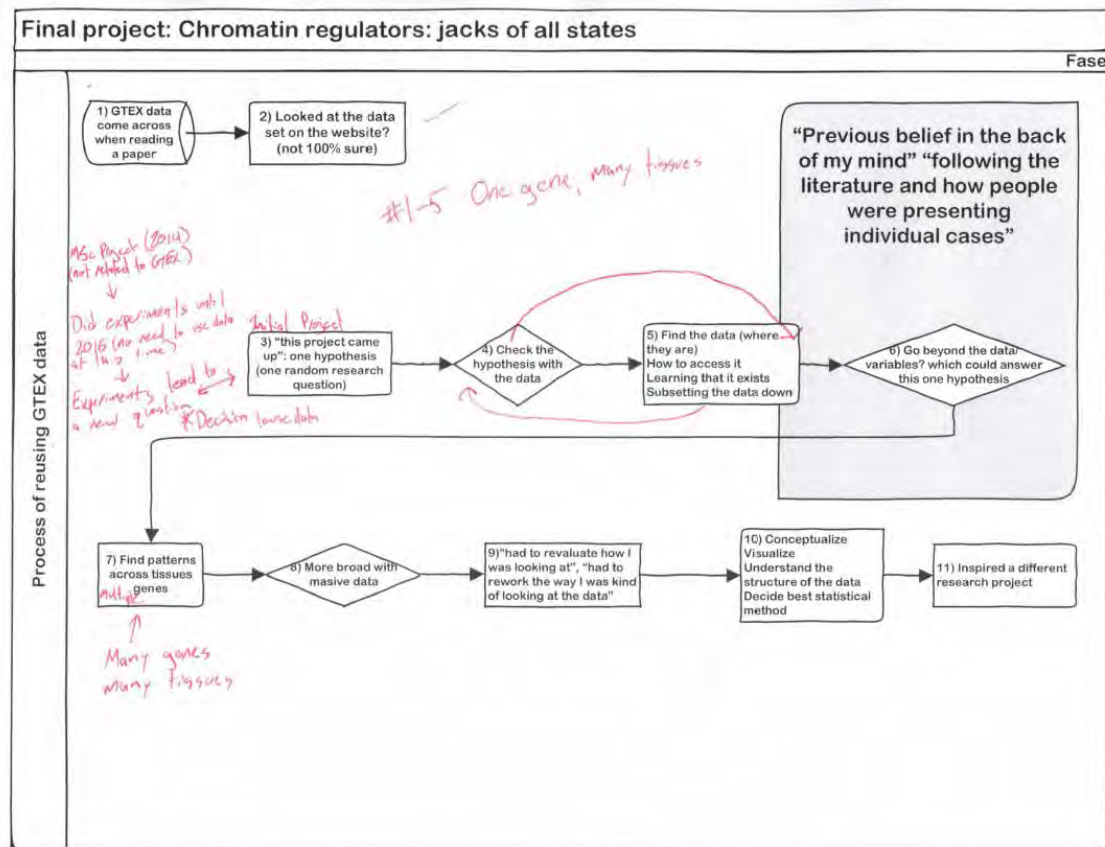


Figure 18 - Workflow diagram of the data reuse process of case study #1

¹³⁵ GTEx data are data released online on the GTEx Portal. *The Genotype-Tissue Expression (GTEx) project is an ongoing effort to build a comprehensive public resource to study tissue-specific gene expression and regulation. Samples were collected from 54 non-diseased tissue sites across nearly 1000 individuals, primarily for molecular assays including WGS, WES, and RNA-Seq. Remaining samples are available from the GTEx Biobank. The GTEx Portal provides open access to data including gene expression, QTLs, and histology images. Source: <https://GTExportal.org/home/>*

When David validated the workflow diagram, he said referring to the first two isolated steps identified with numbers 1) and 2):

David: *Great. So, yeah. I totally agree with kind of how this was laid out because it was this kind of separate thing prior to my actual interest in this project. So, I think one and two were kind of an isolated event, which I like that you put it like that. So, then this kind of data resource became just something in a toolkit in case I needed it. So, that's great.*

So, when I started with that project and when I was first kind of designing the experiments for that project I wasn't thinking about this GTEEx data set at all. It wasn't until I got to -- you know, I did some experiments, and every experiment, while answering certain questions, brings up new questions. So, it was after a bit of just kind of experimentation that we came across the question that required the GTEEx data set.

Yeah. So, I mean, I would say when I saw that -- when I came across GTEEx I did not immediately connect the two. It wasn't until, you know, doing a couple more experiments that I thought maybe I should look into this, and, oh yeah, I remember this GTEEx data set that exists. This would be the perfect tool to look into this.

(C2) Data are obtained

David needed no effort in obtaining the data. GTEEx data are publicly available for reuse, and he has causal powers and liabilities to understand, interpret, and analyze the data.

David: *So, yeah, what it was really great about this is that this data base existed to start with and it was public available and I was able to go and look at.*

[W]hat it was great about this is that, as I said, it was a consortium level project, everything was posted online, wide open to the public, no behind any pay walls or anything like that, you only had to download raw data, you can interact with this data in an online web platform, I mean, very accessible to the public. It's become a very useful tool for these random questions you might have

(C3) Particular secondary data are an initial *satisficing* option

This condition can be two-fold¹³⁶. In this case, this condition is met for C3-SC. On the one hand, *secondary data in general* are *satisficing* for David in so far as his discipline accepts them as evidence of scientific claims, although for only some types of contribution or publications. On the other hand, David finds *some particular secondary data satisficing* in so far as David thinks that he can answer his research question(s) with them, and thus make a scientific contribution.

Although David explained to me that the field of biology is becoming more open to conduct research based on computers and digital data, he opts for a *perspectives article*¹³⁷, and not for a traditional one because in his discipline the use of secondary data is not accepted for all kinds of contributions or scientific claims:

I: *So, in biology, in your field or subfield of molecular biology basic research, is the use of the secondary data a good practice? [...] for publishing a journal or paper, or in top journals. How is it accepted?*

David: *How is it perceived kind of or accepted? Yeah. So, it is very, very accepted and encouraged for people to bring in other people's data. As biology is becoming more computational, we're seeing that in almost every paper. The reason that I was*

¹³⁶ The researcher perceives the secondary data a *satisficing* option in two ways.

- 1) C3-SC) On the one hand, she perceives the option of using some particular secondary data *satisficing* in so far as she thinks she can answer a research question with these data for making a scientific contribution with secondary data, alone or together with primary data. In this case, this perception implies necessarily that she perceives the option of using secondary data in general for making a scientific contribution *satisficing* in so far as she can obtain her expected rewards within the epistemic norms of her discipline.
- 2) C3-BK) On the other hand, she perceives the option of using some particular secondary data *satisficing* in so far as she thinks she can answer a research question with these data only for creating background knowledge, e.g., generate or validate hypotheses with no intention to publish them. In this case, this perception does not necessarily imply that the use of secondary data as evidence of scientific claims is accepted in her discipline.

¹³⁷ Perspectives articles are considered secondary literature since they do not involve original research. Elsevier, for example, for the *Journal of Molecular Biology*, defines a perspectives article as “[...] brief reviews that present a sharply focused view of a rapidly advancing area of research. Authorship is normally by invitation: the Editor-in-Chief or Scientific Editor should be consulted in advance by anyone wishing to submit an unsolicited Perspective.” (Source: <https://www.elsevier.com/journals/journal-of-molecular-biology/0022-2836/guide-for-authors>, consulted December 31, 2019). This is how David explained to me the role of a perspectives article in his discipline:

David: *So, the types of articles that are typically published in basic research journals -- you have the traditional articles. These are multi-figured. I think there's a requirement that you actually have to do like your own experiments in there at least for like the life sciences. So, typically, you won't -- if you're just reanalyzing data, or maybe applying the computational method, or anything, you typically do not publish that as a traditional research article in a journal. There are some exceptions, of course. But most of the time those do not fall into traditional articles because you're not doing experimentation with them. So, the idea of a perspectives article isn't necessarily just to present new findings. It's almost like an opinion article. You're putting an opinion out there --*

saying that I can't really use that data for my PhD or my master's project -- one is that it was just a somewhat different theme or like a different research question than what I specifically laid out for my master's or PhD.

But the reason I said it was a perspectives article rather than a traditional research article... There is a trend in journals that unless you are like writing a new algorithm, or methodology, or something, they typically don't accept submitted articles that are simply reanalysis of preexisting data.

David found that the data in the GTEx portal fitted his research question, and he found no challenges in accessing, and analyzing the data.

David: *uh, and then, that was the main use of this kind of broad public data base of this information. After that, I was "ok, can you go into the literature and see consistent findings and piece together more of the story, uh. So, yeah, what it was really great about this is that this data base existed to start with and it was public available and I was able to go and look at.*

(C4) The idea of collecting particular primary data is not an initial *satisficing* option

David meets condition C4 for this specific research question or project.

David: *We are drafting a manuscript right now...for a... it's a... we are trying to sell it more as a perspective article rather than a basic research. Because for this specific project we are not planning on generating our own data. It's kind of presenting this perspective based on data that already exists.*

(C5) An expected scientific contribution exists and the researcher finds its potential rewards *satisficing*

The research project's tentative name is *Chromatin regulators: jacks of all states*. David's goal is to make a scientific contribution with a perspectives article with secondary data as the only evidence of his scientific claims, that is, the hypothesized outcome 3.

David: *we are drafting a manuscript right now... for a... it's a... we are trying to sell it more as a perspective article rather than a basic research [...]*

so, this one specifically uhhh I was the only one doing actual work on it because it was just primarily computational analysis - digging out things- uhh, like I said, presenting this like a brief perspective- it's not a large scale highly collaborative project - this is more a... I do not want to call it a side project because I do think it's important, but it was a smaller endeavor than some of the other projects. So, I have had a great support system with my lab to bounce ideas of it and discuss what they think about it, but I would say that I am the lead, kind of driver of this project, and the one facilitating all of it. Uhh, when it comes time to actually present this, to submit it, and prepare the final of the manuscript, me and Doctor Barbara Vanderhyden will be working together on the actual manuscript preparation of it. Uhh, but the work itself will be done by me. So, in the end, this will be a two-author publication myself and Barbara.

However, for David, this research project or expected scientific contribution is important, but he did not envision it as part as his PhD dissertation. When I asked David if this scientific contribution was part of his PhD thesis, he answered:

David: *That's a good question. So, it is not part of what is currently my formal thesis, uh, research project. Now, whether it integrates itself into my final thesis itself is different, because, at least in biology, like you never know how's going to shape up. So, in the end, it's kind of "what do I put together?" is my thesis, but it is not my main thesis project.*

And whether the research was subject or contingent upon funding conditions, he said:

David: *luckily, it's a free project*

Conditions at a later time of the decision-making process (Time 2)

Event or outcome: When I interviewed David at the beginning of 2017 he had already reused – downloaded and analyzed– the data from the GTEx portal for a scientific contribution in the form of a perspectives article. At that time, he was already drafting the article. When I contacted him about one year and half later by email, the event was not outcome 3 as David had originally planned, but was outcome 2.

Changes in conditions:

A change in condition C5 had occurred during the time span between t1 and t2. The potential rewards of the perspectives article stopped being *satisficing* for David.

David: *As for the GTEx Project, I kind of gave up on writing it. I thought the work I was doing on the project was informative, but I ended up prioritizing some other projects that I figured may be more beneficial to me.*

The decision-making process

The decision-making process of this case study follows this pattern:

(expected) outcome 3 → (final) outcome 2

While David is doing some experiments in the laboratory, he comes up with a hypothesis and he decides to use a data resource that he discovers a while ago, although, at that time of its discovery, he did not explore in detail. However, he knows for sure that the data are accessible for reuse since released data –or Open Data– are very common in molecular biology.

He approaches the data with a hypothesis in mind, and starts delving into the GTEx repository. His initial idea is to look at a gene in many tissues (see David's handwritten annotation regarding steps 1 through 5 in the workflow diagram “#1-5 One gene, many tissues”). After digging into the GTEx portal, the data inspires him new hypotheses since David observes patterns across many tissues in multiple genes. In this process, it is not clear when David starts envisioning a potential scientific contribution with the new hypotheses and the GTEx data. However, neither the initial hypothesis, which motivates him to visit the GTEx data nor the later hypotheses inspired by the data are related to his doctoral thesis¹³⁸, although they fall into the lab's research program.

David goes on with the idea of making a scientific contribution in the form of a perspectives article until he decides prioritizing other lines of research within the lab, which are possibly more related to what he begins to conceive as a doctoral thesis, which is his priority goal among all his research activities. In other words, the potential benefits for him of publishing a perspective article are not as good [satisficing] as the ones if he focused on other projects within the lab's research program.

¹³⁸ *No. But I think that it was kind of a tangent off of my research. So, I was going along my research trajectory of my master's project. I had very specific aims. This was not part of the aims or objectives at all. It was just kind of a curiosity-driven and inspired question and I just decided to follow up on it because I knew when I was thinking across it that I had a data set that I could access and use.*

Condition C1 happens first in time. C4 maybe happens at the same time as C1, but not necessarily as something reasoned. Conditions C2 and C3-SC happen nearly simultaneously, so David can make informed decisions at once. There is hardly any uncertainty - or none - about data accessibility and about the intellectual and computational effort to analyze them¹³⁹. David has all the information at once. The time horizon of the data-reuse mechanism is short. So the initial expected outcome # 3 arises straightforward [procedural rationality]. During the course of time, condition C5 disappears, while the rest of the conditions continue existing and with the same values. Thus, the initial expected outcome # 3 ends up being an outcome # 2. The only thing that changed is that David stopped seeing the rewards of the perspectives article's contribution as something *satisficing* for him, and he started envisioning other projects' potential contribution as more *satisficing* than the former on the going [procedural rationality]. However, at that time, he does not know exactly what benefits he will obtain from this decision [bounded rationality].

Giving up this potential scientific contribution is an easy decision –or action–, since the efforts and time for accessing and reusing the data are minimal. Once the goal of making a scientific contribution has disappeared, the decision-making process disappears [teleological decision-making theory].

No changes in David's causal powers and liabilities and in the GTEx data's causal powers and liabilities happen during the decision-making process. David's structure remains the same along the process.

¹³⁹*It was the largest of this kind, so it was perfect for... so, essentially what this dataset is a normal experiment we would do in the lab done a couple of thousands times, right? So, we have the power technically to generate these data, but this was done in such a massive scale and it was all, you know done within a standard pipeline and put together, and it is simply the largest one of it. There are other datasets that are similar, but, you know, like I said, this is gene expression in tissues so maybe there is fewer tissues so to work with. So, I simply did this because I wanted to capture as much information about this as I could across as many tissues.*

[...]

So, with regard to mixing and integrating multiple data sets, for example. Luckily this data set for my interest was pretty self contained. Had all the information I needed. So, I didn't have to, you know, sometimes in the field we want to get data set A, and data set B, but they are processed run independently so you have to do statistical things to make sure that at least they are comparable. Uh, luckily with this data set... I should also add that you can, what I ended up downloading from most of this was simply already processed data, so I didn't have, so, you know, they, the developers had done all the, you know, normalization and all that with all the data, so that I didn't have to worry about too many of the issues. It kind of came in a good state for me to just dive in, there are... Sometimes you have to do like mathematical or statistical things just for better stats or visualizations or whatever, maybe some transformations, you know stuff like that.

6.1.2. Case study #2 (GEO Profiles repository)

Conditions at the outset of the decision-making process (Time 1)

Please, refer to David's structure and causal powers and liabilities in case study #1 since they are the same, except for the fact that for the lab experiments of this research project the lab received a NSERC grant (#RGPIN 2018-0653.8).

(C1) The researcher knows that secondary data exist

David knows that GEO Profiles (Gene Expression Omnibus)¹⁴⁰ data exist. This data source has been used by his lab for many years¹⁴¹, and it is one of the most well-known public functional genomics data repositories. David knows GEO since his undergraduate studies. See his own quotes in the validated workflow diagram of the data reuse process that he described regarding the familiarity of his lab with GEO Profiles data (Figure 19 or annex 13). They usually know about specific microarray data sets deposited in GEO by the publications of the MEDLINE database that he finds mainly through search engines, e.g., PubMed¹⁴².

¹⁴⁰ The GEO Profiles database stores gene expression profiles derived from curated GEO DataSets. Each Profile is presented as a chart that displays the expression level of one gene across all Samples within a DataSet. Experimental context is provided in the bars along the bottom of the charts making it possible to see at a glance whether a gene is differentially expressed across different experimental conditions. Profiles have various types of links including internal links that connect genes that exhibit similar behavior, and external links to relevant records in other NCBI databases. Source: <https://www.ncbi.nlm.nih.gov/geoprofiles> [January 1, 2020]

¹⁴¹ David: *We use that on a regular basis uh throughout pretty much every project. Yes I'm using it right now for my PhD project, I haven't in a tiny bit right now, but it was really used a lot (he emphasizes "a lot") in the months leading up to my PhD project to help spark that research question that is part of this proposal. Does that make sense?*

¹⁴² PubMed is a search engine, which mainly access the MEDLINE database. MEDLINE database is a collection of references and abstracts on biomedical and life sciences research topics.

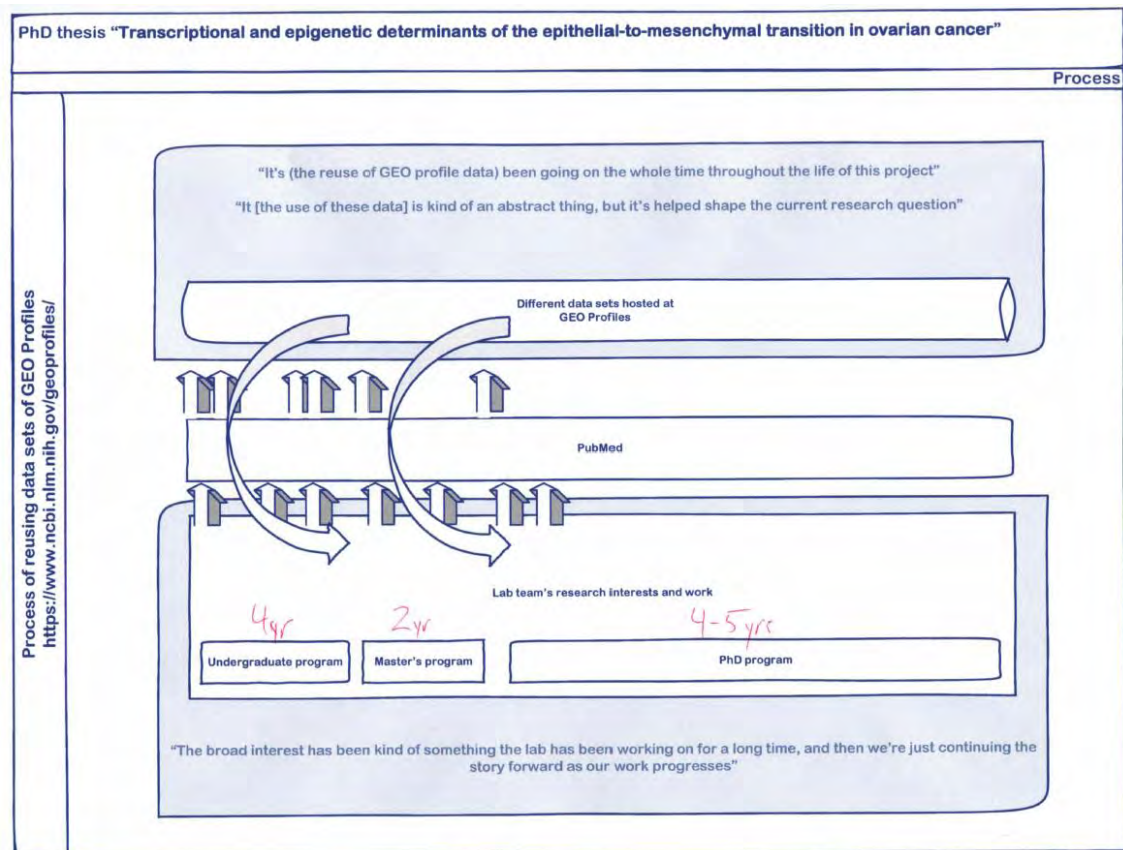


Figure 19 - Workflow diagram of the data reuse process of case study #3

(C2) Data are obtained

Data sets of microarrays in GEO Profiles are publicly available for reuse. There are no restrictions or access walls for using the data. The information about the full availability of the data sets at GEO is clearly stated in GEO Profiles online portal: *Anybody can access and download public GEO data. There are no login requirements. For more information, please read these copyright and data disclaimers*¹⁴³.

(C3) Particular secondary data are an initial *satisficing* option

Some data sets on the online GEO repository were an initial *satisficing* option for his research project, tentatively titled *Transcriptional and epigenetic determinants of the epithelial-to-mesenchymal transition in ovarian cancer*, which is part of his PhD project. David thinks that GEO Profiles portal "is a very rich resource for data sets", and he and the whole Vanderhyden lab have been using data of GEO Profiles continuously for a long time mainly for the creation of background knowledge.

¹⁴³ Source: <https://www.ncbi.nlm.nih.gov/geo/info/faq.html#restrictions> [January 2, 2020]

David: *It was a very key, the usage of data was very important to getting where we are and therefore where the proposal for the project is, but at this time it's not a key part of what's going to be happening in the project, if you know what I mean. So, the usage of data was important to get us all the answers we needed to justify making this proposal.*

At the conception of David's PhD research design, he had a draft idea¹⁴⁴ of the kind of contribution he wanted to make on ovarian cancer, and initially he only knew for sure that he would produce his own primary data at the lab. He perceived GEO Profiles data merely as an information source for helping with his research proposal and other research work at the lab. Secondary data played a second role, and were not going to be used as evidence of scientific claims (C3-BK)¹⁴⁵.

David: *So, then through collecting various data sets that exist, that we come across, we can gather a lot of information from that and then paint a pretty good picture to help make sure we're going down the right path in the lab. So, that was a very, very common occurrence during the early phases of this project where it was just, let's see if this data set exists and whether protein A does job X, something like that. So, we started doing that very, very frequently, and through that it at least gave us things to look at in the lab, and helped us start interpreting our data a little bit better.*

So, it wasn't even really, I would even argue that that data is not really going to be part of the project, or at least the way we present it. So, for example we wouldn't publish their data with our data, but it really helped steer us in the appropriate direction so that we were going through the project well. Umm, that happened more and more, and then it actually helped push the project along quite well.

¹⁴⁴ Actually, this is how a research contribution or project evolves (Abbott, 2004)

¹⁴⁵ David: *It was a very key, the usage of data was very important to getting where we are and therefore where the proposal for the project is, but at this time it's not a key part of what's going to be happening in the project, if you know what I mean. So, the usage of data was important to get us all the answers we needed to justify making this proposal.*

(C4) The idea of collecting particular primary data is not an initial *satisficing* option

For this scientific contribution, David is going to collect primary data. So, this condition is not met. This contribution (*Transcriptional and epigenetic determinants ...*) is part of his PhD dissertation (compendium of articles)¹⁴⁶, which stems from a *web lab* where basic bench research is conducted as expected by the molecular biologists community.

I: *For your PhD, then, you will collect your own data in your lab?*

David: *Yes, absolutely.*

In fact, in his summarized research proposal, he only referred to primary data that the lab would generate:

David: *All data and analyses generated in this project will be made publicly available and will serve as a minable resource for the research community.*

(C5) An expected scientific contribution exists and the researcher finds its potential rewards *satisficing*

David needs to publish articles in order to achieve his current career milestone, namely his PhD degree. The research idea¹⁴⁷ for his PhD dissertation has evolved for several years¹⁴⁸, but it has always fallen under the umbrella of the topic *Transcriptional and epigenetic determinants of the epithelial-to-mesenchymal transition in ovarian cancer*.

¹⁴⁶ David: *That's the idea. In order to do that you have to have two first author publications and a third publication, that's similar. As long as you have those publications, it's so much easier doing it by article because then you just stitch together your papers and there's your thesis. Rather than having to do by chapters where you're synthesizing everything into one smooth document, articles is easier because you just literally dump them into your thesis.*

¹⁴⁷ *The objective of this study is to use loss-of-function approaches and high-throughput genomics to construct a detailed model of the transcriptional and epigenetic determinants of the EMT in ovarian cancer. The work will address several specific aims, including the identification of novel regulators, defining intermediate states, exploring the reversibility of the mesenchymal state, and determining the molecular relationship between distinct EMT-associated characteristics.* Source: David's written summary of PhD research proposal.

¹⁴⁸ David: *So, I mean, it's a continuation of a project that was launched prior to my PhD right? It's actually I worked loosely on part of this when I was in my undergrad actually, and then, because I was working with another student her and it was her PhD project. So, I helped out with hers, and then after I finished my masters I decided to take part of her project and kind of branch off, and do my own direction there.*

Conditions at a later time of the decision-making process (Time 2)

Event or outcome: The use of secondary data happened as evidence of scientific claims together with primary data also being used as the main evidence of scientific claims. This outcome does not match with any of the outcomes that I have theorized for the data-reuse mechanism.

Changes in conditions: In this case, the condition C3 that can be met in two ways (secondary data is satisficing for making a scientific contribution (C3-SC), and secondary data are satisficing for the creation of background knowledge (C3-BK)), starts having the value C3-BK, and later has value C3-SC.

The decision-making process

This case study falls to combination A, and not to combination B (see Table 1). The difference between combination A and B is that the condition “the idea of collecting primary data is a *satisficing* option” is met in the former, but not in the latter. For the initial combination A, I have hypothesized three potential outcomes (a, b, and c). Therefore, the decision-making process of this case study follows this pattern with regard to events:

(expected) outcome b → (final) outcome c

The reason for using primary data for his PhD and, thus, contribute to his field with novel knowledge in ovarian cancer is because David has the causal power to know the epistemic practices of his discipline, molecular biology, and thus knows his scientific community accepts as a contribution. So, when this project started as a research idea, the initial expected use of secondary data from GEO Profiles was outcome b (data were not aimed to be used as evidence of scientific claims, but for creating background knowledge). However, as his work progressed, David saw the opportunity to improve his contribution with secondary data, which made his results more generalizable as he explained to me:

The majority of the data is primary data we generated in the lab, but secondary data was used to further support our findings and assess how generalizable our findings were (ie. if the patterns we observed could also be seen in contexts studied by other groups). Specifically, we used data generated by three separate studies and included the analysis in Extended Data Figure 4b-c. [...] The data was, however, accessed via GEO.

His quote is related to his recent publication: Cook, D. P., & Vanderhyden, B. C. (2019). Comparing transcriptional dynamics of the epithelial-mesenchymal transition. BioRxiv. <https://doi.org/10.1101/732412>

6.1.3. Case study #3 (TCGA data repository)¹⁴⁹

I also asked David to tell me about an occasion when he initially decided to reuse secondary data, but in which he later decided not to¹⁵⁰. This case study refers to a research project or question regarding the reuse of secondary data from the Cancer Genome Atlas (TCGA) repository, which contains publicly released data of about 33 cancer types. The repository contains over 2.5 petabytes of genomic, epigenomic, transcriptomic, and proteomic data from more than 11,000 patients collected by a host of researchers¹⁵¹. This case study has no specific name, although it is related to the research program of the lab on ovarian cancer. I drew no workflow diagram for this case study.

Regarding David's structure and causal powers and liabilities in this case study #3, they are the same than in case studies #1 and #2. David knew that the TCGA repository existed (C1 was met), and that the data were accessible, though only partially (C2 was partially met) [bounded rationality – some information is needed to make decisions]. The reason of meeting C2 partially is that only some of the data were released, but other data –referred to clinical variables– were *stewarded data*, thus, there were some permission walls to access the clinical variables.

David wanted to make a scientific contribution (C5 was met), but not with the TCGA secondary data. On the contrary, the idea of collecting particular primary data was an initial *satisficing* option, since the lab was already producing or wanted to produce primary data in their experiments for the scientific contribution (C4 was not met).

So, initially data from the TCGA project were going to be used as background knowledge (outcome b) for helping the lab with their experiments (C3-BK was met). TCGA data had an initial thematic or conceptual fitness of the data with David's research question(s). However, as he started to get more information about the state in which the data were released, he started to perceive the TCGA data as *non-satisficing* because of two reasons. On the one hand, because they had to make requests and go through several permission walls in order to access the data. On the other hand, because the data were

¹⁴⁹ This is the only case study, where I do not use the three sections for providing the findings.

¹⁵⁰ The reason, as explained in the methods and methodology chapter, is that I wanted to introduce variability in the values of the event to find out what changes in the causal forces of the data-reuse mechanism are necessary to have each of the values of the event.

¹⁵¹ Source: <https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga> [January 1, 2020]

unprocessed data (or *raw data*). Thus, although initially there is thematic or conceptual fitness, David decides that there is no level-of-analysis fitness. David and the rest of the lab thought it was not worth the time required and the extra effort to process the data, compared to the information they thought they could obtain as background knowledge for their own experiment [waning uncertainty; procedural rationality]

So, the expected outcome b (use data as background knowledge) was easy to give up, and thus, outcome a happened (no reuse of secondary data)

However, after some time, he decided to reuse the data from the TCGA repository because some of the conditions changed. The causal powers and liabilities of the TCGA data had changed. They were fully released in a processed state.

David: *There was one example. It wasn't like a core question of the project, it was just that we knew this data set existed or at least that the data existed, and we wanted to look into it to see if you know we could at least mine out some information that we wanted to help with our actual experimental project. I started looking into it. I think I gave you the database part, it's that TCGA database, so it's a bunch of cancer related data. It has some molecular data alongside with clinical information so you look at, you know, can compare the two.*

We wanted to look into that, and at the time I believe the only resource that was available, it was hosted on a government kind of funded database somewhere. There were a bunch of different security levels on the data because it dealt with patient information. So, you know, as researchers we could get access to certain parts of it but we would have to go through an application process just in very brief to get an okay to actually access the data.

So, that was the first barrier we hit where it was like, we just wanted to do this kind of as a quick, easy look into the data experiment. We didn't want to invest a lot of time into this. So, that was the first sign that this was going to be a much greater challenge than we wanted to necessarily pursue. So, I went through that initial step where I applied, I got access to, it was the level three data, so the lowest security data. But, then only to find that it was completely unprocessed data that we would have to do, and we're talking a very, very large collection that like I said wouldn't be easy to do on a single computer. We would have to probably spend a lot of time finding out how to do this, the computation in an efficient way.

So, once we found out that it was only the raw data available, at that point we jumped ship and were like, okay, maybe this isn't going to be the best opportunity for us. It was nice that the project didn't require us to have this information, it was more this

extra information to add to the story. So, that made it easier for us to abandon it. Uh, what was nice though is that after a tiny bit of time, another group or organization, probably just even a lab, took the time to do all that processing, consolidated it all into one easy to manage processed data database that we can now access online¹⁵². So, we just kind of waited it out, and then the processed data, the easy to access data became available.

[...]

Which was nice, so then we revisited it, explored what we wanted to explore.

The decision-making process

The decision-making process of this case study follows this pattern with regard to events:

(expected) outcome b → (final at that time) outcome a → (final) outcome b

The process actually has two decision-making processes with regard to the reuse of the data. In the first one, David decides not to use the TCGA data after obtaining more information about the data's causal powers and liabilities. David finds the latter *non-satisficing*. Condition C1, and maybe conditions C5 and C4 happen first in time, although condition C4 was not met. Conditions C2 and C3-BK are met, but only partially since the information that David has about the data is bounded [bounded rationality]. Conditions C2 and C3-BK become fully met gradually over time, and thus, as new information about the status and accessibility of the TCGA data is obtained [procedural rationality], David changes the value of condition C3-BK, which ends up not being met. Therefore, as TCGA data are not a *satisficing* option, data reuse does not happen.

In the second decision-making process, David decides to use the TCGA data because he becomes aware that the data's causal powers and liabilities have changed [procedural rationality]. The data, including the clinical variables, have been released in a processed way, and thus David perceives that there is now a level-of-analysis fitness with his research question. Condition C3-BK becomes met, and data reuse does happen. Had the data not changed their causal powers and liabilities at a later stage, and had David not known about it, the outcome would have been *a* (no data reuse).

No changes in David's causal powers and liabilities and in the TCGA data's causal powers and liabilities happen during the decision-making process. David's structure remains the same along the process.

¹⁵² The TCGA data's causal powers and liabilities changed after some time.

6.1.4. Case study #4 (GEO Profiles and TCGA repositories)

Conditions at the outset of the decision-making process (Time 1)

In this case study, there are two participants, namely, the principal investigator (PI), Joan Climent, PhD, of the research project based on findings from his own PhD dissertation, and a computational biologist, Jaume Forés (PhD candidate at the time of the data collection), who carries out the actual reuse of the secondary data.

They both have internal relations with their *structures*. They belong to a research institution, to a discipline, and they are subject to a reward system of science. At the time of conducting the interviews, Jaume belonged to the Biomedical Research Networking Center of Mental Health (CIBERSAM)¹⁵³, and both the Spanish and Valencian Governments funded¹⁵⁴ his research with secondary data. Joan Climent belonged to both INCLIVA¹⁵⁵ Health Research Institute located in Valencia (Spain), and to the *Departamento de Ciencias Biomédicas, Facultad de Ciencias de la Salud, Universidad Cardenal Herrera-CEU*¹⁵⁶. They have a common research activity under the direction of Joan Climent, although not related to Jaume Forés' PhD dissertation, which is supervised by other researchers. In their research relationship, they complement each other. Joan mainly provides the guidelines and the research questions to be answered. Jaume does the computational analysis with the data. They both analyze and interpret the computed results, with Joan's guidance.

Both participants have the necessary causal powers and liabilities that I have hypothesized in the data-reuse mechanism. There is plenty evidence of these powers and liabilities in the empirical data. However, as in previous case studies, I will highlight only some of them. Joan and Jaume have both a strong background in biology (Bs in Biology), but they also have knowledge and skills to be considered computational biologists. In fact, Jaume Forés has a master's degree in Bioinformatics, and manages R.

¹⁵³ <https://www.cibersam.es/en>

¹⁵⁴ Grant number PROMETEOII/2015/021 from Generalitat Valenciana and the national grant PI17/00719 from ISCIII-FEDER.

¹⁵⁵ <https://www.incliva.es/>

¹⁵⁶ <https://www.uchceu.es/departamento/ciencias-biomedicas>

(C1) The researcher knows that secondary data exist

Jaume and Joan know that GEO Profiles (Gene Expression Omnibus) and TCGA (The Cancer Genome Atlas) data exist prior to the conception of this research, which has its origins in Joan’s research with primary data at a bench. See step 1) in the workflow diagram (Figure 20 or annex 14). Jaume validated it with red color. Joan validated it in blue color. Condition C1 was met.

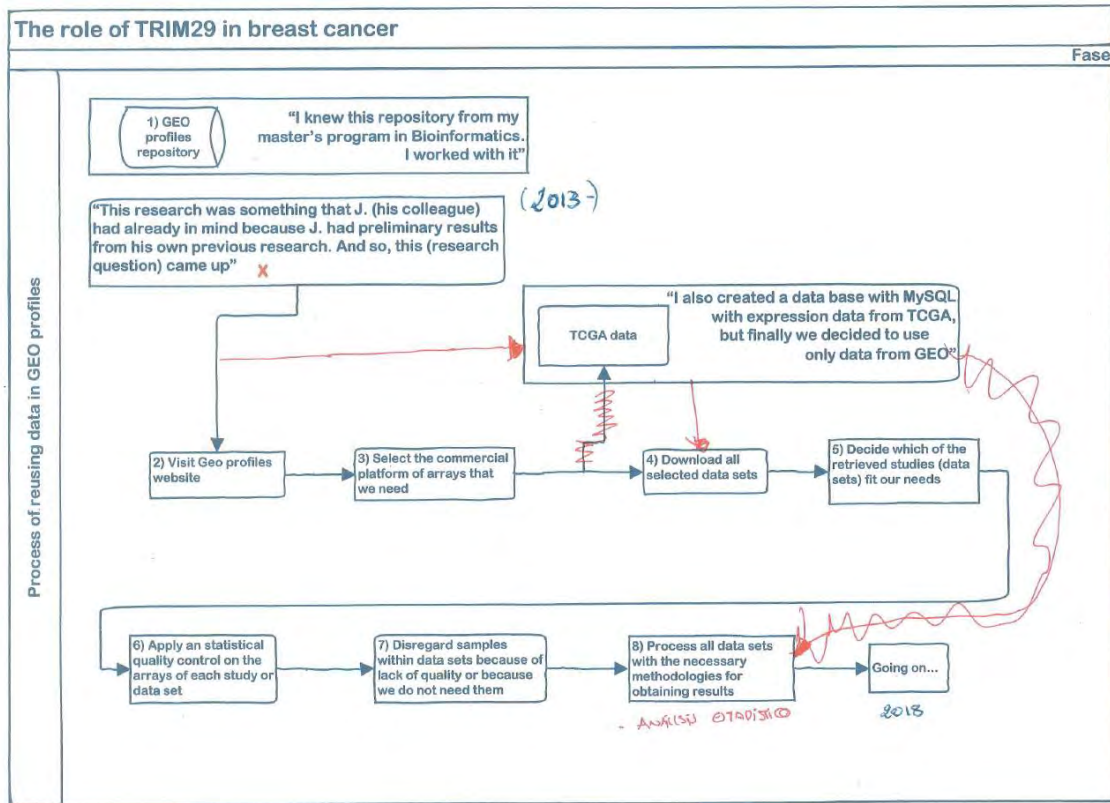


Figure 20 - Workflow diagram of the data reuse process of case study #4

(C2) Data are obtained

GEO and TCGA data are *released data*. Jaume had no issues in accessing and downloading data from GEO Profiles. However, Jaume reported the same issues that David had when accessing TCGA data. Condition C2 was met.

Jaume: *Bueno perquè quan vam escomençar l'accés a fer açò, l'accés a les dades TCGA era més complicat, ara ja...*

Perquè era més... jo crec que hi havia que descarregar les dades a través de l'app, no ho sé, era tot molt més complicat, havies de parcejar els fitxers... ara s'han creat ferramentes, sobretot en paquets de R, que agilitzen molt l'accés a les dades de TCGA.

(C3) Particular secondary data are an initial *satisficing* option

Jaume and Joan needed unprocessed (or raw) data from GEO Profiles and from TCGA. As explained for the previous case studies with David, the reuse of secondary data is accepted in molecular biology, but certain scientific claims and modes of causalities can be only answered with primary data. Joan and Jaume know perfectly their discipline's epistemic practices, and were acting consequently with regard to the scientific claims they wanted to make with secondary data from GEO Profiles and TCGA. Thus, condition C3 was met.

(C4) The idea of collecting particular primary data is not an initial *satisficing* option

This condition has not a straightforward answer. Before knowing whether condition C4 is met or not, we need a previous explanation, which has the underlying philosophical discussion about causality.

The main reason is that Jaume and Joan conduct research in the “moist” zone, that is, the intersection of the wet lab and the dry lab (Penders et al., 2008). Jaume does the computational analysis. Joan and other PhD student do the cell analysis at the web lab on a biological knowledge gap that Joan started to address ten years ago. In order to understand this “moist” zone, or how the dry lab and web lab complemented each other, I drew the diagram of Figure 21 (a full size in annex 15) of his *research project*, which Joan validated. He confirmed me that the three steps are the ideal way for answering biological research questions. However, it does not actually happen like this all the time. Step 2 is where the reuse of data of this case study fits in.

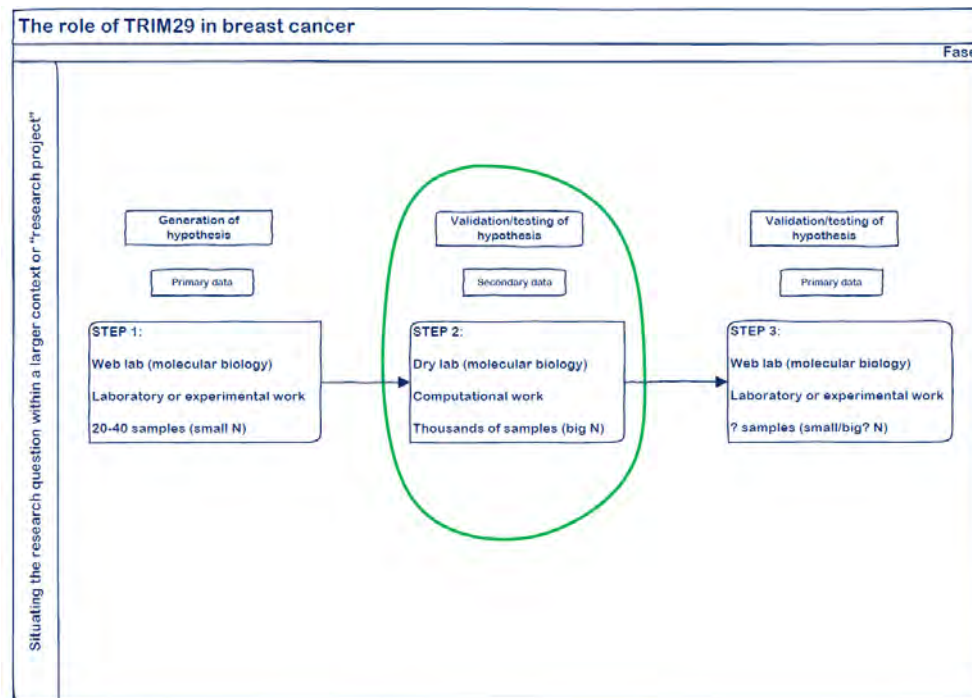


Figure 21 - Situating the reuse of data within a larger research inquiry. Case study #4

Joan's research project¹⁵⁷ (*The role of TRIM29 in breast cancer*) started more than ten years ago. At step 1, Joan created primary data from experiments with a small-N sample for his own PhD research. From these experiments, he was able to generate some hypotheses, which later on and due to the availability of released data and to great developments in analytical tools, he can validate or test with only a computer.

Joan: *la tecnologia ha avançat d'una manera exponencial i brutal de manera que hui en dia hi ha mil tipus de ferramentes que generen dades d'estudi de tot el genoma i cada vegà més punteres, hasta el punt de que hi ha la seqüenciació de tot el genoma RNAC, que és la seqüenciació a nivell del RNA on tens expressió, mutacions. eh... isoformes diferents de cada agent, tens un, en un mateix experiment, per dir-ho aixina, en una mateixa anàlisis, la quantitat d'informació que pots traure és... brutal.*
[04:05]

This testing or validation is mainly the computation work that Jaume carries out (step 2), which is something parallel to the experiments at the web lab by other predoctoral student.

¹⁵⁷ The term "research project" in molecular biology has not the same meaning than in other disciplines. A "research project" is more like a long-life research program in a laboratory, which functions as a big technological and intellectual umbrella for many small research questions aimed to address the same knowledge gap, or solve the same health issue. David Cook, Joan Climent and Jaume Forés coincided when providing me this explanation of a research project, rather a research program of a lab. (Joan: *La part de l'ús de les dades és una part del projecte, el meu projecte va molt més allà, el meu projecte va més en demostrar coses en mostres de ratolins, en mostres de pacients i en línies cel·lulars. Les mostres que jo faig servir públiques és una part, un terç, una quarta part dels projectes que jo porte.*)

Joan: *Lo que trate de dir és, que la hipòtesis que ell vol validar o la part computacional que ell va a fer, fent servir dades públiques, és a partir del que mosatros li hem dit, ell no sap el que hi ha darrere, per això estem dividits entre el que és el biòleg computacional, que és ell, el que és el predoctoral, que està fent la part eh... de laboratori i jo, que vaig sol·licitar eixe projecte i vaig tindre financiació en base a dades que jo mateixa havia generat prèviament en el meu postdoc i durant la meua tesis, vull dir que és una trajectòria molt llarga hasta arribar a que Jaume diga: “-Vale, anem a reunir totes estes mostres i anem a validar-ho tot açò...”*

However, and as a word of caution, the *testing* or *validation* with a bigger N sample does not necessarily translate into direct sound scientific claims. The process of generating hypotheses in step 1 and of validating hypotheses in step 2 is much more complex than the one I have depicted and narrated above, and is also contingent upon the concrete knowledge gap to be addressed. The underlying reason is the confronted view of causation between a probabilistic causal approach (computational biology) and a functional or deterministic causal approach (molecular biology). So, in the end, and to make direct sound scientific claims, researchers need to go back to the wet lab to test or validate what a big N sample has *suggested*, rather than validated.

(regarding step 1 at the wet lab)

Joan: *Però clar, el que tu una cosa no la valides, no significa que no siga certa, sobretot a nivell funcional, perquè poden haver moltes influències tant externes com de manipulació, entones ¿què passa? Però si ho valides, si tu en número elevat de mostres valides la teua hipòtesis, té molta força, entones el que fem nosaltres, com el laboratori és molt menut i en pocs recursos, és fer servir dades públiques per validar idees que tenim tretes o bé de coses que hem estat elaborant, inclús llegint, no fa falta hagen generat la dada en el nostre laboratori, a lo millor a partir d'una dada que han generat al nostre laboratori hem tingut una hipòtesis o hem tingut una idea, l'hem confirmat a base de consultar la literatura, de que fulano fa algo semblant o fulano tamé... però açò no ho ha dit ningú, ¿què passaria si férem esta combinació? I entones anem a les dades públiques a comprovar si la nostra hipòtesis és certa, però eixa hipòtesis pot vindre de dades prèvies o pot vindre simplement d'una idea a base de treballar continuament, projecte, projecte, anem arriscant, entones fem servir les dades públiques per a validar idees.*

(regarding step 2 at the dry lab)

Joan: *Nosaltres el que hem vist és: hi ha un gen x o hi ha una sèrie de gens que determinen un subgrup dins de lo que és la malaltia però la n que tenim és baixeta. Com ja sabem els gens i ja sabem que identifiquen diferents subgrups, anem a agarrar els milers de mostres que puguem recopilar de diferents estudis, de diferents grups i les agrupem en un únic grup de mostres, pa vore si el que nosaltres trobem en unes poques mostres es valida en milers de mostres i si es valida, pues ja tinguem la validació del que nosaltres hem trobat ... no necessitem fer eixes mostres.*

(regarding step 3 at the wet lab)

Joan: *Després clar, sí usem un wet lab, perquè lo que fem és si eixe gen x nosaltres diguem que quan està present identifica mostres que són, tenen una major proliferació o una menor proliferació, i això ho validem en milers de mostres, lo que volem saber és si eixa funció és real o és ocasionà perquè eixe gen va lligat a altres gens i qui realment fa la funció no és eixe gen sinó els altres gens en els quals es relaciona. Lo que fem en el laboratori és, modifiquem eixe gen i allà on s'expressa mos el carreguem i fem que no s'expresse o allà on no s'expressa, l'introduïm i que s'expresse i vegem si, si l'efecte...*

[...]

vegent que l'efecte funcional d'eixe gen és real, el que vegem en les mostres de pacients de que es relaciona amb proliferació, lo que fem és estudis en línies cel·lulars en laboratori on modifiquem el gen i vegem si la proliferació es modifica. Entonces... el que ens indica l'estudi de mostres ho comprovem funcionalment en el laboratori.

Therefore, overall, condition C4 is not met for Joan's whole *research project*, which includes the three steps. Two of the steps do require primary data. However, if we only refer to step #2, namely for the computational work done by Jaume Forés, as it is the case, C4 was met.

(C5) An expected scientific contribution exists and the researcher finds its potential rewards satisfying

Both Jaume and Joan wanted to make a scientific contribution with only secondary data regarding the gene and biological relationship between autism spectrum and cancer and the role of TRIM29. So, condition C3 is a C3-SC.

When I asked Jaume why he was doing computational analysis for a research work that was not related to his PhD dissertation, he gave me three reasons. One of them was the potential rewards he could obtain with the publication of an article:

Jaume: *Jo crec que serien 3 respostes en realitat. En primer lloc, l'interés que desperta qualsevol qüestió científica. En segon lloc, la relació personal que m'unix amb Joan. Jo ara mateix no treballo per a ell de manera directa però m'agrada col·laborar amb ell per la relació personal que tenim. I en tercer lloc, donç, eh, si tot va bé, l'estudi conduirà a una publicació que des de el punt de vista professional, donç és... m'interessa.*

I ja està. Eixos serien els tres motius.

Conditions at a later time of the decision-making process (Time 2)

Event or outcome: The use of secondary data from GEO Profiles and TCGA happened as the only evidence of scientific claims as originally planned. The scientific contribution has consisted in three main scientific contributions. At the time I asked participants, only one of the contributions was published:

Forés-Martos, J., Catalá-López, F., Sánchez-Valle, J., Ibáñez, K., Tejero, H., Palma-Gudiel, H., ... Tabarés-Seisdedos, R. (2019). Transcriptomic metaanalyses of autistic brains reveals shared gene expression and biological pathway abnormalities with cancer. *Molecular Autism*, 10(1), 1–16. <https://doi.org/10.1186/s13229-019-0262-8>

According to Jaume, this publication only includes data from GEO Profiles. The other two were under preparation at the end of November, 2019, and included TCGA data:

Jaume: *En el paper que adjuntes no es va utilitzar TCGA però en la resta si que hem gastat les dades de TCGA com a cohorts de validació y també hi han datasets d'array express (pocs).*

Com comentava abans tenim dos articles més que están en fase de redacció!

Changes in conditions: There were no changes in any of the five conditions or in the researchers' structures and causal powers and liabilities. Neither in the data's causal powers and liabilities.

The decision-making process

The decision-making process of this case study follows this pattern:

(expected) outcome 3 → (final) outcome 3

The time horizon of the data-reuse mechanism is relatively short. Condition C1 is met before the idea of a potential contribution with the data. Conditions C2, C3, C4, and C5 happen most probably nearly simultaneously. Jaume found the same bureaucratic issues for accessing TCGA data [bounded rationality], but short after that he discovered that the access process had been simplified, and thus, was much easier [procedural rationality]. There was a complete level-of-analysis fitness with the data, because Jaume wanted to do the computational analyses with unprocessed data (or *raw data*). The reason why Jaume did not finally use TCGA data in the first contribution that was published in 2019 is that the TCGA data he needed are RNA-Seq, which is a different methodology from expression arrays in GEO Profiles. He needed the same type of data that Joan originally used previously, and which were based in expression arrays data. He needed a *technological fitness*¹⁵⁸ between the data used previously by Joan and the data he aimed to use. Also, both Jaume and Joan knew from the outset that GEO Profiles data had limitations, e.g., some variables are missing, some data are redundant, there are errors in the variables, etc., and, thus, they know what type of research questions they can formulate [*unbounded* rationality]. They always found the potential rewards of their scientific contribution(s) (C5) satisficing [teleological decision-making theory].

6.2. Case studies reusing *stewarded data*

Stewarded data are data available for reuse, and this is known by the secondary user, but are not publicly released or published (the data are available for others to reuse them, but there may be some type of walls, e.g., payment walls, confidentiality walls, technical walls, etc., or conditions on the reuse).

The stewarded data of the three case studies under this category are data from BORN Ontario data repository (Better Outcomes Registry & Network / Registre et Réseau des Bons Résultats dès la naissance) –BORN from now on–. BORN¹⁵⁹ is an Ontario prescribed data registry for health and care

¹⁵⁸ Jaume: *És més reproducible utilitzant la tecnologia més semblant, és més fàcil de resumir [33:01], és més senzill reproduir resultats inicials utilitzant la tecnologia més semblant perquè no estas introduint altres elements tècnics que distorsionen els possibles resultats.*

¹⁵⁹ <https://www.bornontario.ca/en/about-born/about-born.aspx> [January 19, 2020]

issues related maternity, newborns and children under the Personal Health Information Protection Act, 2004 (PHIPA 2004)¹⁶⁰. BORN is funded by the Ontario Ministry of Health and Long Term care, and administered by the Children’s Hospital of Eastern Ontario (CHEO), and its main goal is to facilitate quality care for families across the whole province, for which it collects, interprets and shares data. BORN also shares its data in order to support research and innovation in maternal, child, and youth health under the registry mandate. Thus, data do not belong to BORN. BORN stewards the data.

BORN data have all the causal powers and liabilities that I have theorized in the data-reuse mechanism. In fact, BORN data are carefully curated data to be shared for caring, research, and innovating in the field of maternal and child health. Data privacy and quality is BORN’s highest priority for which its staff follows a Data Quality Framework (DQF) with five dimensions, namely timeliness, accuracy, comparability, usability and relevance, and their respective elements and sub-elements as shown in Figure 22¹⁶¹.

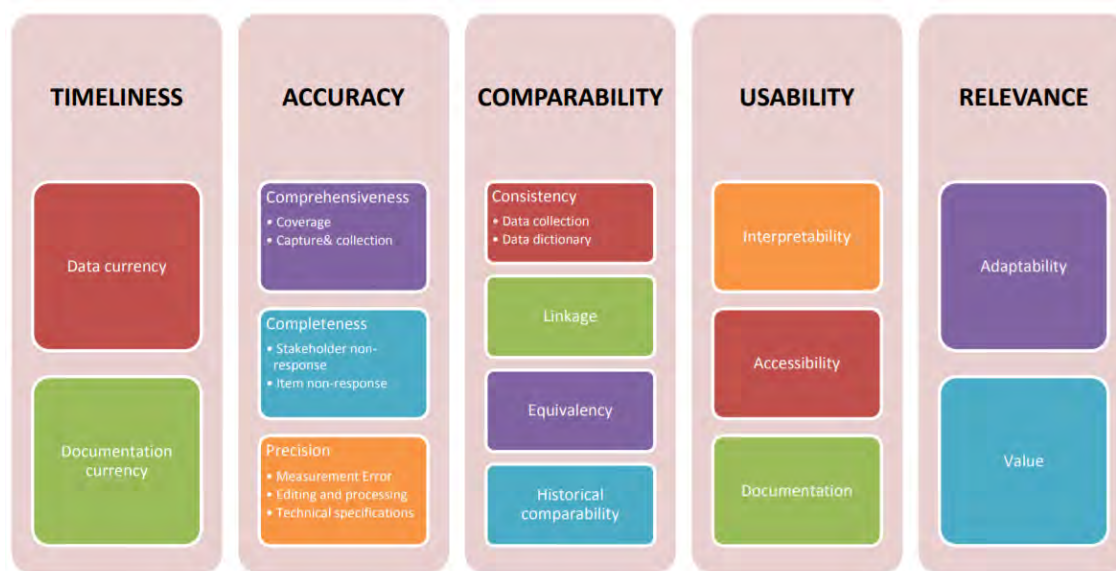


Figure 22 - Dimensions, elements, and sub-elements of BORN’s Data Quality Framework. Source: BORN’s DQF

Case study #5 also used data from ICES, which also hosts and manages stewarded data. ICES is the Institute for Clinical Evaluative Sciences in the province of Ontario¹⁶². Its data repository consists of several record-level, coded and linkable health data sets for about 13 million people. ICES data are

¹⁶⁰ <https://www.ontario.ca/laws/statute/04p03> [January 19, 2020]

¹⁶¹ Source: BORN’s DQF <https://www.bornontario.ca/en/data/resources/Documents/BORN-Ontario-Data-Quality-Framework-Summary.pdf> (page 2) [January 19, 2020]

¹⁶² ICES is a not-for-profit research institute encompassing a community of research, data and clinical experts, and a secure and accessible array of Ontario’s health-related data. <https://www.ices.on.ca/About-ICES/Mission-vision-and-values> [January 19, 2020]

very varied, and the system is fed from different interactions with patients, namely *physician claims submitted to the Ontario Health Insurance Plan, medical drug claims submitted to the Ontario Drug Benefit Program, discharge summaries of hospital stays and emergency department visits, claims for home care, long-term care*¹⁶³, etc. This makes ICES a powerful data source for research, but also a difficult-to-use one. According to my interviewee working at ICES as data specialist:

ICES data staff: *Data are extremely complex. We find it takes a typical analyst - so someone who we have hired to work with a researcher to analyze the data - it takes them about a year before they get comfortable with putting the data together.*

ICES data are also curated, but unlike BORN data's curation goals, ICES's curation is not targeted for specific research projects.

ICES data staff: *[...] All of the data at ICES are curated as you would call it. They're all standardized and cleaned, but the point is they're cleaned to remove sort of artifacts in the data. But what they're not cleaned for or what they're not prepared for is for any one given research study.*

With administrative data, [...], it's being collected for the purposes of billing. There's no flag in there that says someone had a heart attack. Certainly, when we receive hospitalized individual data we would clean it. But what cleaning means and curating means is ensuring that each of the variables has a standard format and ensuring that each of the data points if there's lots of missing that we look at why they're missing in some of the elements and we try and clean it from that perspective.

What it doesn't mean is we are not going to go through that data and identify everybody who has had a heart attack because there's no question in administrative data saying, "Do you have a heart attack or not?" So, when the clinician researcher who is interested in cardiology wants to analyze the administrative data they have to figure out who has a heart attack. So, it's those types of things where the data is definitely cleaned, but we cannot predict every single question that's going to be asked by a PI so we can't create flags for thousands of diseases in the data. Instead, the researcher has to go through and look at the hospital discharge codes and identify those codes that may be heart attack related. And those codes - some are very specific. Some clearly identify someone that's had a heart attack. Others will specify possible heart attack.

¹⁶³ <https://www.ices.on.ca/Data-and-Privacy/ICES-data> [January 19, 2020]

So, it's the PI who really needs to make the decision as to what they are going to use as an indicator of heart attack. Is that somewhat clear?

Therefore, although ICES data have all the causal powers and liabilities that I have theorized in the data-reuse mechanism, it is evident that a potential user of ICES data need more or other types of causal powers and liabilities than a potential user of BORN data.

6.2.30 Case study #5 (BORN Ontario data and ICES data)

In this case study we have the same researcher's structure and causal powers and liabilities and two different secondary data repositories' structures and causal powers and liabilities –BORN Ontario data and ICES data.

A word of caution is needed for explaining conditions in time 1 and in time 2 in this case study, titled *The relationship between 2009 pandemic H1N1 influenza during pregnancy and perinatal outcomes in Ontario*. When I interviewed Deshayne Fell, time 2 had already occurred for the research project in which she used secondary data, originating from BORN Ontario (BORN from now on) and ICES. In other words, the use of both BORN and ICES secondary data had already happened, and the outcome of the reusing process was already known when I interviewed her in March and April 2017, thus I had to track her process of using secondary data retrospectively, which started in 2012¹⁶⁴ (time 1) and ended in March 2016 (time 2). So, I have not been able to track changes prospectively in the conditions or causal forces of the mechanism from time 1 to time 2. However, despite potential biases on my side and on Deshayne's side about what were decisions and actions, and despite Deshayne's potential omissions about changes in the conditions, I would say that the evidence that I have collected about the process is accurate, though may be not complete. Yet, I have traced progress of this case study until January 2020, when I have validated results with Deshayne and searched for the publication of her manuscript.

Despite that time 1 (2011) and time 2 (2016) happened in the past, I sometimes use the present tense when reporting about this case since some conditions are still the same, for instance, Deshayne's causal powers and liabilities.

Although this case study is about the above mentioned manuscript, there are some references in the participant's narrative about her PhD, since the manuscript was part of her PhD dissertation.

¹⁶⁴ Deshayne Fell started her PhD in 2011, but it was not until 2012 that she started developing the design of her study with BORN Ontario data and ICES data.

Conditions at the outset of the decision-making process (Time 1)

In time 1 (2011), Deshayne Fell was working as a Perinatal Epidemiologist at BORN. Actually, she started to work there in 2009. She already had a master's degree in epidemiology at that time and had lot of experience working as an epidemiologist, including extensive experience working with secondary data even before starting working at BORN. She decided to pursue a PhD degree in 2011. As a result of her willingness to pursue a PhD degree, she has also *necessary internal relations* with other *structure*, McGill University, which provides her with the causal powers and liabilities to carry out research as a PhD candidate. Deshayne knows the reward system of science for academic researchers in epidemiology, and that her research career depends on this system.

For the topic of her PhD dissertation, she was inspired by both her prior working experience with other administrative data repositories and her work at BORN, more specifically by a surveillance project requested by the Canadian Federal Government to BORN in which she was the leader. The topic was the 2009 flu pandemic involving the H1N1 influenza virus, namely how the virus and the vaccination against it could affect pregnancy and perinatal outcomes.

Deshayne Fell has, with no doubt, all the causal powers and liabilities that I have theorized in the data-reuse mechanism for using both BORN data and ICES data. She is a very experienced researcher, and tries to perform her research with the highest level of rigor. Before starting her PhD, she knew perfectly the limitations of doing secondary analysis of BORN data and ICES data, at least when used independently, although she was not aware of the constraints she had to face when linking both data at the individual level due to privacy and confidentiality reasons. She also had the causal power and liabilities of accessing the most restricted level of BORN Ontario data because of her *necessary internal relation* with BORN Ontario.

(C1) The researcher knows that secondary data exist

Deshayne Fell knew both BORN data and ICES data before starting her PhD, as the validated-by-participant workflow diagram in Figure 23 (full size in annex 16) and part of our conversation reflect:

I: *So [both BORN and ICES] w[ere] a resource, you know, in the corner of your mind that you knew that you were going to use?*

Deshayne: *Yes.*

I: *Okay.*

Deshayne: *Yeah, I didn't have to go doing a lot of homework. I already knew what was there.*

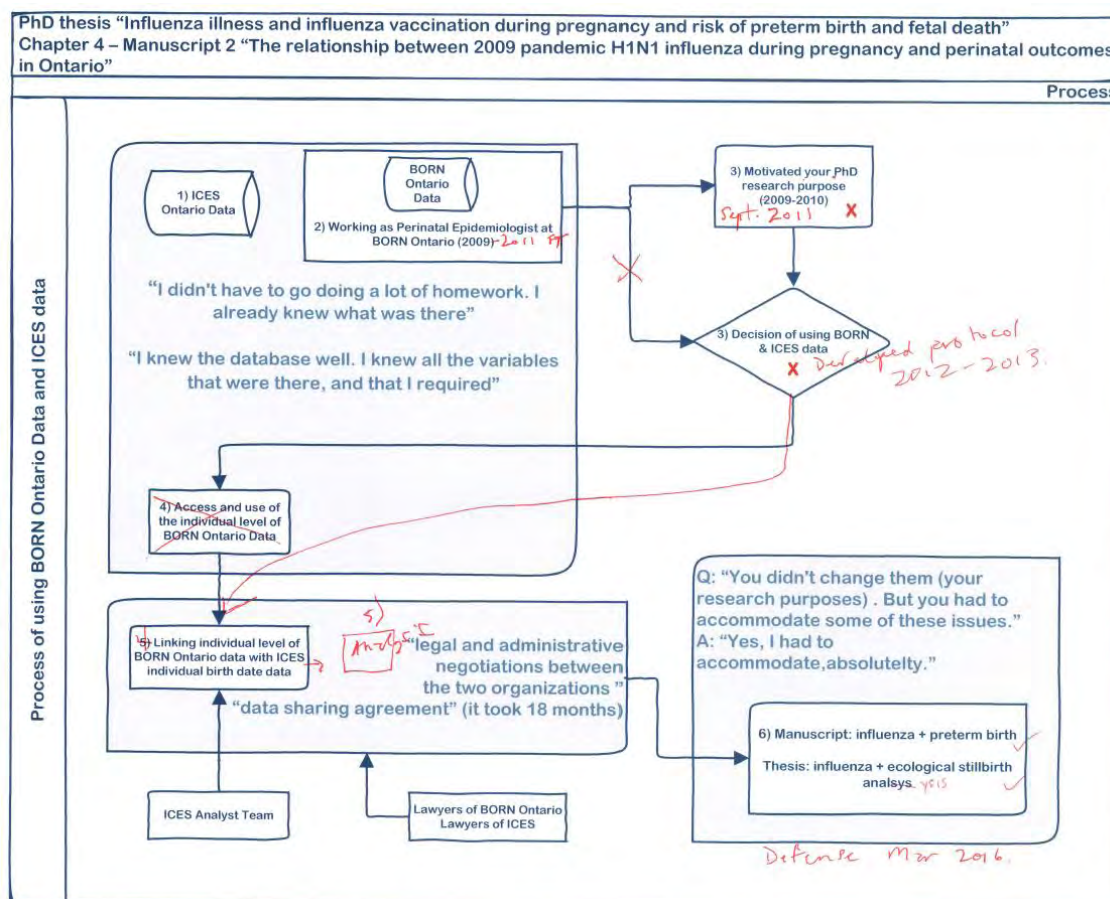


Figure 23 - Workflow diagram of the data reuse process of case study #5

(C2) Data are obtained

Deshayne needed several variables from the BORN data base related to her PhD research topic (H1N1 influenza, vaccination, pregnancy, and perinatal outcomes). Access to BORN data was easily obtained since Deshayne Fell was an employee at BORN, and the internal relation with BORN (her main structure) provided her with the causal power to have the restricted access to the individual data, although not after having obtained first ethics clearance from three institutions (McGill University, Children's Hospital of Eastern Ontario, the Ottawa Hospital Research, OHSN).

However, Deshayne could not obtain ICES data in the strict sense of the word *obtain*. Due to privacy and confidentiality reasons, ICES could not provide Deshayne with the individual data in order to match variables with the individual data from BORN. Instead, she had to securely transfer BORN data

to ICES, and rely on someone else to do the record linkage for her. She also had to analyze the outcomes of the linkage onsite at ICES, within a secure network environment.

Deshayne: *I mean the biggest challenge was when you work over at ICES, they also have their own review that happens, an internal review. They have their own privacy office and they review every project. Kind of, they're almost like a REB as well, so I had their approval to go through as well, and of course they're the ones who raised the red flag about having the exact date of birth and postal code, so they wouldn't let me have those variables.*

(C3) Particular secondary data are an initial *satisficing* option

At time 1, both BORN data and ICES were an initial *satisficing* option. Condition C3-SC is met since both data were to be used for a PhD dissertation and a journal article.

I: *You knew you were going to use [BORN data]- that they were useful?*

Deshayne: *Yeah, absolutely.*

I: *It had useful information. Okay.*

Deshayne: *Not only the BORN data, but the administrative data over the Institute for Clinical Evaluative Sciences, ICES.*

I: *Okay.*

Deshayne: (minute 12.09) *So, yeah, and I already, although I had never maybe done a project there [at ICES] yet, that type of a data warehouse and those big administrative data sets is something I already had a lot of experience with in the past, so I knew the value of shipping it over to link it up and how I could use the data.*

(C4) The idea of collecting particular primary data is not an initial *satisficing* option

Condition C4 was met at time 1. Deshayne, as an epidemiologist used to manage huge amounts of data, she never thought or has thought of collecting primary data for her manuscript, nor for her PhD dissertation.

I: [...] *-- so in some disciplines even related with health the use of secondary data is not well seen. I mean, a PhD student would never be allowed to present a thesis with secondary data. In your case, it is in epidemiology. Is that correct?*

Deshayne: *Yes.*

[...]

Deshayne: *Very.*

[...]

Deshayne: *Yeah. Classic epidemiology really is about the population. And now with databases being so available, more and more you're seeing big data research. So, I mean, I think -- I bet you if you went to McGill, which is where I did my training -- you know, if you went there 20 years ago, I bet it was probably 75 percent of students probably did primary data collection and maybe 25 percent did some sort of secondary data. And now I would say it would be 80 percent secondary data and 20 percent primary. Maybe even higher for secondary data.*

I: *Even higher? Yeah, I guess. Okay.*

Deshayne: *People just aren't doing that anymore, which is a bit of worry I think. Because these databases are fantastic for having big numbers and a whole population, but what you gain in size you lose in depth in terms of the clinical information that's there. And there are something that you just can't do well with the limited information because they're always assembled for other purposes. [...]* *But with that being said, that's all I ever do. So, I don't do any primary data collection.*

(C5) An expected scientific contribution exists and the researcher finds its potential rewards *satisficing*

Deshayne decided to do a PhD degree in the field of epidemiology. She opted for a manuscript-compendium dissertation in which she finally included three manuscripts from the four studies that she conducted. Only manuscript number two involved BORN and ICES data. Both her dissertation (titled *Influenza illness and influenza vaccination during pregnancy and risk of preterm birth and fetal death*) and the manuscript she expected to write and publish from her study were *satisficing* for her in time 1 (2011). She expected to submit the manuscript for publication to the Canadian Medical Association Journal. However, she and her co-authors published it in the journal *Epidemiology*. So, condition C5 was met.

This is what Deshayne wrote in her dissertation:

Prior to commencing my doctoral studies, I worked as a Perinatal Epidemiologist for a number of years and during this time was encouraged by several mentors to pursue a PhD. (Fell, 2015, p. viii)

Manuscript 2: Fell DB, Platt RW, Basso O, Wilson K, Kaufman JS, Buckeridge DL, Kwong JC. The relationship between 2009 pandemic H1N1 influenza during

pregnancy and perinatal outcomes in Ontario. To be submitted to the Canadian Medical Association Journal.

I developed the original protocol for this study, which evolved over time through consultations with Dr. Platt and Dr. Basso. I provided advice to the Institute for Clinical Evaluative Sciences (ICES) on technical aspects of the record linkage and conducted my analyses onsite at ICES in Ottawa. Due to restricted access to date of birth information, which was required to implement the time-varying methodology, some of my programs were submitted on my behalf by Robin Ducharme, a Research Analyst at ICES. I compiled and interpreted all the results and drafted the manuscript. Dr. Platt, Dr. Basso, Dr. Wilson, Dr. Kaufman, Dr. Buckeridge and Dr. Kwong all contributed to the interpretation of the results and provided methodological advice. All of the authors critically reviewed the article for intellectual content. (Fell, 2015, p. x)

Conditions at a later time of the decision-making process (Time 2)

Event or outcome: The use of secondary data from BORN and ICES happened as the only evidence of scientific claims as originally planned. The final outcome (Deshayne's manuscript 2 in her PhD dissertation) was finally published in January 2018 in Fell et al., 2018:

Fell, D. B., Platt, R. W., Basso, O., Wilson, K., Kaufman, J. S., Buckeridge, D. L., & Kwong, J. C. (2018). The Relationship between 2009 Pandemic H1N1 Influenza during Pregnancy and Preterm Birth: A Population-based Cohort Study. *Epidemiology*, 29(1), 107–116. <https://doi.org/10.1097/EDE.0000000000000753>

Changes in conditions: There were no changes in any of the five conditions nor in Deshayne's structures and causal powers and liabilities. Neither were changes in BORN's causal powers and liabilities from time 1 to time 2.

The decision-making process

The decision-making process of this case study follows this pattern:

(expected) outcome 3 → (final) outcome 3

The time horizon of the data-reuse mechanism of this case study (Deshayne's manuscript #2 of her dissertation) is long with a lot uncertainty about whether condition C2 could finally be met with regard to ICES data. In fact, this case study did not meet condition C2 in the strict sense of *obtaining* the data since Deshayne was not allowed to match the individual variables of BORN data with the individual variables of ICES data by herself.

She wanted to make a scientific contribution for which she needed variables from both BORN data and ICES data, and the variables had to be linked at the individual level. For linking BORN and ICES data she had to obtain *ethics clearance* from ICES, apart from obtaining ethics clearance from the above mentioned institutional REBs (Research Ethics Boards). The only way that the linkage was possible was by securely transferring BORN records at the individual level to ICES. For the shipment and the data linkage, a legal data sharing agreement between both institutions was required. The process of signing this agreement between BORN and ICES took one year and a half, but Deshayne never knew how long it was going to take [bounded rationality]. She waited with uncertainty for eighteen months while working at BORN and preparing other manuscripts for her dissertation. The latter was possible because she did not need any data shipment for the other manuscripts [procedural rationality].

Deshayne: *But what's interesting is that there was never a point prospectively at which I knew how long I was going to have to wait. Like it wasn't like they said to me in 2011, "This data sharing agreement will take 18 months." You know what I mean? It was always like, "Oh, well, it should be done soon. It should be done soon." And so you never -- like I never knew when I would ultimately --*

I: *Yeah. You didn't have that information.*

Deshayne: *No.*

[...]

Deshayne: *Yeah. I didn't know how long I would have to wait ultimately. Yeah.*

In Deshayne's account of the fact that both institutions needed eighteen months to have a data sharing agreement between them, there seems to be an acceptance of the fact [satisficing], though we have to keep in mind that her account takes place during our face-to-face interviews (April and May 2017). Indeed, everything was over at the time of the interviews because, at that time, she already had obtained her PhD degree (her defense was in March 2016). However, I suggest that waiting for so long was *satisficing* for her for two main reasons. On the one hand, she never knew how long she had to wait. Time was passing away while she was waiting for the data sharing agreement to be signed by all parties. On the other hand, she could invest her time and efforts in accomplishing other parts of her PhD dissertation.

Deshayne: *So, buuuut, the challenge that I faced was the fact that I wanted to take the data and send it over to ICES, which is a separate entity and have it linked to the administrative databases. And so, as it turned out, at that time, so I had my PhD study to look at influenza disease in pregnant women, but concurrently we were also funded by CIHR to do a study where we wanted to look at infant outcomes over one year following their birth and again comparing women who had the H1N1 vaccine in pregnancy and those who didn't. So, for both of those projects, which I was involved with, we needed to take the BORN data and ship it over and get it linked up. I kind of got... uh, a little bit, uh, what's the word? "Caught" in the legal and administrative negotiations between the two organizations that had to happen in order to allow that transfer of data because BORN is a prescribed registry under the provincial legislation and ICES is also what's called a prescribed entity under the provincial privacy legislation.*

So, they have to be so careful about confidentiality and privacy and all of those things with the data. So, in the end it took 18 months to get a data sharing agreement that all the lawyers in both organizations were happy with before the data could go.

[...]

I waited 18 months to get access to the data. And so that was my study number two. I had three studies. I actually had four studies, but I only included three in my dissertation.

My first one was a systematic review. So, at least I did that one kind of first while I was waiting. And then my third study I was able to do as well, it was using a different data set but here at BORN. There were no linkages. No agreements were required.

[...]

It was relatively easy. I just needed usual approvals and REB, of course, but that was it. That was very statistically complex, that study, so that kept me busy for a while. And so, it all worked out okay in the end, but imagine if that study was my only study, I would have been waiting for nothing, well not for nothing, I would have been waiting with nothing to do for 18 months. That's how long it took to get the agreement. So, that was a huge challenge. Uh, once the agreements were in place, it started to go more smoothly, like the data from BORN were securely transferred. I worked quite closely with the linkage team over at ICES to do the linkages, which were also complicated. And, uh, yeah, but I then eventually I got access to the link data set on that site. [...]

Deshayne also faced other relevant challenges when using ICES data. One had to do with the linkage: about 20% of the stillbirth cases for her study time period could not be linked with the ICES databases. This linkage issue with the stillbirth records posed an important challenge¹⁶⁵ with respect to the independent variable “pandemic exposure”. She was not able to access the variable regarding diagnosis, which was in the ICES database. Instead, she had to use the time period of the pandemic as a proxy for exposure (ecological exposure), which according to Deshayne “is okay, but not great” [satisficing]. Again, she never knew that she was going to face those challenges, or rather the consequences of those challenges for her research question. However, when she faced those challenges, she *accommodated some issues* instead of giving up the usage of ICES data, and thus giving up her study. For instance, one of the other accommodations was that Deshayne dropped the stillbirth outcome for manuscript #2, and instead she provided analysis results only on the preterm birth outcome for the manuscript. However, she still included the stillbirth analysis in her PhD dissertation, but it was an ecological study. Both solutions were satisficing for her, although not what she wanted to achieve.

These two challenges were the most relevant, but other unexpected major challenges appeared along the process¹⁶⁶. Yet, none of them prevented Deshayne from using BORN and ICES data and from linking the two of them. C5 was constantly being met from the outset until the end of the process despite Deshayne’s uncertainties and the challenges she faced. She was always able to adopt strategies to counteract the challenges that she was constantly facing.

In sum, with regard to BORN data, condition C1 was met before the decision of pursuing a PhD degree (C5). I would even say that condition C4 was met also before C5, according to Deshayne’s own words (*So, I don’t do any primary data collection*). After C1 and C4 were met, condition C5 was met followed by condition C3-SC with maybe a time span of several months. Condition C2 for BORN

¹⁶⁵ For the main analysis (preterm birth), the linkage was good and Deshayne was able to use the diagnosis information from the ICES databases.

¹⁶⁶ **Deshayne:** *I don’t know if I told you last time that like initially when I went in and I started my study in the student lab, I had sort of access to all of the information I required to do my study. And then after about two months there was a change in policy and they said, “Okay, you can’t have this one variable,” and that one variable was essential to the type of analytical framework I was using. I needed to have it and they absolutely said, “No, you cannot. You can’t have it.” And so I had a breakdown and then thought about it. And what I think -- did I tell you this last time? [...] Yeah. And so then what I ended up having to do was parts of the analysis I could do myself in the lab, but I would say 70 percent of my statistical models I had to -- I had to write all of the code and then send it to their analyst. She would run it because she could access all of the information that was required and then she would get the output and transfer it back to me so I could look at the output. So, it’s very -- [...] But still it’s really hard when you have hard complex, long programs of syntax and complex models to code blind. Like usually -- at least the way that I work is I’m very iterative with my data, right? So, I’ll run something, look at how the output is, and do a lot of checks to make sure it’s doing what I want it to do. And I wasn’t able to be as iterative with the data and so I had to be more thoughtful about putting certain steps into place in my code that would give me some information that the analysts would produce that would still give me that feedback that the way I had coded and set things up was working properly. So, it just -- I just had to be a lot more thoughtful, careful, and organized than I would’ve if I was doing it all myself. So, that was a big challenge, but...*

data was met immediately after C5, since Deshayne had permissions to access individual level data. Yet, she strictly used only the variables that she needed for her manuscript.

The sequence of conditions for ICES data is the same as for BORN data. However, condition C2 for ICES data was partially met [satisficing] and only after one year and half [bounded rationality].

6.2.4. Case study #6 (BORN Ontario data)

In this case study, which is related to the relationship between maternal obesity and stillbirth and neonatal death, when I interviewed Mary Smith, PI of the research project, time 1 and time 2 had already occurred in the past. Time 1 occurred in summer 2014 and time 2 occurred just before I interviewed Mary Smith in March 2017. Therefore, as it happened with case study #5, I was able to trace the conditions of the process of the use of BORN data only retrospectively.

For this case study, I drew two diagrams. One represented the process of using BORN data, which was a very simple and brief one (Figure 25 or annex 18). The other diagram represented the context in which the use of BORN data took place (Figure 24 or annex 17). The participant validated none of the diagrams because I only interviewed her once¹⁶⁷, and thus I had no opportunity to show her the diagrams¹⁶⁸.

Conditions at the outset of the decision-making process (Time 1)

Mary Smith had all causal powers and liabilities of the data-reuse mechanism at both times 1 and 2 because she was an associate researcher at the Ottawa Health Research Institute (OHRI). However, she belonged to two more structures (The Ottawa Hospital (TOH) and the University of Ottawa (uOttawa)). The internal relations that Mary had with these two structures, provided her with specific causal powers and liabilities. For instance, she could teach, and thus could supervise and hire students for research projects since she was assistant professor at uOttawa. Also, she knew BORN data quite well, at least the variables related to the health problem that she studied, because of her role as clinician at TOH. As a clinician in maternity issues, she was used to enter data in BORN platform.

¹⁶⁷ I emailed Mary Smith several times in which I asked her for the second meeting, and for the possibility of interviewing the student who actually conducted the little analysis of BORN data. However, I never received a reply.

¹⁶⁸ However, in view of the minor corrections of workflow diagrams by participants in the rest of the nine case studies, I would suggest that the WD I prepared of this case study would have received also minor correction, if any.

However, only two of these structures, namely uOttawa and OHRI were the ones that allowed the use of BORN data for the specific research project that I interviewed Mary about, and which I tentatively titled *The effect of maternal obesity on stillbirth and neonatal death*. The diagram in Figure 24 represents these two structures and their relationship with the use of BORN data by a hired student during summer 2014. She is subject to the reward system of science for academic researchers, and she knows the rules.

Unlike Deshayne Fell (case study #5), Mary's structure did not include BORN Ontario, so she did not have the causal power and liabilities of accessing directly any level of BORN Ontario data. The only way that she could access BORN data was by following BORN's the request data protocol¹⁶⁹, either for aggregated data or record-level data. Mary was an *external* user of BORN data.

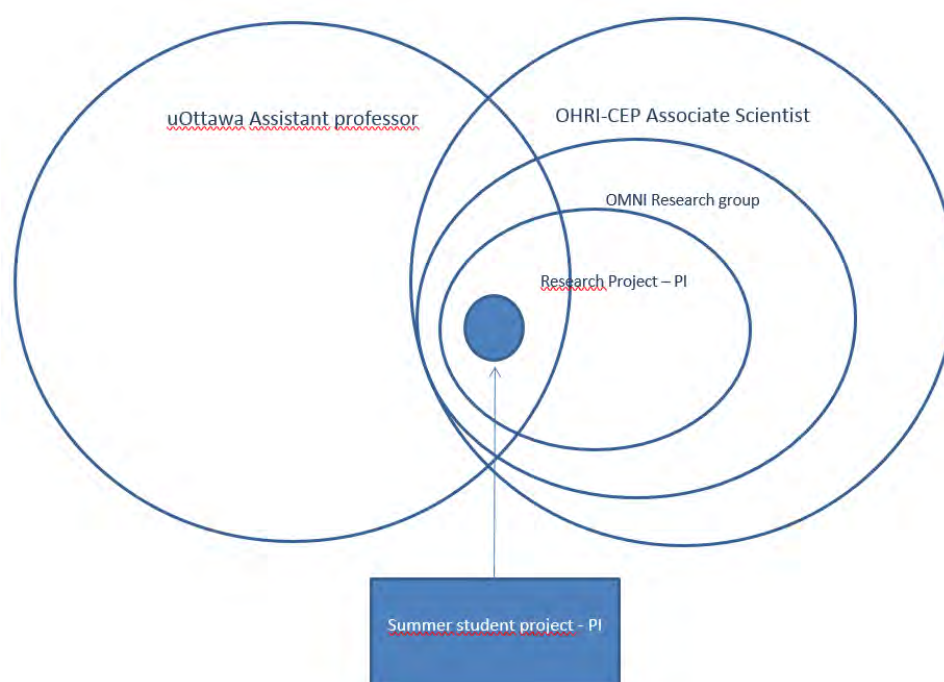


Figure 24 - Situating the reuse of data within a larger research project. Case study #6

(C1) The researcher knows that secondary data exist

Condition C1 was met at time 1, and even before of the conception of the research project as the workflow diagram shows. Although it was not validated by Mary, her verbatim words shadowed in grey confirm that she first new about BORN when she was a master student, and that she later used BORN platform or system as a clinician since she has to report data about pregnant women.

¹⁶⁹ <https://www.bornontario.ca/en/data/requesting-data.aspx> [20 January 2020]

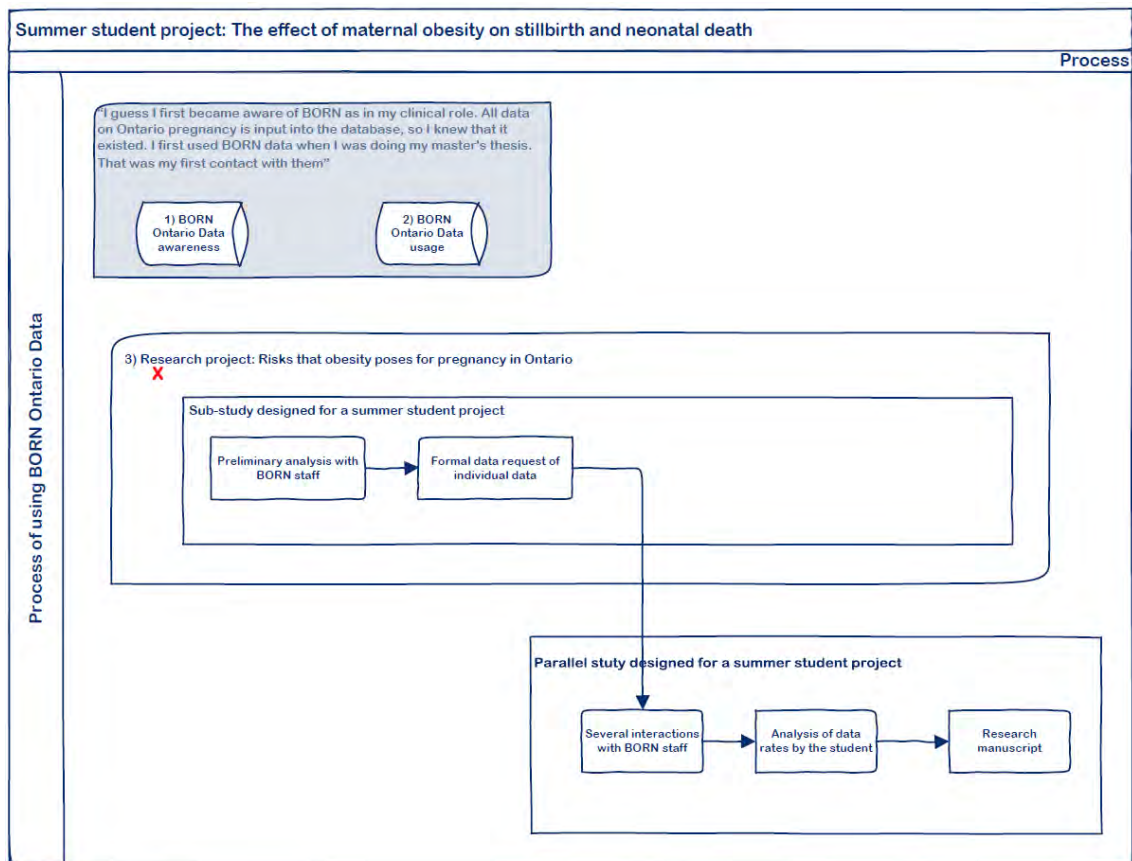


Figure 25 - Workflow diagram of the data reuse process of case study #6

(C2) Data are obtained

Condition C2 was not met for the data details and variables that Mary Smith and her student needed. They needed record-level data and variables by means of which patients could be identified. Since the summer project only lasted four months, they could only have rates, not even aggregated data. They would have needed between 12 and 18 months to have the record-level data from BORN staff.

Mary: [...] *We had, as I said, a fairly detailed analysis planned. When we went to them and said, "What are the timelines? How much time, how much money would this involve?" It became clear that for it to be doable in a four-month time span, it had to be very limited. And, so, at the end of the day, we were only given aggregate data, not individual level data in any way. In fact, we were given rates rather than whole numbers. And so, that is the essence of that project and what data we obtained from BORN.*

[...]

(minute 17:02) *Probably the biggest problem I encounter is the time involved. Typically, it takes 12 to 18 months to get data out. And that can be a major problem.*

[...]

(minute 17: 43) *To get what I would have originally wanted was going to be a year-and-a-half. I had a summer student coming in a very short period of time, and so I had to say to them, "Can you then at least give me some basic data so that the student has something to work with?"*

I: *Ok, so you didn't request everything you needed. You adapted to the timeframe of the student.*

Mary: *Originally, I've requested what I needed, but we had to modify that based on the timeline.*

[...]

(minute 19:02) [...] *Like I said, because of how this project ended up, there was not a lot of analysis. [The student] was basically given the numbers back. And so, she put those into a table and generated some graphs of that data. She subsequently finished the write-up and has presented the data.*

(C3) Particular secondary data are an initial *satisficing* option

BORN data were indeed a satisficing option. Condition C3 was met.

Mary: *I guess I first became aware of BORN as in my clinical role. All data on Ontario pregnancy is input into the database, so I knew that it existed. I first used BORN data when I was doing my master's thesis. That was my first contact with them. That was the first time I learned how to access data, what the limitations to accessing the data was. Then, because I had been through that experience, when I came to this research question, I was fairly confident I'd be able to get what I needed by going through that data set.*

I: *By the time that you were accessing those data for this project, you knew quite well what was there and-*

Mary: *At least within my realm of the questions that I would ask, yes.*

(C4) The idea of collecting particular primary data is not an initial *satisficing* option

I did not have the opportunity to find out about this condition with the participant. However, considering the time constraint for obtaining and analyzing record-level BORN data in four months, I would say that the collection of primary data was not *satisficing* at all. I would even say that this option was impossible, since it is not feasible to collect data prospectively during a summer project in health topics like the one at hand. A REB process for collecting primary data would take at least from 2 to 4 months, at least at OHRI where the PI is an associate researcher¹⁷⁰.

(C5) An expected scientific contribution exists and the researcher finds its potential rewards *satisficing*

Mary conceived a major project to find out the effect of maternal obesity on stillbirth and neonatal death, for which I give for granted that she wants to make a scientific contribution. However, within the major project, there are sub-projects or parts, which tackle different aspects such as, for instance, the four-month summer student project that focused on delivery times of obese women. This project was a kind of preliminary exploratory study. Mary and her student submitted the scientific contribution of this part of this major project carried out with BORN data at time 2, just before I interviewed Mary.

Mary: *So, the project overall is looking at the risks that obesity poses for pregnancy, and this part specifically on timing of delivery of women who have significant obesity. There are multiple data sources now that suggest that there's an increased stillbirth rate in the obese population. And for many other populations, we induce people early, or deliver them early, if they're at an increased risk of fetal demise. And so, I was interested in finding out in Ontario what the chances of demise are in a woman who's obese and at what gestational age that risk substantially increases.*

[...]

We had planned very detailed analysis. However, before embarking on that, we just wanted some very, very simple data that shows that there is a difference in the population in Ontario. And so, we went to BORN and said, "Can we just have numbers for stillbirth and neonatal death by body mass index?" Uh, it was a project initially that was designed for a summer student. We had, as I said, a fairly detailed analysis planned.

¹⁷⁰ I know this by my own experience, and by the rest of interviewees of the case studies, which needed ethics clearance for using secondary data.

Conditions at a later time of the decision-making process (Time 2)

Event or outcome: The use of BORN data happened as the only evidence of scientific claims, although not for the initial research question –RQ (a)– as originally planned due to the constraints of not having the record-level data. Instead, a new research question –RQ (b)– was answered with BORN data.

I have been able to find only these two publications from my Mary Smith on her Researchgate profile. However, I am not sure if the scientific claims in these two journal articles are the result of the research project at hand, although they are based on BORN data.

El-Chaar, D., Guo, Y., Corsi, D., White, R., Gaudet, L., Walker, M., & Wen, S. W. (2019). Caesarean delivery on maternal request in Ontario: trends and determinants. *Journal of Obstetrics and Gynaecology Canada*, 41(5), 716.
<https://doi.org/https://doi.org/10.1016/j.jogc.2019.02.179>

Guo, Y., Miao, Q., Huang, T., Fell, D. B., Harvey, A. L. J., Wen, S. W., ... Gaudet, L. (2019). Racial/ethnic variations in gestational weight gain: a population-based study in Ontario. *Canadian Journal of Public Health*, 110(5), 657–667.
<https://doi.org/10.17269/s41997-019-00250-z>

In an email message from my participant in February 4, 2020, she informed me that “[t]he project we discussed is STILL ongoing. I remain hopeful that one day soon it will be submitted for publication”. However, I deduct that she refers to the umbrella research project, which included the summer student project.

Changes in conditions: There were no changes in any of the five conditions from time 1 to time 2, maybe because the summer research project lasted a short period (four months), and thus there were not many opportunities for changes. There were not changes in the student’s and Mary’s structures and causal powers and liabilities. Neither were there changes in BORN data’s causal powers and liabilities.

The decision-making process

The decision-making process of this case study follows this pattern:

(expected) outcome 3 with RQ (a) → (expected) outcome 3 with RQ (b) → (final) unknown outcome

All four conditions C1, C3-SC, C4 and C5 happened simultaneously and all were met at time 1. Condition C2 was not met at time 1, but Mary knew this fact from the very first moment. Mary and the student had no uncertainty regarding what they could obtain from BORN staff for a four-month student project in order to answer RQ (a) because they checked the options at the outset of the summer project with BORN staff [*unbounded* rationality]. As they knew that, they could not have the individual data, Mary deployed an alternative for the summer project [*procedural* rationality]. Mary decided to formulate a new research question RQ (b) which could be answered with the data they would obtain by the student.

I: *Ok. With regard to the project itself, which started with being this project for a student, did that at any point change the initial question? Did you have an initial research question? That will be my first question.*

Mary: *Yes, we did. And yes, I would say we modified it based on the data we were able to obtain in the timeframe that we needed.*

[...]

Mary: *I actually contacted them first and confirmed that they would be able to offer me the data. Body mass index in particular is a little bit problematic through BORN. There's a high missing rate. We went through a little preliminary analysis where we looked at pregnancies where data was missing and pregnancies where data was available, and just made sure that there were not any huge discrepancies between those two groups. We're basically making sure that the data was representative. Once we knew that, we put in a formal data request and then we went through several iterations of what they would actually be able to give us in the timeframe we needed it.*

I: *Can you develop a little bit more about these interactions that you said with BORN Ontario? What are they about? Why did you need several interactions with them for the data?*

Mary: *It's always a conversation about respecting privacy and not taking advantage of patients whose data has been collected for a totally different reason. I feel like as a researcher, when we request things, we are asking for our dream list (minute 12:43), which often includes things that cannot be provided while respecting patient privacy. Uh, and so, that's where acting with one of the BORN analysts, we go back and forth and say, "Look, I'd like to have this," and they say, "I can't give it to you because that's identifying information." Um, and then, we work out ways around that. Either I have to give that up or maybe I need to give up some uh, some of the uh, uh, sort of the strict criteria that I might normally use [satisficing]. For example, I might like to have patient's birthdate, but they can't give that to me because that may identify the patient, so we'll go to birth month or birth year. You know, there's lots of back and forth on that.*

I: *Ok, trying to adapt the way that they could give you the data with your-*

Mary: *To try to match the question as carefully as possible.*

Time 2 happens four months after time 1. There was no change in Mary's decision. She did in time 2 what she decided in time 1, which consisted of going forward with using BORN data with the new research question RQ (b), for which she designed a new *parallel* study for the student (see the workflow diagram).

A hypothetical value regarding condition C2 (no data could be obtained at all, so C2 would have not been met) would have ended with this decision-making process:

(expected) outcome 3 with RQ (a) → (final) outcome 1

I: *Uh, what would have happened if[...] you know, imagine that BORN Ontario couldn't have been able to give you those specific data or would have given you the data in a way that you finally could not adapt them for that summer project?*

Mary: (minute 27:11) *I would have given her a different summer project.*

I: *A different one, okay.*

Four months is not a long time. It's not like I can look through another data set. It was either going to work or we had to give her something else to do.

6.2.5. Case study #7 (BORN Ontario data)

This case study needs two caveats. On the one hand, when I interviewed my participant, time 1 (2014) had already past, and time 2 (2017) was still going on since the analysis of BORN data was going on and the scientific contributions started to be presented at conferences. On the other hand, in time 1, I refer to this project as a *surveillance project* or public health practice project, and not as a *research project*. Public health practice and public health research follow different research and ethics protocol, legislations and norms.

[Public health] [p]ractice is about protecting the public's health. It includes epidemiological investigations, surveillance, programmatic evaluations, and clinical care for the population. These activities are the essence of what public health people do [...]. Underlying many of these activities is the collection and analysis of identifiable health data by a public health authority for the purpose of protecting the health of a particular community.

Public health authorities, however, also design and conduct research involving human subjects for the purpose of generating knowledge that often benefits those beyond the participating community who bear the risks of participation. Public health practitioners engage in research activities for reasons similar to any researcher's interests: they seek to explore hypotheses, advance current knowledge, and contribute to the welfare of persons beyond the study itself (Hodge & Gostin, 2004)

Conditions at the outset of the decision-making process (Time 1)

My participant in this case study, Sarah Wilson, holds a BSc, a MPH (Public Health and Epidemiology master's degree), and a PhD. She has all the causal powers and liabilities that I have hypothesized in the data-reuse mechanism since Sarah knows how to conduct research. Moreover, she has specific causal powers and liabilities to carry out *public health practice* since she works in a public health unit. In this role, her work and *research* performance is not subject to the reward system of science, but to the public health surveillance system and its norms. Yet, Sarah knows the reward norms of science in the field of epidemiology, and has the liability to adapt to these norms.

At time 1, and for the purposes of Sarah’s surveillance project in epidemiology related to maternal and neonatal health¹⁷¹, she belonged to two structures. On the one hand, she worked at one of the public health offices in the country. On the other hand she was a BORN Ontario *agent*. A BORN agent is someone who is not BORN staff, but is given this status to access the restricted level of a portion of BORN data for a specific study during a specific time of period. BORN agents work very closely with BORN staff, and are trained in the PHIPA regulations. Sarah became a BORN agent in order to carry out a specific study on maternal and neonatal health for which she needed some additional data on the health problem that the public health office wanted to address. Her role as BORN agent provides her with specific causal powers and liabilities to access BORN data at the record-level.

(C1) The researcher knows that secondary data exist

This condition was met before Sarah started to work with BORN data in 2014 (time 1). She knew BORN Ontario (originally named Niday) when she was a student. See her own quote in Figure 26 (a full size in annex 19).

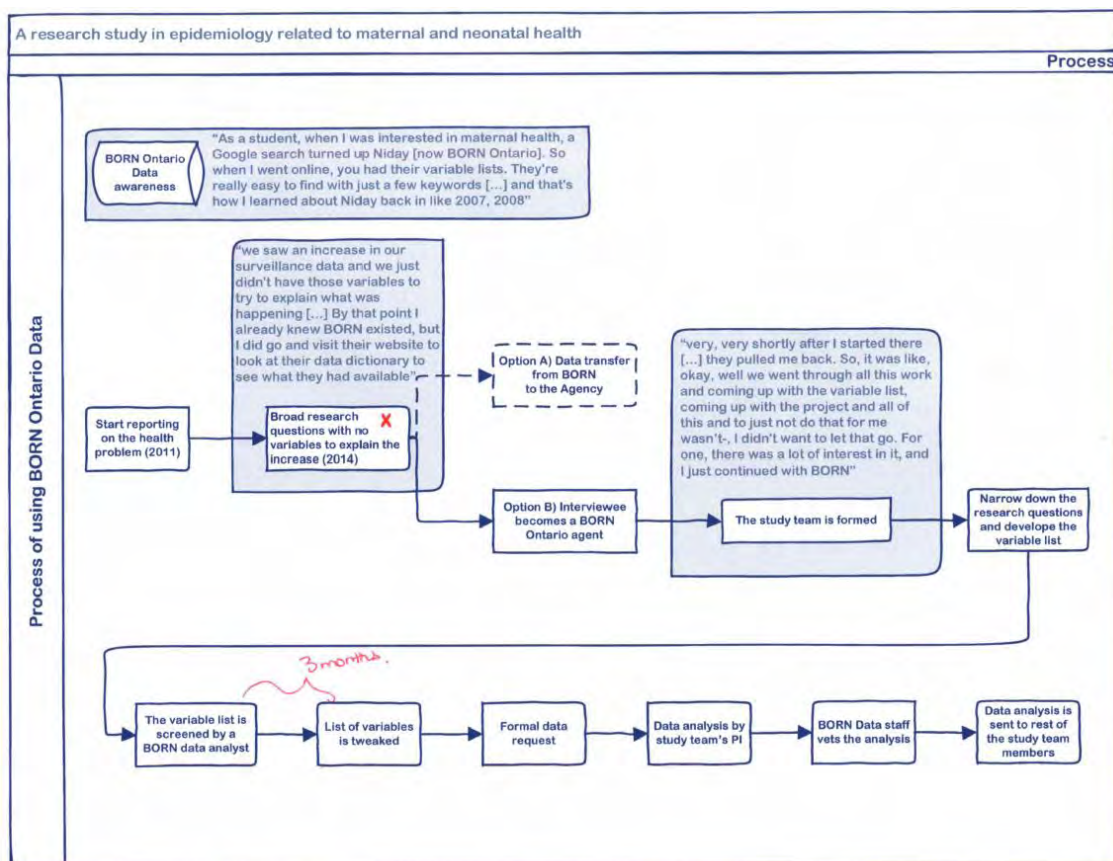


Figure 26 - Workflow diagram of the data reuse process of case study #7

¹⁷¹ I must not provide more details about the type of research that my participant conducted, since she could be identified despite anonymization.

(C2) Data are obtained

Condition C2 was met from the very moment when Sarah became a BORN agent, since she had the same permissions, rights and duties with the data as any BORN employee. However, she only had access to the data and variables at the record-level that she strictly needed for the study.

(C3) Particular secondary data are an initial satisficing option

Condition C3 was met for BORN data.

Sarah: *(minute 19:31) Okay, so as I explained, we saw an increase in our surveillance data and we just didn't have those variables to try to explain what was happening. We suspected strongly that [some factors] there were likely contributing factors to this, but also we know that there is genetic susceptibility to [...], as well as repeat pregnancies, that kind of thing. We were limited into what we can look at. By that point I already knew BORN existed, but I did go and visit their website to look at their data dictionary to see what they had available. When I looked at it then, it had evolved from when I knew it as Niday. Niday was very limited in terms of what they were able to produce. I don't even recall if [...] was a variable that they were able to report on back then. I don't recall 100%, but I don't believe it was there. Looking forward, it was like okay, all of these elements that we want to look at are there [at BORN Ontario].*

Sarah also needed ICES data (the same that Deshayne Fell in case study #5 used)

Sarah: *[...] As we started to narrow down the research questions, there was actually a fourth question which didn't make it in because we realized that BORN didn't have the data we needed. We actually needed to go to something called ICES. ICES, I don't know if you're familiar-*

However, ICES data were disregarded as a potential source for answering the research questions:

Sarah: *[...] ICES is different because I wanted to do a linkage between BORN and ICES, apply for a grant, and explore my question a bit further. What I found difficult with ICES is it wasn't straightforward.*

[...]

Around the time [2014, time 1] I started this. Because when we saw that I was missing a specific variable it was suggested to me to go to ICES. So, that's all I know about ICES [...]

(C4) The idea of collecting particular primary data is not an initial satisficing option

Condition C4 was met in time 1.

Sarah: [...] *But the realities are: collecting primary data is just unrealistic. It's so expensive. It's so time consuming. By the time you start, get to analyze your data, and use it, depending on your study design it can be anywhere from like a year, three years, or five years. [...] it's just such a time commitment. Trying to answer your question using something that exists is much more appealing and a lot simpler. Yeah.*

(C5) An expected scientific contribution exists and the researcher finds its potential rewards satisficing

Condition C5 was not met at time 1. The research, rather surveillance activity was part of the work, and thus what Sarah was doing was public health practice. It started in 2014 as a public health surveillance exercise, although the inception of the reporting on the health problem at hand was in 2011.

Sarah: [...] *I work with the [...] unit. Specifically I look for mainly maternal health issues. [The health issue] wasn't something we typically reported on. We knew that the rates were going up, so I was asked to go ahead and look at this. This was years ago, this was in 2011. I started-*

[...]

Like what's going on with [...]. Once we pulled it out, we started to notice that the trend was increasing. [...] we have no risk factor information, we have no personal information, nor should we, but I can't access [some factors or variables], things that might have a predisposition to developing [the health problem]. This is how this came about.

[...] *Then it was like okay, what else are we missing? This is where the BORN project came up. [...] I was going to look at the sociodemographic profile of [certain type of]*

women and what else is going on. Let's look at [women's characteristics and test results], because this helps really define what's going on with [the health problem]. We had theories, but there was nothing for me to substantiate that.

I: No empirical data.

Sarah: Yeah, exactly.

I: Okay.

Sarah: So then BORN happened. And so I formed this research team because they were interested in [the health problem] [...]

Conditions at a later time of the decision-making process (Time 2)

Event or outcome: Sarah was already reusing and analyzing BORN data when I interviewed her in 2017 (time 1). In January 2020 (time 2), Sarah confirmed me by email that some scientific contributions are planned to be sent for publication in scientific journals in short. Co-authors are currently reviewing the manuscripts.

Changes in conditions: Between time 1 and time2 there was a change in one of the conditions. Condition C5 was met short after Sarah started her research at BORN. She stopped reusing BORN data for the health problem at her work, but continued analyzing BORN data for the health problem at hand as only a BORN agent or independent researcher.

Sarah: That means for all intents and purposes I'm a BORN employee. [...] So, I'm a BORN agent, therefore I have the same permissions and rights as any employee.

The decision-making process

The decision-making process of this case study follows this pattern:

(expected) outcome 2 by an employee at a public health unit → (final) outcome 3 by an independent researcher

When Sarah stopped conducting the *surveillance project* as an employee of a public health unit very shortly after starting it. Yet, she decided continuing it as an independent researcher by keeping the status of BORN agent [procedural rationality]. She was very motivated about the health topic she was

surveilling and found a scientific contribution about it an interesting one (Condition 5 is met), despite not belonging to an academia setting [teleological decision-making theory].

Sarah: [...], and very, very shortly after I started there, I think I was in week three of my one day a week, they pulled me back [bounded rationality]. So, it was like, okay, well we went through all this work and coming up with the variable list, coming up with the project and all of this and to just not do that for me wasn't-, I didn't want to let that go. For one, there was a lot of interest in it, and I just continued with BORN. So I do that on my own time, or I started doing that on my own time with the study team. In a way it was I think better. I say it's better because [...] we don't make recommendations for intervention, and we don't make specific policy recommendations because that's not my role.

[...] Yeah, right? I just can't. I love research and my current work allows me to kind of straddle both roles, which I really enjoy. But on the flipside is when I'm just - I don't know how to frame this. But I started something and I can't not finish something I've started. I just felt to a certain degree not a personal responsibility because I'm not doing anyone any favors necessarily, but to just go through all of this and have a really neat project and not do it to me was like, "Well, that's wasteful." And they gave me an opportunity to do that, so I'm going to take it because I got to go to a conference and I got to present what we've done. And there's a lot of interest in this work because people aren't looking at it. [...] I feel like if I can do that kind of contribution I think it's important for the topic at hand, right? We can explain that.

[...]

But then on the flipside, like I'm also not dumb. If I got publications out of this, this absolutely helps [...]. But it also keeps me current, it keeps my research skills up, and it allows me to do something outside of my day job. So, to just be able to switch that mindset is for me really gratifying. Yeah. So, it's self-serving and also I think important.

Sarah faced some challenges when going through the process of reusing the data as a BORN agent, and she did not know about them until she faced them [bounded rationality]. However, conditions C1, C2, C3, C4 and C5 were met at a very early stage of the reuse process.

6.3. Case studies reusing *proprietary data*

Proprietary data are data that have not been publicly released. Researchers can know that the data exist, but the availability of the data for being reused is uncertain. So, proprietary data cannot be necessarily obtained upon request.

All three case studies under the category of proprietary data are IPD MA (individual patient/participant data meta-analysis). Case study #9 and #10 are a specific case of IPD MA, namely an IPD NMA (individual patient/participant network meta-analysis). The main difference between an IPD MA and an IPD NMA is that the former is limited to only two competing treatments, while the latter includes studies comparing different sets of treatments (Hummel et al., n.d.).

IPD MAs fulfill with the definition of reuse of data in this dissertation, despite reusing data for the same research goal or health problem than the original randomized clinical trials (RCTs) or non-randomized studies (NRSs) which collected the data (primary data). This fact ensures that there is a conceptual or thematic fitness of the secondary data with the research question prior to having the data, although it does not ensure necessarily other types of fitness (e.g., measurement fitness, level-of-analysis fitness, etc.) unless the information regarding other types of fitness are clearly stated in the journal article.

Unlike a meta-analysis (MA), that is carried out from aggregated data found in publications, an IPD MA, and its variants (i.e., IPD NMA) need to access and analyze the individual data that previous empirical studies or CTs have collected in order to be carried out. Most of data sets of RCTs and NRSs are not usually publicly shared, and thus accessing these data is a rather overwhelming and unsurmountable task most of the times, which may end up in unattainable IPD MAs. Sometimes, when it is not possible to obtain all data sets at the individual-record level, IPD MAs and IPD NMA mix in their analysis, when possible, both individual data, and aggregated data.

IPD MA studies are usually carried out by a team of researchers, and usually one of the team members, who is not necessarily the study PI, coordinates all activities related to the retrieval of the data from the original studies. Researchers, who aim to conduct an IPD MA, know in advance which data sets are needed because they develop, or at least should develop, a previous *protocol for a systematic review and individual patient data meta-analysis*, which sometimes follow both a protocol for a systematic review and meta-analysis (with aggregated data (AD)), and the meta-analysis (with AD) itself. In most cases, an IPD MA is led by a researcher or a rather small group of researchers, who invite researchers who collected the primary data to participate in the IPD MA study. This invitation is an international good practice followed in medical disciplines. The primary data collectors,

depending on their level of participation in the study, become authors or their studies are only cited by the IPD MA's publication¹⁷².

The fact that IPD MA studies are carried out by a team of researchers, although there is always at least a PI in the team, it makes it rather difficult to know who decides what and whether all team members agree fully with the decision or action of using secondary data. For the purposes of providing an account of these case studies, I consider the “collective rationality” of the team of researchers a *collective action*, that is, nothing else *than an aggregation of individual decisions* (Townley, 2011, p. 189)¹⁷³. Difficulty aside, and although I only interviewed one participant in each of the three case studies of released data and that I did not check whether the decisions were agreed by all research members, I have tried to give voice to the *actual* decision maker in each of the case studies. I justify the *actual* decision maker at the beginning of each of the case studies' accounts. Yet, to know the *actual* decision maker is not relevant for the purposes of answering the research questions of this dissertation.

6.3.3. Case study #8 (IPD MA)

In this case study, I refer to “the team” or “they” as the actual decision maker, although I only interviewed Nicole Langlois at the Ottawa Health Research Institute (OHRI) from all the research members of the team, and thus evidence of the conditions of the theorized data-reuse mechanism are only from Nicole's verbatim words. There is one main reason why I emphasize “the team” or “they” as the *actual* decision maker. In this case study, there was no unilateral invitation to participate in an already made decision about conducting an IPD MA study as it usually happens in IPD MA studies. The decision of conducting the IPD MA study was made jointly by the primary data collectors of RCTs, although the proposal, and funding application was led by M.A. Rodger, MD (Ottawa Blood Disease Center). There was a proposal from OHRI to the to conduct together the IPD MA, who most of them had already participated in a previous *protocol for a systematic review and individual patient data meta-analysis* for the same health problem.

¹⁷² See how the International Committee of Medical Journal Editors (ICMJE) defines authorship and contributions at <http://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html> [23 January 2020]

¹⁷³ For a conceptual or theoretical discussion on “collective rationality” or group decision-making, see, for instance, Townley (2011) or Kilgour & Eden (2010)

Conditions at the outset of the decision-making process (Time 1)

When I interviewed the participant of this case study, Nicole Langlois, she was a Senior Research Associate in the Clinical Epidemiology Program (CEP) of the Ottawa Hospital Research Institute (OHRI), where she does research in Clinical Trials, Epidemiology and Hematology. She has both a bachelor and an MSc in nursing research, and she has worked in the Thrombosis Research Program since 2001. She has coordinated multicentre studies, international research collaborations, and a multidisciplinary research program focused on thrombophilia. She and the rest of the research team had all the causal powers and liabilities of the theorized data-reuse mechanism. They all were subject to the reward systems of science and followed the epistemic norms of their discipline.

Time 1 (2013) and time 2 (2016) had already passed at the time of the interviews (April 2017). The outcome (#3) was also known at the time of the interviews. The workflow diagram (Figure 27 or annex 20) shows the main steps of the process and how the publication of the scientific contribution was published in *The Lancet* in November 2016.

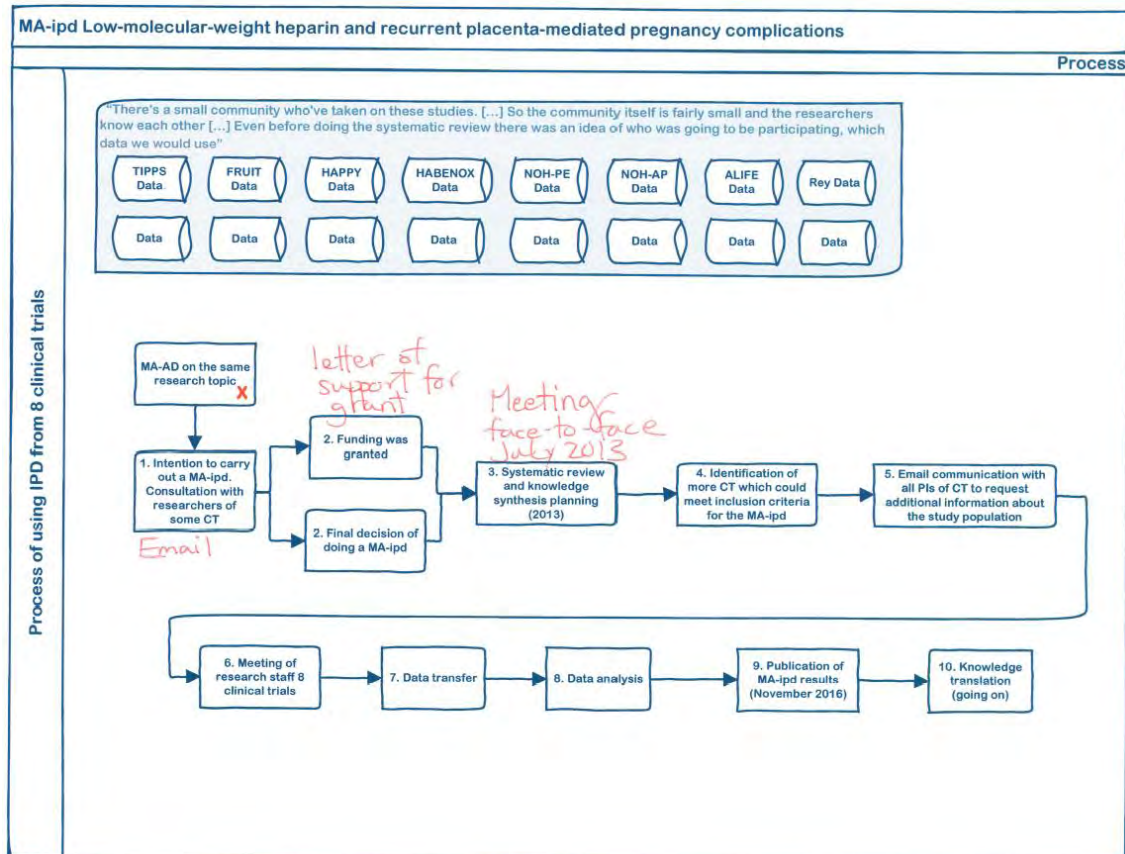


Figure 27 - Workflow diagram of the data reuse process of case study #8

(C1) The researcher knows that secondary data exist

The team knew that the eight data sets of clinical trials that they needed for the IPD MA existed, not only due to the protocol that she co-authored (Rodger et al., 2015), but also due to a previous collaboration of researchers, who had conducted a meta-analysis of aggregated data about the same health problem (Rodger, Carrier, et al., 2014)

Nicole: (minute 05:35) [...] *So, the community itself is fairly small and the researchers know each other. So, when a new trial comes out in the area everybody sort of says, "Oh, yes there's something that's been added." Even before doing the systematic review there was an idea of who was going to be participating, which data we would use.*

I: Okay.

Nicole: (minute 06:25) *And then there had been a meta-analysis. A past collaboration with the same group of authors so they did do a regular meta-analysis. And one of the problems is the studies have composite outcomes and the population is a little bit mixed so there might be women with slightly different characteristics who are enrolled, so the impetus for doing an individual patient data meta-analysis was to be able to look more closely at some of the sub groups of women and to look more closely at some of the outcomes. And so, the meta analysis gave an answer but there more questions that came up so the next step was individual patients. All of these authors, not all, but most had already collaborated, knew each other, had collaborated for the original meta-analysis.*

(C2) Data are obtained

Condition C2 was met shortly after making the decision of reusing the eight RCTs data sets. The decision was made by the whole research team, which included researchers from the eight RCTs, thus they sent the data sets to the study's manager at OHRI, Nicole Langlois.

Nicole: (minute 16) *They [researchers from the eight RCTs] were interested and there was definitely buy in and they were wanting to participate [in carrying out the IPD MA]. [...]*

[...] (minute 17:25) [...] *We had terms of reference that described authorship at the beginning, it described use of the data, and security measures, and those sorts of things. We had contracts that were data transfer agreements so formal agreements before the data were transferred. We provided everybody with a secure USB stick and that's how ... it was an iron key that was sent back and forth so we did it ...*

(C3) Particular secondary data are an initial *satisficing* option

Condition C3 was met as previous publications to the IPD MA show, both the meta-analysis of AD and the protocol for an IPD MA. The team wanted to make a scientific contribution with secondary data (C3-SC).

Rodger, M. A., Carrier, M., Le Gal, G., Martinelli, I., Perna, A., Rey, E., ... Gris, J. C. (2014). Meta-analysis of low-molecular-weight heparin to prevent recurrent placenta-mediated pregnancy complications. *Blood*, 123(6), 822–828. <https://doi.org/10.1182/blood-2013-01-478958>

Rodger, M. A., Langlois, N. J., de Vries, J. I. P., Rey, É., Gris, J. C., Martinelli, I., ... Kaaja, R. (2015). Low-molecular-weight heparin for prevention of placenta-mediated pregnancy complications: Protocol for a systematic review and individual patient data meta-analysis (AFFIRM). *Systematic Reviews*, 3(1), 1–11. <https://doi.org/10.1186/2046-4053-3-69>

For the IPD MA, the team identified sixteen potentially trials, but finally included only eight trials, excluding the other eight for the following reasons:

“wrong population, trial ongoing (EPPI, HEPEPE, HOPPE trials), inability to confirm eligibility of participants, low-molecular-weight heparin intervention stopped too early in pregnancy, and no response from the principal investigator. Additional details about included and excluded studies are in the protocol”. (Rodger et al., 2016, p. 2631)

(C4) The idea of collecting particular primary data is not an initial *satisficing* option

Condition C4 was also met. The team did not think of collecting primary data for studying this health problem.

Nicole: (minute 05:35) *So going back to the original trials, they're trials in a very difficult population, with pregnant woman who've had past complications. So, the population is difficult to study. So there's not a whole lot of research. There's a small community who've taken on these studies. The studies have been quite difficult to complete. They've taken a long time and been hard to recruit for.*

[...]

I think it depends very much on the clinical area. And in this case, some of these studies, as I mentioned how hard they were to do, this one took 13 years to complete. So, already, we're going back. From the earliest participants, we're going back a long time. And the reason it's still a relevant question is there are no other treatments. This question hasn't been answered. There's no alternative. So, there hasn't been a huge amount of progress. Again, we're talking about women who are pregnant, people don't want to do a lot of experimental trials, "Let's just try this or try this." But, so for preeclampsia, for example, there is no effective known treatment.

(C5) An expected scientific contribution exists and the researcher finds its potential rewards *satisficing*

Condition C5 was met.

I: [...] *So, you know, at that point when they were trying to figure out if we are going through this individual patient data meta-analysis or not, were they thinking of the journal already?*

Nicole: *Not necessarily the specific journal. They wanted to go for a high impact journal, because they thought it would be practice-changing and an important finding because it was such a big collaboration. And it was a question that hadn't been definitively answered.*

Conditions at a later time of the decision-making process (Time 2)

Event or outcome: The use of secondary data happened as evidence of scientific claims (Rodger et al., 2016), and except for some small issues, everything happened as originally planned.

Changes in conditions: There were no changes in the conditions C1, C3, C4, and C5 in this IPD MA study. Nicole's and the rest of the research team's structures and causal powers and liabilities did not change along the process of reusing individual data from the eight RCTs from time 1 to time 2. However, if we take into account that the research team aimed to include sixteen RCTs, then condition C2 was not met at time 2.

The decision-making process

The decision-making process of this case study follows this pattern:

(expected) outcome 3 → (final) outcome 3

In this case study, conditions C1, C2, C3-SC, C4, and C5 are met nearly simultaneously. The time horizon of the data-reuse mechanism of this IPD-MA is long (3 years) but there is nearly no uncertainty about any of the value of the conditions, since they are met at the outset of the process. However, there were challenges along the process of reusing the data, but the team always found the way to overcome them.

6.3.4. Case study #9 (IPD NMA)

In this case study, I will refer to my participant, Claire Johnson, as the actual decision maker, rather study leader, of the process of reusing data, although the IPD NMA was conducted by a whole team of researchers. However, the actual decision maker of conducting an IPD NMA study on mass deworming of children in developing countries is the Bill & Melinda Gates Foundation (BMGF)¹⁷⁴. The foundation decided to fund a two-year IPD NMA study in order to find out if the effects of mass deworming on child welfare outcome (e.g. growth, attendance in school, attention span in school) vary

¹⁷⁴ <https://www.gatesfoundation.org/> [24 January 2020]

across important characteristics that could help with better targeting and delivery of deworming, e.g. socioeconomic position, gender, nutritional status. SickKids¹⁷⁵ submitted a proposal, which was peer-reviewed and finally accepted by the BMGF in November 2015. The proposal also included a NMA-IPD of deworming for pregnant women and an analysis of geospatial data. SickKids, in turn, sub-contracted my interviewee's team to actually coordinate the data retrieval and conduct the analysis. Thus, although Claire made decisions about the process of retrieving and analyzing the data from the trials, Claire and her team or SickKids had no authority to come to a decision of not finishing the study. Ultimately, this decision could be only be taken by the BMGF. However, the foundation had no role in the methods, but were invited to comment on draft and interim results at two points.

When I interviewed Claire twice in May 2017, time 2 of the decision making process had not happened yet, and neither the outcome was known. This is the main reason I have followed up with her about this IPD NMA on mass deworming for more than two years.

¹⁷⁵ <http://www.sickkids.ca/> [24 January 2020]

Conditions at the outset of the decision-making process (Time 1)

At time 1, Claire Johnson, PhD belonged to two structures, a research one and a teaching one. With no doubt, she and her team had all the causal powers and liabilities to conduct the IPD NMA. In fact, she and some members of the IPD NMA had previously conducted an aggregated data meta-analysis (AD NMA) on the same health issue, which got published in 2017 (Welch et al., 2017). Claire knows the epistemic norms of her discipline, as well as the co-authoring norms by the International Committee of Medical Journal Editors (ICMJE), and the norms of the reward system of science.

(C1) The researcher knows that secondary data exist

Condition C1 was met even before of being sub-contracted by SickKids with the BMGF's funding. Claire had been long interested in these kinds of health issues, and she had already conducted an AD NMA about this topic:

Welch, V. A., Ghogomu, E., Hossain, A., Awasthi, S., Bhutta, Z. A., Cumberbatch, C., ... Wells, G. A. (2017). Mass deworming to improve developmental health and wellbeing of children in low-income and middle-income countries: a systematic review and network meta-analysis. *The Lancet Global Health*, 5(1), e40–e50. [https://doi.org/10.1016/S2214-109X\(16\)30242-X](https://doi.org/10.1016/S2214-109X(16)30242-X)

However, in order to conduct the IPD NMA, her team had to review and update their search strategy for finding studies because in the AD NMA they did not include studies focused on one type of worm that was of interest to the BMGF proposal. From the 280 primary trials that they found to assess effects of deworming, they appraised only those, which had some measure on the intensity of the infection. Thus, they narrowed the list to 67 studies. When I interviewed Claire for the first time in May 17, 2017, her team only had 10 responses from the 67 primary trials that they contacted.

(C2) Data are obtained

Condition C2 was met. However, only 19 RCTs of individual participant data out of finally 41 eligible ones were accessed. Also, data from the AD NMA including 29 RCTs were included in the study at hand.

Individual participant data (IPD) was obtained from 19 out of 41 eligible randomised trials. These 19 trials included 31,945 participants and had an overall low risk of bias. A secondary analysis added new data to the meta-analysis of STH deworming versus placebo of a previous Campbell review by the same authors. This analysis included 29 randomised trials, with data from two studies which had not published weight gain data and updated effect estimates from three studies based on the data provided by authors (Welch et al., 2019, p. 3)

However, the process of obtaining the data was tedious and cumbersome, and thus required lot of effort and time, as the validated-by-participant workflow diagram shows (Figure 28 or annex 21).

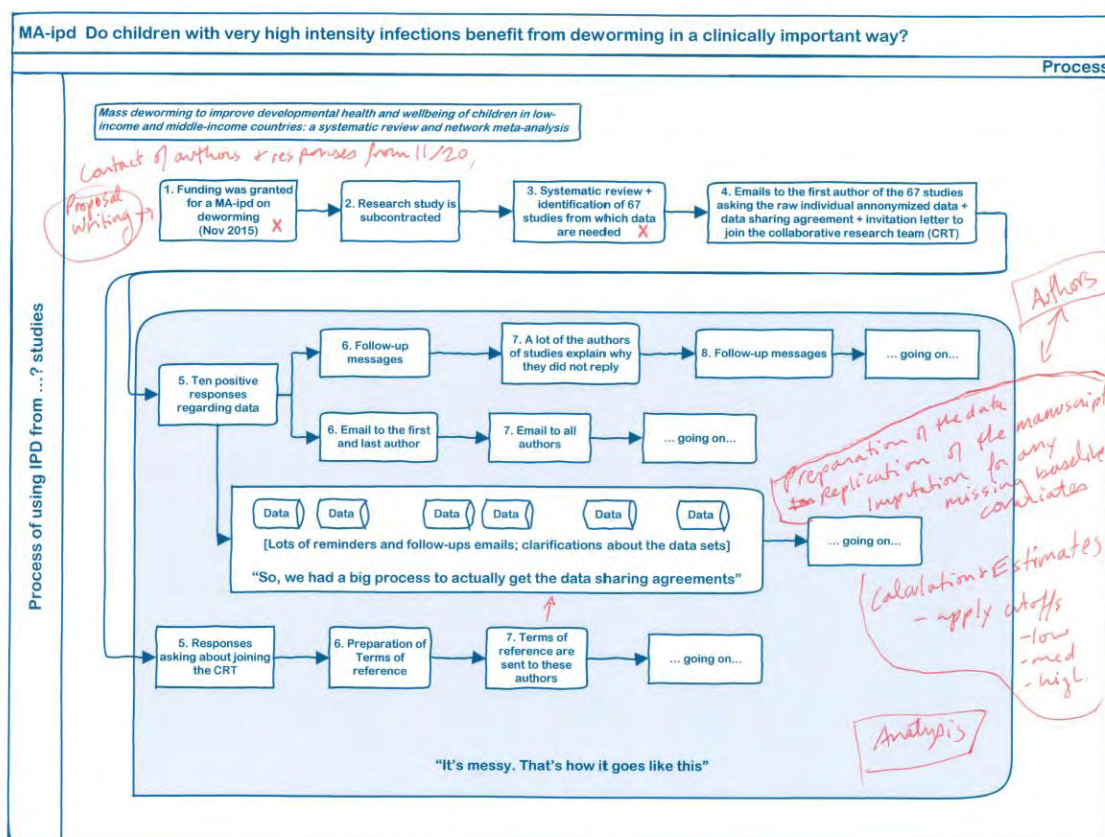


Figure 28 - Workflow diagram of the data reuse process of case study #9

(C3) Particular secondary data are an initial *satisficing* option

Condition C3 was met for a scientific contribution (C3-SC). The research team made sure that the studies that they needed were *satisficing*, at least at time 1. However, neither Claire nor her team were aware of the challenges that they were going to face when accessing or retrieving the data from primary trial researchers [bounded rationality].

Claire: (minute 13:04) *So, we weren't getting responses from the generic email that my team was sending, I started to send personal emails from my email and the authors told me that they didn't know what to do with this very technical form. So, we had a big process to actually get the data sharing agreements. One person had to change the data sharing agreement because her institution required a clause about mediation. One author had their own data sharing agreement, so we used their data sharing agreement, because they've shared their data with other teams.*

What else did we do? And then, because we only got ten out of 67, then we wrote to the last author and the first author of every team, and then we wrote to all the authors of every team so we had to find the email of every author. And then for those that we didn't find, we phoned the institutions to ask, "Do you know what happened to this author?" [procedural rationality]

And then we have studies going back to 1975 and so some of them have moved on (she laughs). So, one of them had retired, one of them gone with no forwarding address, and there are still 10 that we haven't identified a contact.

[...]

So, once we do get the data, I open the files right away to check ...

[...]

So, I think fitting in the phone calls because life is busy, but that what ... with people who said yes, I've phoned to say, "You said yes, but ..."

These people aren't usually in their office, right? So, that has been difficult to fit in, but that's just part of the process of following through [satisficing]

[...]

It's messy. That's how it goes like this. [satisficing]

[And when I asked Claire about ethic protocols, she replied:]

They haven't said they had to go through their ethics, but some have gone to their contract officer, like the one that changed the consent for ... the data sharing agreement, and had to go to their contract dean at the institute, because sometimes the institutes own the data.

(minute 16:04) *Everything is, we asked for proof of ethical review of the studies and for studies conducted before 1990, the ethical review process was not the same as now, so they don't have an ethics certificate number, they don't have their consent*

forms, and so we had ... we asked for an addendum to our ethics and they granted it, that authors could simply write us a letter to explain what ethical procedures were followed and why they can't provide it.

[Regarding amount of data sets from trials, analysis, Claire knew perfectly the limitations of using secondary data]

Claire: *And what we will do is compare what, we'll compare our results with the studies that we receive, we'll compare with the studies that we haven't received to see if they're different in some systematic way that would affect the analysis. But we're optimistic that we will have enough [bounded rationality. She does not really know]*

We won't have a full ... we definitely won't have all of them. And we won't have a lot from before 2000, so it'll be a limitation and ...

(C4) The idea of collecting particular primary data is not an initial *satisficing* option

Condition C4 was met although I have no evidence from the participant's own words. They finally got nearly 32,000 participants, although initially expected to have 40,525 from 41 studies, thus it is inconceivable, and not only *not satisficing*, that the team of researchers led by Claire could collect primary data on short-or-long-term effects of deworming in 40,525 participants in developing countries within the two years required by the BMGF.



FIGURE 3 Yield of STH studies and participants. STH, soil-transmitted helminthiasis

Figure 29 - Number of eligible studies and participants. Case study #9. Source: (Welch et al., 2019, p. 12)

(C5) An expected scientific contribution exists and the researcher finds its potential rewards *satisficing*

We can deduct from the funding initiative by the BMGF that the foundation wanted an answer to the health problem they addressed, whether in a scientific journal or in any other kind of publication. Had the foundation foreseen no benefits (rewards) from the answer to the research question, they would not have offered any funding. For Claire, and her research team, the potential reward of a scientific contribution was also *satisficing* since they are subject to the reward system of science. However, the potential rewards, for instance, citations and prestige, were not the only potential rewards for Claire. She had other both personal and professional interests in the health issue at hand, and thus in conducting the study. When I asked her about the benefits that she perceived from this IPD NMA study, she replied:

Claire: *It's time-intensive.*

I: *Exactly. It's time-intensive. It's going to be or it has been exhausting. [...] Now, tell me why a researcher gets into this issue that is complicated and into this project.*

Claire: *Huh. Well, I think the funding was obviously one reason. That was the initial contact, actually, is that the funding was likely available from the Gates Foundation to do this work. And it addresses a gap that we identified in our published meta-analysis that there's really a huge uncertainty about whether subgroups of children could benefit more.*

I: *Could benefit? Mm-hmm.*

Claire: *Yeah. Benefit in an important way more. And so I guess there's the funding. There's the answering unanswered questions, which is what scientists like to do. I think also an opportunity to work with some really interesting people. You know, the advisory board is quite incredible, and honestly the authors have been quite incredible. Like nice to work with.*

I: *The authors? You mean the people sharing the data with you?*

Claire: *Yes. Yeah, yeah. They helped share the data. I think for the institute this is an important publication. It will be for sure a high impact journal that would like to see this published, so that's good career-wise for me for my institute to support my career*

and I think it's also been an opportunity to bring people on to the team, which is always nice. [...] So, co-op students and one PhD student is helping now with the main publication. So, in that sense, that's good for me because I like to support students. I like to have funding for them and something for them to do.

I: *So, you are paying those students?*

Claire: *Yeah. They're all being paid. Yeah. I think those are all the benefits.*

Conditions at a later time of the decision-making process (Time 2)

Event or outcome: The IPD NMA publication on mass deworming came out in 2019 (Welch et al., 2019), so secondary data have been finally reused for a scientific contribution as initially planned in time 1.

Welch, V. A., Ghogomu, E., Hossain, A., Riddle, A., Gaffey, M., Arora, P., ... Wells, G. (2019). Mass deworming for improving health and cognition of children in endemic helminth areas: A systematic review and individual participant data network meta-analysis. *Campbell Systematic Reviews*, 15(4). <https://doi.org/10.1002/cl2.1058>

Changes in conditions: No changes in the conditions C1, C3, C4, and C5 happened during the process. The only change happened with condition C2, which was not met because not all eligible data could be obtained.

The decision-making process

The decision-making process of this case study follows this pattern:

(expected) outcome 3 → (final) outcome 3

As I have suggested hereinbefore, the decision of not achieving outcome 3 depended exclusively on the BMGF or in the fact that enough individual participant data could be obtained.

At time 1, conditions C1, C3, C4, and C5 were met. However, the value of condition C2 (data are obtained) was not known until the very end just before Claire and the research team made the analysis.

So, at time 2, no researchers' causal powers and liabilities had changed, but condition C2 was not met. Yet, the research team decided that they had enough data to make a scientific contribution (outcome 3). Claire and her team were determined to achieve the goal [teleological decision-making theory], and all their actions and/or decisions were managed [procedural rationality] in such a way to achieve the scientific contribution even knowing, rather sensing, that they would not be able to obtain individual participant data from all eligible studies.

Claire: *We won't have a full ... we definitely won't have all of them. And we won't have a lot from before 2000, so it'll be a limitation and ...*

[...]

(minute 34:23) *The only way ... so what we'll do to try to mitigate that limitation is to compare the studies we receive with those that we don't, and that'll be the best that we can do.*

We still think it'll be worth publishing and I think it'll still make some interesting conclusions.

Had not they obtained enough individual data, they would have still published a contribution with no answer to the initial research question asked by the BMGF about the actual effects of deworming, but on the issue of why they would have not been able to conduct the IPD NMA as others have done, i.e., Jaspers & Degraeuwe, 2014.

Claire: *Yep. I think we felt that if we -- so, now, we have 17 data sets and I think we have all except for 3 that have been published after the year 2000. So, I think it will be published. We had thought that if we had only five we probably would still do the analysis and probably publish it as here's what we tried to do, but we only got 5 studies and this is how they are in relation to the other 62 studies. So what are the characteristics? Especially if we had had only a few participants, I think because some of these studies are very large, like 2,000 children. So, if we had only been able to get data sets from smaller studies - we'll say 100 to 200 children - I think the publication would be more a publication about how it wasn't possible to get enough studies.*

But would we publish the findings of those small studies? I'm not sure. I think maybe we would have not have. But I think at this point we have enough data that it will be interesting to publish; even interesting to compare with those that we didn't receive the data sets. So...

The time horizon of the data-reuse mechanism in this case study is of 2 years, as it was a requirement by the funding institution, although the scientific results have seen the light later on due to reasons related to both scientific publishing and scientific incentives.

6.3.5. Case study #10 (IPD NMA)

This case study refers to an IPD NMA on Alzheimer's dementia, in which there is a team of researchers leading the study. In this study, my participant, Areti Angeliki Veroniki, is the *actual* decision maker of the process of reusing data, so I refer to her most of the times, if not all, although she was not the only one conducting the IPD study.

When I first interviewed Areti Angeliki Veroniki (from now on *Argie*) in September 2017, the IPD NMA study was at a very early inception (time 1), so due to this reason and to my participant's other commitments, the second interview happened fifteen months later (9 January 2019). However, time 2 of the decision making process had not happened in our second interview (January 2019), and neither the outcome was known. So, I had followed up with Argie about this IPD NMA on Alzheimer's dementia until September 2019, and again in January 2020 for sharing my findings about her case study and to ask about the actual outcome of the IPD NMA study.

Conditions at the outset of the decision-making process (Time 1)

Argie had all causal powers and liabilities that I have theorized in the data-reuse mechanism at both times 1 and 2. She is a mathematician, holds an MSc in Statistics and Operations Research, and a PhD in Epidemiology¹⁷⁶. At time 1, Argie was a post-doctoral fellow in Knowledge Synthesis with a Banting post-doctoral fellowship (2015-2017) at the Knowledge Translation Program of the Li Ka Shing Knowledge Institute of St. Michael's Hospital (Toronto). She was also a co-Convenor of the Cochrane Statistical Methods Group. She is a very experienced and motivated researcher and knows the limitations of working with secondary data, despite being the leader of an IPD NMA study for the first time. She knows the epistemic practices and norms of her discipline, and is subject to the reward system of science.

¹⁷⁶ Source: <https://esm.uoi.gr/en/argie-angeliki-veroniki/> [5 February 2020]

(C1) The researcher knows that secondary data exist

Argie knew for sure that the secondary data she needed for the IPD NMA existed. She developed the study protocol (Veroniki, Straus, Ashoor, Hamid, et al., 2016), from which she identified 108 potentially eligible RCTs. She knew quite a lot on the topic and the types and number of RCTs on the topic from previous studies in which she and other of her team members were involved, for instance, see Tricco et al. (2018, 2012). See Figure 30 or annex 22 for the workflow diagram of her process of reusing data, which she validated. She added no details or corrections.

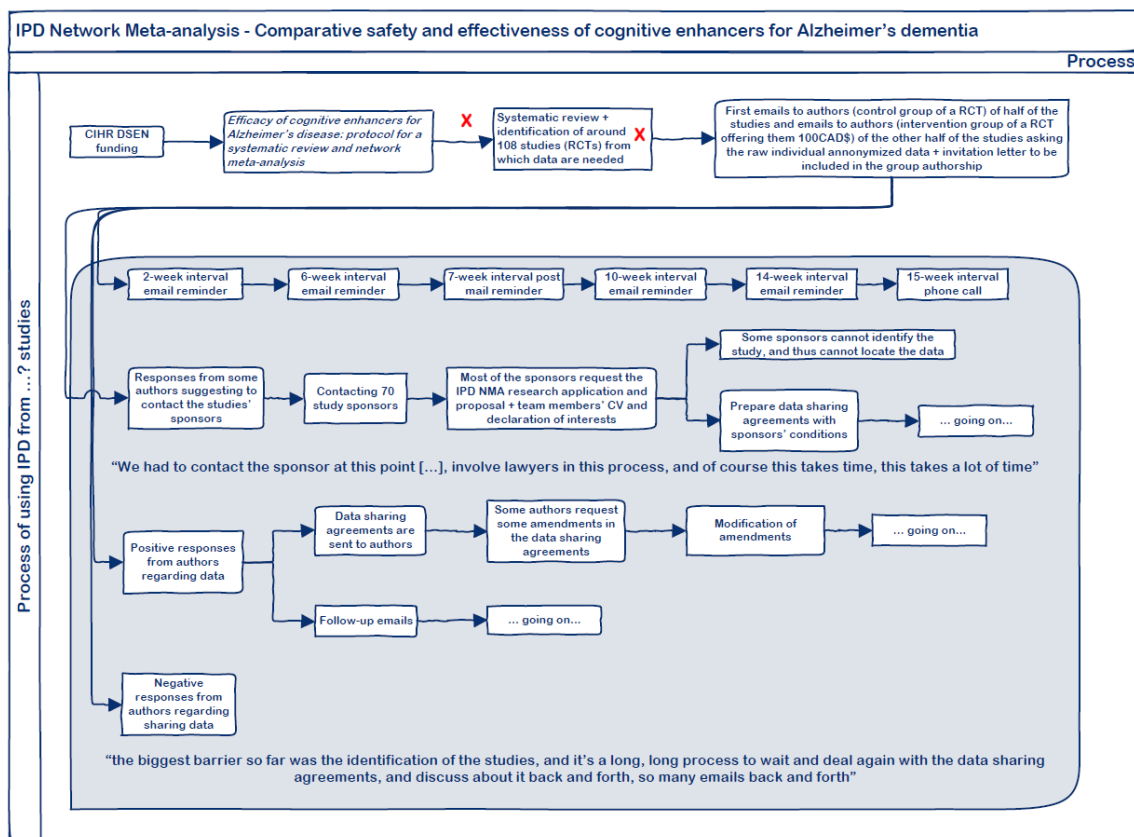


Figure 30 - Workflow diagram of the data reuse process of case study #10

(C2) Data are obtained

At time 1, data from the eligible RCTs were not obtained since Argie was still requesting the data to most of the primary data collectors.

(C3) Particular secondary data are an initial *satisficing* option

Condition C3 was met. However, Argie was not aware of all of the different challenges she was going to face when accessing, retrieving and analyzing the data, especially regarding the data from RCTs, which were sponsored by pharmaceutical companies.

(C4) The idea of collecting particular primary data is not an initial *satisficing* option

Condition C4 was met. When I asked her about collecting her own primary data for answering the research question, she answered:

Argie: So, instead of... you mean instead of conducting a network meta analysis why I don't conduct well a clinical trial, a randomized controlled clinical trial?

I: Yes, right

Argie: (minute 31:43) Well, the issue is first of all, with the meta-analysis and network meta-analysis that we are conducting now we have the opportunity, uh, to collect all the previous data that have probably suggested different things and make sure why [she emphasized "why"] they suggested different things, and uh, search for this more, and then we have the opportunity to include in the network meta-analysis in the statistical analysis all the different treatments that have been suggested to date. So, in a clinical trial it would be very expensive and probably sometimes is not ethical to compare all the different, uh, treatments by themselves. So, let's say that there are studies that compare placebo to specific treatments, but in that case, no. Let's say that we have the comparison of the treatment 1 versus treatment 2, right? So, in that case we wouldn't know if treatment one is better than placebo because sometimes is unethical to conduct those studies. In a network meta-analysis we are able to infer for, uh, treatment comparisons that were never conducted before. And, we would have to spend more money to conduct those clinical trials, and sometimes it would be unethical to compare, to include patients in specific treatments. But the thing is that with the clinical trial we also have uh, we are usually able to include a certain number of patients. This is the third point that I can raise. So, the first one is the ethical approach, the other one is that we increase the number of patients that we include in our analysis. So, let's say that we include ten trials of a hundred people. Then we would be able to include 1000 patients in our analysis. However, to include 1000 patients in a clinical trial would be probably something not very easy to do, and

the third one would be the money, that we would have to spend a lot of money. That's why I believe that is better to conduct at this point a network meta analysis, and as I said, uh, we are able to compare all the treatments that have been suggested so far, that are available on the market. Where in a clinical trial we are limited on this. We usually compare two, three maximum four treatments I believe, unless of course we conduct an observational study, but in a RCT it is difficult to conduct to compare treatments more than four, I believe. At least this is what I have seen so far.

(C5) An expected scientific contribution exists and the researcher finds its potential rewards satisfying

Condition C5 was met. In fact, Argie did not view the IPD NMA only as a contribution to clinical epidemiology in mental health, but also as a contribution to methods research communities.

I: *Where would you put it? Would you put it in clinical epidemiology? In mental health? You know...*

Argie: *Yeah, I think mental health for Alzheimer's dementia, at least. But also the important also thing is the analysis, at least to me. So, I think the specific models that I'm going to apply are of interest and, apart from the clinical-*

I: *Research findings.*

Argie: *Clinicians will be interested. I think the mathematicians, the statisticians will be interested in the models that we are going to apply and we are going to provide the models in appendix. So hopefully this will be food for more research, if I can say that. Food for thought.*

Conditions at a later time of the decision-making process (Time 2)

Event or outcome: The outcome of this case study at time 2 (February 2020) is, rather will be, a published scientific contribution. Argie confirmed me, in her email message dated 3 February 2020, that she and her team will submit their manuscripts for publication:

Argie (in an email message): *As an update, I am happy to let you know that we are currently conducting our network meta-analysis models (2nd stage of our analysis), and that we plan to submit our findings for publication soon. That means that all the IPD analyses have been finalized - well, only through the sponsors' platforms, which restricted us from combining all IPD studies in a more advanced network meta-*

analysis model. You may also take a look at our published study on the barriers we identified from retrieving the IPD in Veroniki et al. (2019). Of course, there are more specific issues we identified in the Alzheimer's research study, and we will add them in our final publication.

Changes in conditions: Keeping in mind Argie's email message, it seems that there has not been any change in any of the four conditions C1, C3, C4, C5, or in any of Argie's and of the research team members' causal powers and liabilities. At time 2, Argie's structure¹⁷⁷ is different from her structure at time 1, but she still belongs to several research institutions and to a research discipline and is subject to the reward system of science.

Regarding condition C2 at time 2, I have not been able to find out whether Argie and her team have been able to retrieve or access the individual participant data from the 108 studies that were identified in the study protocol (Veroniki, Straus, Ashoor, Hamid, et al., 2016) as potentially eligible for their IPD NMA study. However, from her email message, it is inferable that they could not retrieve all IPD, but only access to them through the sponsors' own platforms, which most of them were pharmaceuticals.

Decision-making process

The decision-making process of this case study follows this pattern:

(expected) outcome 3 → (final) outcome 3

The reuse of individual participant data has happened for a scientific contribution (outcome 3), if it is accepted and published by a scientific journal, which is very probable. This contribution is possible even if she has not been able to combine all IPD studies in a more advanced network meta-analysis model [satisficing]. Actually, Argie and her team have already made two contributions, i.e., Veroniki et al. (2019), and Veroniki, Straus, Ashoor, Stewart, et al. (2016), from her experience retrieving or accessing the data for this IPD NMA on Alzheimer's dementia and from other IPD NMA on type 1 diabetes that she is also leading.

¹⁷⁷ She is a Research Fellow at the University of Ioannina in Greece, a Research Associate Statistician at the Imperial College in London, UK, and an Affiliate Scientist at St. Michael's Hospital in Toronto, Canada. Her research focuses on the statistical modelling for evidence synthesis and the methodology of systematic reviews. She is a co-Convenor of the Cochrane Statistical Methods Group and an Associate Editor for the BMC Systematic Reviews journal and the BMC Pilot and Feasibility Studies journal. Source: <https://esm.uoi.gr/en/argie-angeliki-veroniki/> [5 February 2020]

So far, the time horizon of this case study has lasted nearly three years. All conditions, except condition C2 (data are obtained), were met at time 1. When I interviewed Argie in January 2019 she was still requesting data from some RCTs' sponsors and authors. While she was requesting the data and waiting for some pharmaceuticals' or other sponsors' responses, she did not know whether she was going to retrieve or access the data [bounded rationality], but she kept insisting on getting the data and continuing the study [procedural rationality]. Every time I interacted with Argie, she was determined to achieve her goal [teleological decision-making theory]. In fact, in my last follow up email with her in September 2019, she confirmed me that the IPD NMA was currently under conduction and that they were trying to finish it. However, sometimes, it was not easy for her to make a decision when she was encountering so many challenges in obtaining the data. In fact, when I asked her once whether she had ever thought of giving up the study, she answered:

Argie: *Of course I've thought about it. I was desperate at some point. I said, "Oh my god." Especially for the Type I diabetes that we needed that license. I said, "Yeah, it's not worth it." Something that I have to give up. But then, I said, "one last try" to ask if they would give us the data without the license. But then I was thinking they will not. They will not give up the license, so what can I do? The systematic review is already outdated. So, we have to rerun everything from the beginning. I was thinking, "Is this helpful for someone? The results will be helpful for someone or not?" So, yeah, I thought about it. Now, I have a hope again that they are going to give us the data. Fingers crossed. So we probably won't hear back from them this month. We will contact them again, yeah.*

Her decisions, though, were based on *satisficing* options.

I: *Yeah. Okay, so their suggestion is just to publish and to get the results whatever they are, depending on what you have? Okay.*

Argie: *That is what I am suggesting. So some of my collaborators agree. Some of them say, "Yeah, but why don't you wait? Why don't you press them more?" Briefly, I've been waiting for more than two years, and we have been pressing them. [...]. They're not going to give us any more data so this is what they are going to share with us. So, this is what we have. If they give us more data in the future that would be more welcome. We'd probably conduct the analysis from scratch, but, for now, we have to wrap up.*

6.4. Summary of analysis of the ten case studies

I have tested the theorized data-reuse mechanism as a causal explanation of the use of secondary data in ten case studies in health sciences. All case studies have served to test the data-reuse mechanism except case study #6, where although X (researcher's expected goal or scientific contribution) is known, the final outcome is unknown (Y). In order to test a mechanism, both X and Y have to be known (Beach & Pedersen, 2013, p. 147). See Table 6.

Table 6 - Summary of outcome and conditions of findings in case study #6

	Expected and final outcomes	Changes in conditions or in the data's or researcher's causal powers and liabilities between time 1 and time 2
Case #6	(expected) outcome 3 with RQ (a) → (expected) outcome 3 with RQ (b) → (final) unknown	C2

6.4.1. When condition C4 of the data-reuse mechanism is met

Condition C4 (the idea of collecting particular primary data is not an initial satisficing option) is met in case studies #1, #4, #5, #7, #8, #9, and #10 at both time 1 and time 2 of the data-reuse process.

Regarding the other four theorized conditions of the data-reuse mechanism, they are met at both time 1 and time 2 of the data-reuse process in only two case studies, #4 and #5. In both these cases, the expected outcome 3 of making a scientific contribution with only secondary data is achieved. See summary of these two case studies in Table 7.

Table 7 - Summary of outcomes and conditions of case studies #4 and #5

	Expected and final outcomes	Changes in conditions or in the data's or researcher's causal powers and liabilities between time 1 and time 2
Case #4	(expected) outcome 3 → (final) outcome 3	No changes
Case #5	(expected) outcome 3 → (final) outcome 3	No changes

In case studies #8, #9, and #10 changes occur in condition C2 between time 1 and time 2. Although condition C2 is met only partially at time 2, the final or actual outcome 3 is equal to the initial expected

outcome 3 (condition C5), and thus, a scientific contribution is made with only secondary data. See summary in Table 8.

Table 8 - Summary of outcome and conditions of case studies #8, #9 and #10

	Expected and final outcomes	Changes in conditions or in the data's or researcher's causal powers and liabilities between time 1 and time 2
Case #8	(expected) outcome 3 → (final) outcome 3	C2
Case #9	(expected) outcome 3 → (final) outcome 3	C2
Case #10	(expected) outcome 3 → (final) outcome 3	C2

An explanation of the data-reuse mechanism with final outcome 3 (a scientific contribution is made with only secondary data) and with condition C2 being partially met lies in the fact that these two case studies needed many data sets from different data owners, and at least some of these data sets were obtained. Although researchers in these case studies have not obtained all possible data sets, researchers find the data sets obtained *satisficing* for making their scientific contribution even to the point of sacrificing some of the responses, findings or conclusions regarding their initial research questions. In these case studies, I suggest that the fact of making a scientific contribution (condition C5) was so strongly pursued, that researchers made all necessary efforts in order to make the contribution, although without reaching the point of *unacceptable compromises* (Clarke & Cossette, 2000, p. 111). In their processes of reusing the data, researchers judged what strategies were best in order to make a scientific contribution despite the fact they had not obtained all desired data sets or, having obtained them, the data were not in the format they needed.

From the review of the literature, we know that researchers can adopt once of these two¹⁷⁸ strategies: adjust or tweak the research question to fit data (Doolan & Froelicher, 2009), or acknowledge limitations in their findings and conclusions of their publications¹⁷⁹. Researchers in case studies #8, #9, and #10 cannot adjust or tweak their research question(s) to fit the data, since they have developed and published a study protocol. Both epistemic practices and good research practices in their disciplines require them to follow this protocol. Thus, the strategy that they adopt is to be transparent with all the challenges they found in obtaining and analyzing the data, acknowledge limitations in their findings, and highlight differences with their original study protocol.

¹⁷⁸ Actually, there are five, but two of the strategies consist of giving up the reuse of the secondary data, and the third one consists on formulating the research question after having looked at the data. These three options are not possible for these case studies.

¹⁷⁹ Garmon Bibb (2007)

Although with the same outcome 3, there is an important difference between case studies #4 and #8 and case studies #5, #9 and #10. In case study #4, access to data and thus to information about all types of potential fitness needed between the research question and the dataset happen at the outset of the process of reusing data (time 1). A similar situation with regard to data access and fitness happens with case study #8 because there was an initial agreement among all the primary data collectors and users to share data with the leading institution (OHRI) of IPD MA.

However, in case study #5 there was a lot of uncertainty about how access could be granted and in which conditions, although Deshayne Fell was very familiar with BORN data and ICES data and what type of analysis she could do with them. She was confident that she was going to have access to these data, but she never knew when, until it actually happened. In case studies #9 and #10, the level of uncertainty was very high regarding access to data and some kinds of fitness throughout the whole process of reusing data. However, the PI and the rest of the team persevered until the scientific contribution was achieved, reaching the *satisficing* point of accommodating some issues and of renouncing some findings and conclusions in their studies.

Case study #7 also confirms that the data-reuse mechanism works empirically with secondary data being used as the only evidence of scientific claims (outcome 3). In fact, it confirms that the appearance of condition C5, and the fact that condition C3-BK (secondary data are *satisficing* for the creation of background knowledge) changing to C3-SC (secondary data are *satisficing* for making a scientific contribution) is what activates the mechanism. In this case study, the reuser's structure also changes between time 1 and time 2. Sarah Wilson started to perceive a scientific contribution's potential rewards to be *satisficing* as soon as she belonged to a research structure. However, in case study #7, data reuse could have also happened while Sarah Wilson was an employee at a public health unit, although data had not been used as evidence of scientific claims, but as evidence for surveillance purposes. See Table 9 for a summary of outcome and changes in conditions of case study #7.

Table 9 - Summary of outcome and conditions of case study #7

	Expected and final outcomes	Changes in conditions or in the data's or researcher's causal powers and liabilities between time 1 and time 2
Case #7	(expected) outcome 2 by an employee at a public health unit → (final) outcome 3 by an independent researcher	C3-SC, C5, and reuser's structure

In case study #1, the five conditions of the data-reuse mechanism are met in time 1. However, condition C5 disappears along the process of reusing data as shown in Table 10, and thus, the final outcome in time 2 is not a scientific contribution (outcome 3) as expected and planned in time 1, but final outcome 2 (the use of secondary data serves for the creation of background knowledge). The

final outcome of this case study #1 could have been outcome 1 (use of secondary data does not happen at all after having tried or considered the option) if condition C5 had disappeared before the researcher reuses the data. However, since condition C5 disappears after reusing the data, the outcome is 2.

Table 10 - Summary of outcome and conditions of case study #1

	Expected and final outcomes	Changes in conditions or in the data's or researcher's causal powers and liabilities between time 1 and time 2
Case #1	(expected) outcome 3 (final) outcome 2	C5

6.4.2. When condition C4 of the data-reuse mechanism is not met

In case studies #2 and #3, all conditions except condition C4 of the data-reuse mechanism are met in time 1. Condition C4 is not met at time 1 because the researcher finds the idea of collecting primary data *satisficing* for making a scientific contribution. The researcher's initial intention with secondary data was their reuse for creating background knowledge (outcome b).

However, as summarized in Table 11, only in case study #3, the final outcome b in time 2 is the same as the expected outcome b in time 1¹⁸⁰, and thus, secondary data are used for the creation of background knowledge and primary data are used as evidence of scientific claims.

In case study #2, the researcher wanted initially to use secondary data for the creation of background knowledge (outcome b) to make better decisions with primary data at the lab. However, later on along the process, he changes his mind and decides using secondary data to support his scientific claims done with primary data in his publication (outcome c).

¹⁸⁰ Although there is an intermediate decision of not reusing TCGA data because after gathering some information about the data, they are not *satisficing* for David.

Table 11 - Summary of outcome and conditions of case studies #2 and #3

	Expected and final outcomes	Changes in conditions or in the data's or researcher's causal powers and liabilities between time 1 and time 2
Case #2	(expected) outcome b → (final) outcome c	C3, C5 C4 was unmet in time 1
Case #3	(expected) outcome b → (final at that time) outcome a → (final) outcome b	C2 (this condition had different values along the process) C4 was unmet in time 1

Table 12 on this page summarizes the analysis of the ten case studies in orange color in time 1 and time 2. The grey color represents unknown or uncertain values. The red color represents unmet conditions of the theorized data-reuse mechanism at time 1. The blue color represents new values of conditions at time 2, due to changes during the data reuse process, which may have affected the outcome.

Table 12 - Summary of analysis of the ten case studies

		Time 1							Time 2							
	Case study	Researcher's CP&L	Researcher's structure	C1	C2	C3	C4	C5	Researcher's CP&L	Researcher's structure	Event	C1	C2	C3	C4	C5
RELEASED DATA	#1					C3-SC		Outcome 3			Outcome 2					
	#2					C3-BK		Outcome b			Outcome c			C3-SC		
	#3					C3-BK		Outcome b			Outcome b					
	#4					C3-SC		Outcome 3			Outcome 3					
STEWARDED DATA	#5					C3-SC		Outcome 3			Outcome 3					
	#6					C3-SC		Outcome 3			Unknown					
	#7					C3-BK		Outcome 2			Outcome 3			C3-SC		
PROPRIETARY DATA	#8					C3-SC		Outcome 3			Outcome 3					
	#9					C3-SC		Outcome 3			Outcome 3					
	#10					C3-SC		Outcome 3			Outcome 3					

6.5. How does the data-reuse mechanism work?

In spite of the proposed data-reuse mechanism and the results of the analysis of the ten case studies, we are still in the dark of how the causal forces of the mechanism link the hypothesized cause(s) X with the outcome (Y) of secondary data being used as evidence of scientific claims (outcome 3 and outcome c). We do not know either how the causal forces are linked in such a way that there are different outcomes, namely 1, 2, a, and b.

I have suggested above that Sayer's structure of a causal mechanism (2000, 2010) is an appropriate conceptual tool in order to identify entities capable of doing things, and the conditions under which these capabilities lead to the studied final outcome (Y). Yet, it needs to be complemented with process-tracing methods. The reason is that, while Sayer's structure of a causal mechanism lets us identify the parts of the mechanism, process-tracing methods (Beach & Pedersen, 2013) lets us identify how the parts of the mechanism are causally linked with each other.

6.5.1. When are secondary data used as evidence of scientific claims (outcome c or #3)?

Based on results of the analysis from case studies #2, #4, #5, #7, #8, #9, and #10, in which secondary data are used as evidence of scientific claims, the following paragraphs describe the process of how the data-reuse mechanism works empirically, which I have depicted in both Figure 31 (full size in annex 25) and Figure 32 (full size in annex 26). In Figure 31, secondary data are used as the only evidence of scientific claims because condition C4 is met. In Figure 32, secondary data are used to support the evidence of scientific claims made with primary data since condition C4 is not met.

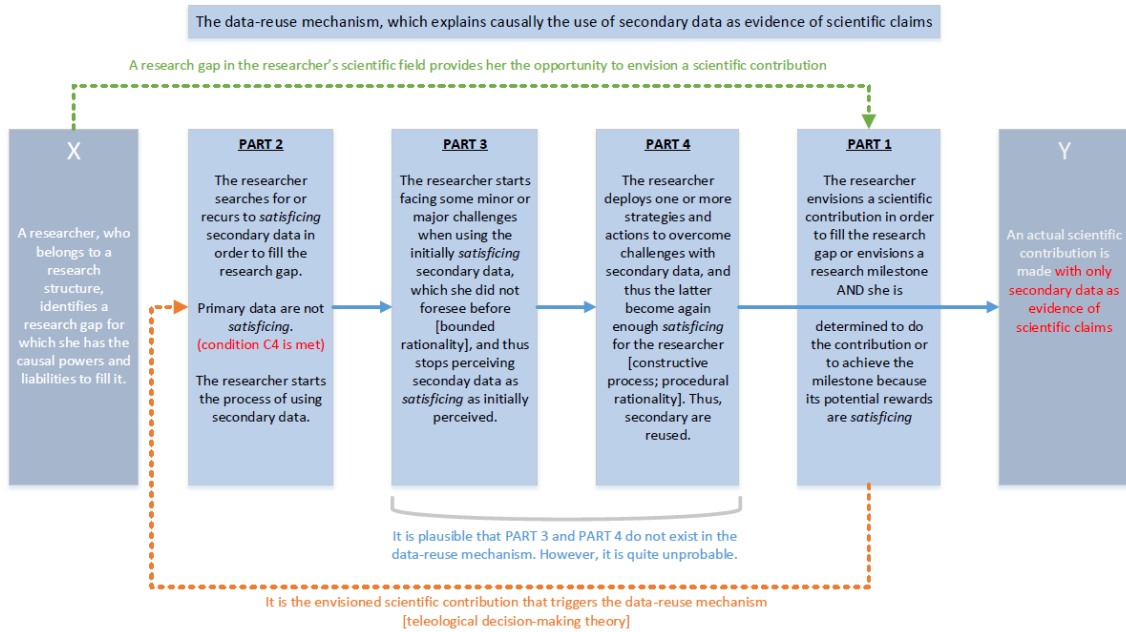


Figure 31 – Process-tracing of the data-reuse mechanism when data are reused as the only evidence of scientific claims

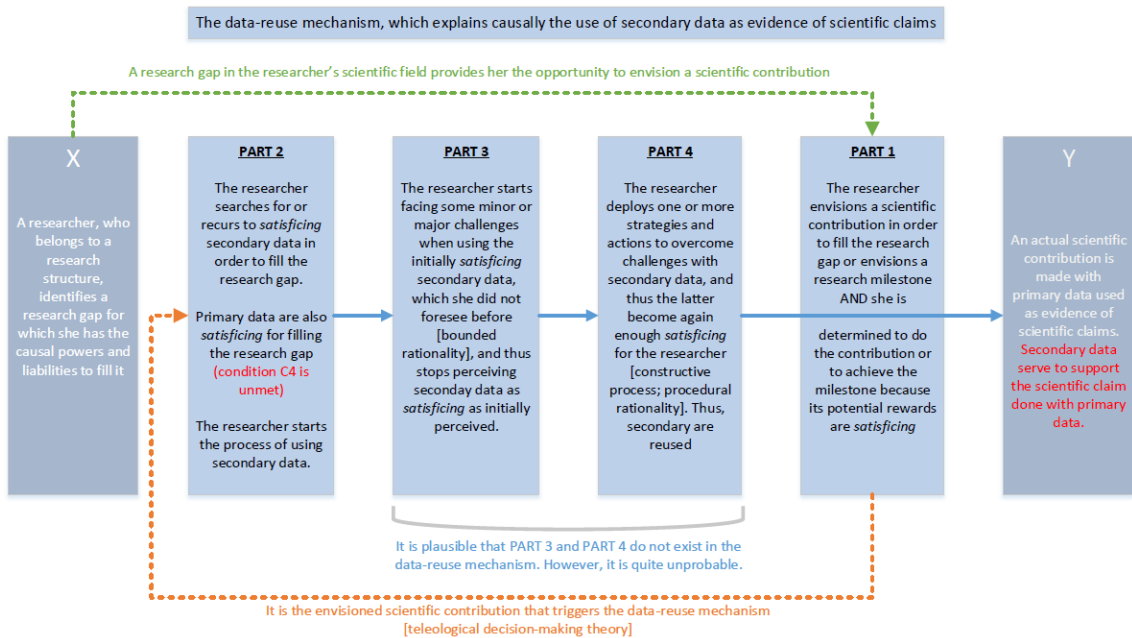


Figure 32 - Process-tracing of the data-reuse mechanism when secondary data are used to support scientific claims done with primary data.

First, a researcher, who belongs to a research or scholarly organization, identifies a research gap for which she has the knowledge and skills required to fill it. In bigger projects, which require

more resources, the researcher might decide to conduct the study with other researchers, and thus all members of the research team are those participating in the process of reusing data. Whether a single researcher or a team of researchers, there is always one leader of the study. The researcher –or members of the research team– is willing to solve a real problem, satisfy their knowledge curiosity, fill a knowledge gap, provide jobs for students and early career researchers, or work collaboratively in teams, or a mix of some or all of these reasons.

Then, in Part 1 of the mechanism, the researcher envisions a scientific contribution in order to fill the research gap and she is determined to achieve it because its potential rewards are sufficient or, rather, satisficing as expounded in the *bounded individual horizon (BIH) model*. The expected goal and its expected resulting scientific contribution (X) triggers the data-reuse causal mechanism. It is the motivation to make a scientific contribution and obtain rewards for it (condition C5), which triggers the causal forces of the mechanism.

In Part 2 of the mechanism, the researcher searches for satisficing secondary data in order to fill the research gap and thus make a scientific contribution. However, other conditions have to exist. The data she obtains will depend on which data she knows about the existence of and whether the researcher is successful in obtaining these data. At the same time, primary data that could substitute for the need to reuse data are not a viable alternative.

In Part 3 of the mechanism, the researcher faces some minor or relevant challenges with the initially satisficing secondary data. These challenges are not known in advance and the researcher has to make some initial efforts in order to have some information about the conditions of access and fitness of the data with her research question(s). Therefore, any decision made will therefore occur within the framework of all the information available to the researcher at the time. When facing these challenges, the researcher may reach a point where they stop perceiving the secondary data as satisficing.

In Part 4 of the mechanism, the researcher deploys one or more strategies and actions in order to overcome challenges with secondary data, and thus these data again become satisficing for the researcher. These strategies or actions are part of the ongoing struggle toward realizing the final goal of making a scientific contribution, which uses secondary data as the evidence of scientific claims. At this point, it is important to acknowledge that the reuse of data as evidence of scientific claims is not a formally validated contribution until a scientific journal or publisher accepts it for publication. However, I suggest that this last step is not part of the data-reuse mechanism.

6.5.2. When some conditions of the data-reuse mechanism are partially met

However, some conditions are not working empirically as initially theorized. Initially and theoretically, the data-reuse mechanism has five conditions with dichotomous values, as happens in a crisp-set logic (Goertz, 2012; Rihoux & Meur, 2012). However, we now know from the analysis of the ten case studies, that condition C3 admit more than two values as it happens in a fuzzy-set logic (Ragin, 2012). In condition C3, researchers may find secondary data less satisficing at the end of the process than at the outset of process of reusing data or vice versa, but researchers still keep perceiving secondary data satisficing. When data stop being satisficing, researchers deploy several strategies with the data and/or the research question(s) in order to perceive data satisficing again, as long as Part 1 or condition C5 exists. If Part 1 or condition C5 disappears once the process has started, then researchers do not proceed to make adjustments with the data and their research question(s).

Condition C2 may have also several values and not only dichotomous values. However, this is only true when the amount of data obtained is satisficing enough for answering the research question or making sound conclusions, as it happens for example with case study #8. The research team conducted the IPD MA with 50% of the eligible IPD studies.

Conditions C1, C4 and C5 admit only dichotomous values (crisp-set logic) in the data-reuse mechanism in light of the analysis of the ten case studies.

6.5.3. When are secondary data used for the creation of background knowledge (outcome b or #2)?

Case study #1 fulfills the theorized conditions of the data-reuse mechanism at time 1, but has the outcome 2 at time 2 (use of secondary data happens, but their reuse is not shared with the research community and these data do not end up being used as evidence of scientific claims). Although the researcher's initial goal was outcome 3 (make a scientific contribution with only secondary data) at time 1, condition C5 is not met subsequently (time 2). The researcher does not find the expected scientific contribution with secondary data and its potential rewards satisficing after some time. The use of the secondary data has happened, but because condition C5 stops being met after it, then the use or analysis of secondary data end up serving as background knowledge (outcome 2). Figure 33 (annex 27) depicts how the disappearance of Part 1 (condition C5) of the data-reuse mechanism in time 1 leads to secondary data finally being used for the creation of background knowledge once the mechanism has been initiated.

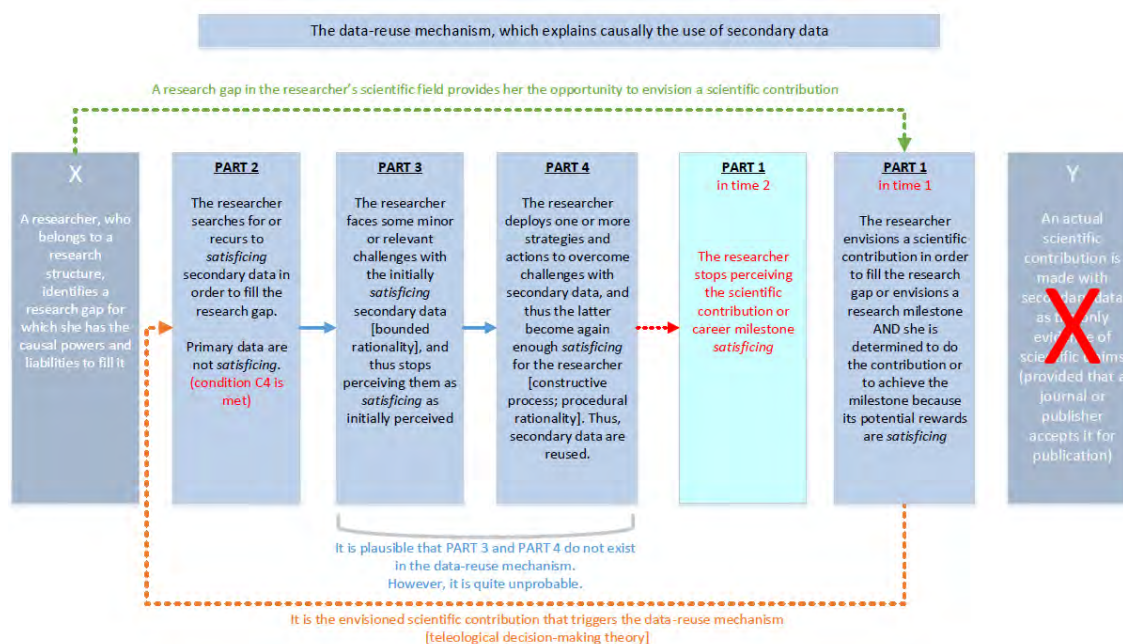


Figure 33 - Process-tracing of the data-reuse mechanism when condition C5 is not met in time 2

Other hypothetical changes could happen in the data-reuse mechanism that could lead to Y=secondary data being used as background knowledge. For instance, in Part 2, condition C4 (primary data are not satisfying) is initially met, but during the process of reusing data, C4 can be stop being met, and thus the researcher decides to use primary data for making the scientific contribution and to use secondary data only for the creation of background knowledge.

Another hypothetical example could be in Part 3, if condition C3 (particular secondary data are an initial satisfying option) is not met after some time, although C3 is met at the outset of the process. The researcher might find that secondary data are not satisfying for making a scientific contribution, and thus decide that she would use secondary data only for the creation of background knowledge. In Part 4, it could happen that, in spite of trying to use different strategies to perceive data as again satisfying, the researcher finds that the strategies are not good scientific practices in her discipline, but still uses the secondary data for the creation of background knowledge.

6.5.4. When are secondary data not finally used (outcome a or #1)?

None of the ten cases gives us the opportunity to know how the causal forces have to be linked in order that the effect of the data reuse process is outcome 1 or a, where data reuse does not happen after having tried or data reuse does not happen at all, as none of the case studies has this outcome.

6.5.4.1. *After having tried using secondary data*

However, we know from case study #3, that David Cook, after trying using TCGA data, decided not to use them. Had I interviewed him at the time when he made that decision, the outcome would have been *a*, and indeed, it was outcome *a* at some middle-point in time of the process of reusing data. The reason for being outcome *a* at a specific time of the process was a change in condition C3. At time 1, David found TCGA data satisficing, but after facing two main challenges, he stopped perceiving TCGA data as satisficing. In other words, he perceived the effort required to use TCGA data to be higher than the benefits of the background knowledge that he and the rest of the lab members could obtain with the reuse of these data. Therefore, we can infer that a change in condition C3, that is stopping to perceive secondary data as satisficing, would be a cause for data not being reused after an attempt has been made.

In addition, in case study #6, when I posed Mary the hypothetical situation of not being able to access any aggregated or individual BORN data, she answered:

I: *Uh, what would have happened if [...] you know, imagine that BORN Ontario couldn't have been able to give you those specific data or would have given you the data in a way that you finally could not adapt them for that summer project?*

Mary: (minute 27:11) *I would have given her a different summer project.*

I: *A different one, okay.*

Mary: (minute 27:19) *Four months is not a long time.*

[...]

It's not like I can look through another data set. It was either going to work or we had to give her something else to do.

So, we can also infer from Mary's answer to the hypothetical situation I laid out regarding condition C2 not being met, that if secondary data are not obtained at all, then data reuse does not happen after trying to access them. This may sound too obvious. However, as discussed above with case studies #8, #9, and #10, the outcome of a decision regarding the reuse of data is not so obvious when only part of the dataset (or part of the variables) that the researcher needs are obtained, while other part of the dataset are not obtained. A decision of whether to reuse or not to reuse is not so straightforward, it

depends on the type of expected scientific contribution expected, the sacrifices with regard to findings and conclusions the research is willing to accept, and at what level the researcher calculates her satisficing threshold to be at each stage of the process.

6.5.4.2. The use of secondary data does not happen at all

In case study #1, it can be hypothesized that, had David Cook stopped perceiving the expected scientific contribution with secondary data and its potential rewards as satisficing much earlier, the reuse of the data would have not happened at all, even if conditions C1 and C2 were met. So, we can infer that a change in condition C5, that is, stopping to perceive a scientific contribution and its rewards as satisficing before starting to reuse the data, would be a cause for data not being reused.

There can be several hypothetical values and combination of the conditions C3, C4 and C5 of the data-reuse mechanism that can lead to secondary data not being reused at all. Table 13 shows four potential combinations.

Table 13 - Several combination of conditions C3, C4 and C5 that do not lead to the reuse of data

Potential initial combinations of conditions C3, C4, and C5.	Potential outcomes
<ul style="list-style-type: none"> ▪ The researcher does not know about the existence of some particular secondary data, thus, condition C3 is not met. ▪ The idea of collecting primary data is a satisficing option (condition C4 is not met). ▪ An expected scientific contribution exists and the researcher finds its potential rewards satisficing (condition C5 is met). 	<ul style="list-style-type: none"> ➤ Use of primary data happen. ➤ Use of secondary data does not happen. <p>(I suggest that the search for secondary data does not happen because the collection and analysis of primary data is satisficing, then the researcher does not try to find secondary data for making a scientific contribution. Yet, it can be plausible that a researcher searches for secondary data in order to support her scientific claims made with primary data. If she finds the secondary data and uses them, then this is combination A of conditions and which has been explained in section 4.2 and depicted in Figure 4.)</p>
<ul style="list-style-type: none"> ▪ Particular secondary data exist and the researcher knows that secondary exist but secondary data are not a satisficing option (condition C3 is not met). ▪ The idea of collecting primary data is a satisficing option (condition C4 is not met). ▪ An expected scientific contribution exists and the researcher finds its potential rewards satisficing (condition C5 is met). 	<ul style="list-style-type: none"> ➤ Use of primary data happen and are used as evidence of scientific claims. ➤ Use of secondary data does not happen.
<ul style="list-style-type: none"> ▪ Particular secondary data does not exist or have not been found by the researcher (condition C1 is not met, and thus C3 cannot not met). ▪ The idea of collecting primary data is not a satisficing option (condition C4 is met). 	<ul style="list-style-type: none"> ➤ The researcher rejects the research question. I hypothesize that the researcher does not even consider or formulate the research question.
<ul style="list-style-type: none"> ▪ Particular secondary data exist, but are not a satisficing option (condition C3 is not met). ▪ The idea of collecting primary data is not a satisficing option (condition C4 is met). 	<ul style="list-style-type: none"> ➤ The researcher rejects the research question. I hypothesize that the researcher does not even consider or formulate the research question.

Chapter 7

Findings

It is correct that summarizing case studies is often difficult, especially as concerns case process. It is less correct as regards case outcomes. The problems in summarizing case studies, however, are due more often to the properties of the reality studied than to the case study as a research method. Often it is not desirable to summarize and generalize case studies. Good studies should be read as narratives in their entirety (Flyvbjerg, 2013, p. 195).

Indeed –as the above quotation asserts–, I find presenting summarized findings¹⁸¹ of the ten case studies a difficult task, especially when the process and changes during the process play a relevant role in the outcome of the phenomenon studied. Thus, in following Flyvbjerg’s advice, I try to present findings in a narrative, though also structured way according to the main findings.

An association between *high effort* and a *scientific contribution*

As I have argued above, and according to the definition of the use of secondary data or the reuse of data in this dissertation, the use of secondary data is not an end in itself, but a means to achieve a goal, namely to make a scientific contribution or to create background knowledge.

¹⁸¹ I am deeply grateful to Grit Laudel, PhD for her advice on disregarding interesting but not useful empirical data for answering my research questions and for presenting findings.

Results of the analysis presented in Chapter 6 show that researchers wanted initially to use secondary data as the only evidence of scientific claims or to support scientific claims made with primary data in a scientific contribution in eight of the ten case studies. The goal of making a scientific contribution is represented by condition C5 in the theorized data-reuse mechanism. However, in one of these eight case studies, namely case study #1, the use or analysis of some secondary data ends up serving to increase researcher's background knowledge. This happened because the goal of making a scientific contribution (condition C5) disappears after accessing and analyzing the secondary data. The disappearance of the goal of making a scientific contribution (condition 5) was not due to challenges in accessing and using the data since case study #1 is a case of using *released data* (or *Open Data*) or because the researcher had not the knowledge and skills to access and analyze the data. What happened is that the researcher stopped considering the potential rewards of a scientific contribution (condition 5) good enough (or *satisficing*) in comparison with the potential rewards of making other scientific contributions.

It is interesting to note that there is association between “high efforts” when reusing data and “a scientific contribution”, and between “very little effort” when reusing data and “background knowledge”. From the definition of *released data*, *stewarded data* and *proprietary data* in this dissertation, and from the analysis of the ten case studies, we know that researchers have to make none or very little effort in accessing secondary *released data*, while they have to invest medium to very high effort in accessing and using secondary *stewarded* and *proprietary* data. From the eight case studies that initially aimed to make a scientific contribution with secondary data, six were of *stewarded data* or *proprietary data*. In fact, participants of these six case studies faced much uncertainty and many challenges in order to access and analyze the data, and in some cases, the process of accessing the data lasted from one to three years. Only two case studies (#1 and #4) aimed to make scientific contributions with secondary *released data*.

The only two case studies (#2 and #3) that initially aimed to use secondary data for the creation of background knowledge were cases using *released data* (or *Open Data*) as shown in Figure 34.

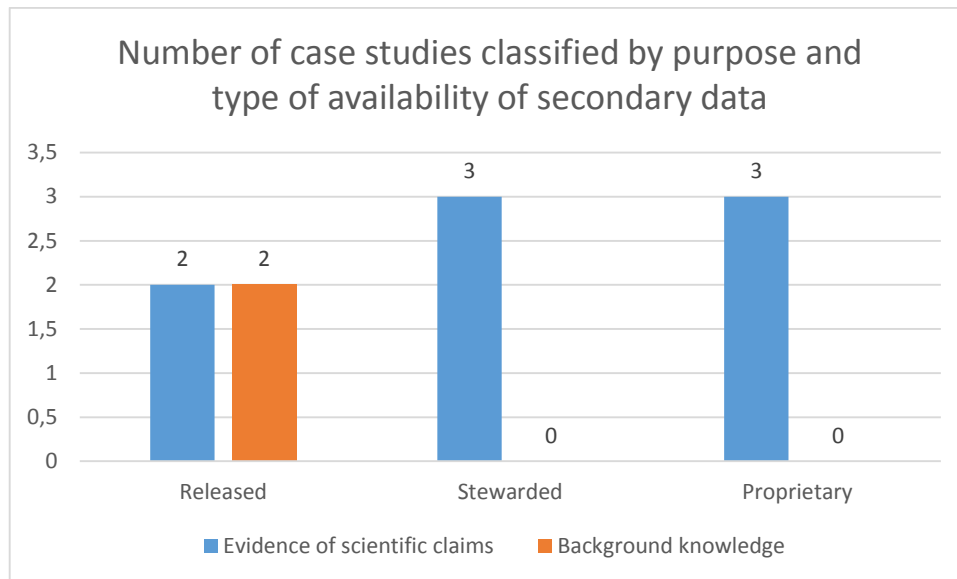


Figure 34 - Association between high effort and the goal of making a scientific contribution in the ten case studies

An association between low effort and the withdrawal of a research goal, and between high effort and the persistence in achieving a research goal

There is also an association between high effort, and thus investments in both human resources and time in order to overcome challenges when reusing data and the persistence in making a scientific contribution. Researchers, who have faced huge challenges and long periods of uncertainty about the outcomes of their decisions and actions during the process of accessing and using the data (for instance, in case studies #5, #9 and #10), were very determined to achieve their goal of making a scientific contribution.

Conversely, when the effort invested is low, it is related with a withdrawal of a research goal. In case study #3, the researcher decided to not use the secondary *released data* at some point of the process due to some small challenges in accessing the data because of some privacy and ethics reasons, and because of the format of the data¹⁸².

In case study #1, the researcher withdrew his research goal, namely a scientific contribution, in a clear example of how decisions on data reuse are nested in broader decisions or are dependent on other decisions.

¹⁸² Later in time, the privacy and ethics walls disappeared and the format in how the data were released changed. So, finally, the researcher decided to reuse the data.

Researchers' strategies for reusing data despite challenges and why they deploy these strategies

Researchers in the ten case studies were not aware of all the challenges that they would face during the process of reusing data before starting the process or after having started it. Thus, they could not really calculate or foresee the efforts that they would have to make in order to make a scientific contribution or increase their background knowledge. Yet, despite the uncertainty and challenges, researchers in the eight of the case studies, who aimed to make a scientific contribution, persevered, deployed all necessary strategies, and took action in order to overcome the challenges and make the scientific contribution. Often this was not exactly as they initially planned, but one that they considered good enough (or *satisficing*). The reason is that researchers adapted constantly their satisficing threshold every time they faced a challenge. In these cases, condition C5 of the data-reuse mechanism was present during the whole process of reusing data. In other words, every time researchers faced a situation that was not good enough for them, they reacted and took action until the situation was again good enough for them. This was the case even if they had to sacrifice and regulate their expectations regarding the kind of contribution they had initially conceived, even to the point of eventually making a scientific contribution, which could not answer their initial research question.

“The order of factors alters the product”

Findings regarding conditions show that when all conditions (C1, C2, C3, C4, C5) are met, the order of conditions and change of the value of the conditions throughout the process of reusing data can affect the process and, thus, the outcome of the process. The skills and knowledge of researchers can also change during the process. The properties of the data can also change over time, as happened in case study #3. After some time, TCGA data were released at a different level of analysis (raw to processed) and the way of accessing the data also changed. Also, contextual situations or the structure of the reuser can change as happened in case study #7. Here, the reuser was initially using secondary data as a public health employee and ended up using secondary data as a researcher. All these changes can affect the outcome of the process.

Furthermore, different properties of the data may also affect the order in which the five conditions are met in time. For instance, with *released data*, the fact that a researcher finds particular data an initial good enough (*satisficing*) option (C3), implies that the researcher knows that secondary data exist and are obtained (C1 and C2 are met previously). They can be antecedent conditions of C3. Yet, conversely, the fact that C1 or C2 is met, does not imply that condition C3 is met because the researcher may find particular secondary data non-satisficing.

With *stewarded data* something similar happens. However, condition C2 (data are obtained) is not necessarily met before the researcher finds the data satisfying (C3), but at least the researcher knows that data under this category is obtainable. With *proprietary data*, there is much uncertainty as to whether data can be obtained and this can happen after several years. So, all of the other four conditions can be met during the process of reusing data and the outcome may not exist until data are obtained (until C2 is met).

Condition C2 (secondary data are obtained) does not have necessarily to be *fully* met. I hypothesized the type of values of the five initial conditions as dichotomous in the data-reuse mechanism. However, after testing the mechanism with the ten case studies, only conditions C1, C4, and C5 presented dichotomous values along the whole process of reusing data. In other words, they were met or unmet. However, conditions C2 and C3 could admit several (ordinal) values. Yet, condition C2 can have ordinal values when the answer of the research question depends on the reuse of more than one data set. When the answer of the research question depends on only one data set, then condition C2 only accepts dichotomous values: secondary data are obtained or data are not obtained.

The condition that activates the reuse of the data, and causes researchers persevere in the reuse of the data is the expected scientific contribution (condition C5), provided that the other parts of the data-reuse mechanism exist.

Data reuse: a nested, embedded and inter-related decision or action

Researchers made constantly decisions or took action at every stage of the process of reusing data, and depending, not only on the challenges they encounter with accessing and using the data, but also on other environmental contextual situations in which they were embedded, and on personal or professional relations that researchers had with others. Researchers in the ten case studies had personal life in mind when making decisions about their research goals. Their values, personality, beliefs and feelings affected their decisions. Yet, they were not directly related to data reuse decisions since data reuse decisions were nested in broader decisions, namely making a scientific contribution or deciding what research line was the best for the lab.

Chapter 8

Discussion. Limitations, and opportunities for further research

Where is the wisdom we have lost in knowledge?

Where is the knowledge we have lost in information?

[Where is the information we have lost in data?]

In “The Rock” by T.S. Eliot, 1934

Decisions on data reuse are nested in broader decisions

The decision-making of reusing secondary data is a small nested decision within a larger one, namely to make a scientific contribution or to broaden one's knowledge in a topic. As I have expounded in Chapter 3, the *bounded individual horizon* (BIH) is a model of the scientific actor, which explains among other things the structuring of scientists' work and careers by the motivation to produce scientific contributions and the rewards system that prioritizes this. Scientists struggle to achieve the objective of creating new findings and receiving recognition and rewards within a frame of limited information, time, and particular institutional, social and other contextual factors. As part of this ongoing contextualized, embedded struggle to produce scientific claims, opportunities present to reuse data. However, data reuse can be a challenging undertaking as previous studies on this topic have shown. Previous studies on data reuse have found many challenges affecting the reuse of data, but some factors, or rather causes that have to do with receiving recognition and reward have been overlooked.

A mechanistic causal explanation has been offered that enables us to understand the data reuse process and its potential outcomes, the creation of background knowledge or new findings. The *data-reuse mechanism*, as it is called, enables us to understand how the satisficing behavior that characterizes scientific decision-making applies to the specific conditions and processes of data reuse, and how this behavior is impelled and driven by scientists' research goals of creating background knowledge or new findings. However, from these two research goals, it is a self-allocated scientific contribution that a researcher expects to achieve, which, not only triggers the reuse of the data, but also maintains researchers motivated in reusing data despite the challenges that they face when reusing data.

Researchers decide to reuse and keep reusing data despite the challenges they face when their goal is to make a scientific contribution with secondary data, either as the main evidence of scientific claims, alone or together with primary data, or as supporting evidence of claims done mainly with primary data. On the contrary, when the research goal consists of creating background knowledge, researchers' willingness and motivation to create background knowledge disappears partially or fully when they face challenges and, thus, they stop the process of reusing data.

Research goals and level of effort

Findings show that there is an association between high efforts in reusing *stewarded data* or *proprietary data* and the completion of a scientific contribution. However, in one of the case studies, namely #4, the scientific contribution is made with *released data*, which require relatively low effort, at least in accessing data, compared to the effort required in accessing *stewarded data* or *proprietary*

data. It would be mistaken to think that scientific contributions cannot be made with relatively low effort. In fact, with relatively low effort, both research goals can be achieved, namely the creation of background knowledge and a scientific contribution. Major investments of time and resources to overcome challenges are only considered worthwhile when the research goal consists solely of making a scientific contribution.

Some limitations and (other) opportunities for further research

The data-reuse mechanism has been shown to be valid and useful for understanding why data reuse leads to scientific contributions in some circumstances and in others not. However, there may be other factors relevant to the data-reuse mechanism which contribute to the reuse of data in producing scientific claims. Further developing the data-reuse mechanisms requires more, and more varied, case studies in order to identify what other changes in the conditions of the data-reuse mechanism are necessary to obtain different outcomes. However, this variability will not be easy to identify without going “into the field” and conducting deep and fine-grained analysis of the decision-making process of reusing data. This detailed analysis is also necessary because data reuse happens in many forms. Some data reuse happens in small projects embedded in larger ones, and only the latter end up in publications, as happened in case studies #4 and #6. In such cases, it may be possible that the data reuse process never becomes known, and this poses a problem not only for studying data reuse but also for the reproducibility of science itself.

The empirical part of this study has been conducted in health disciplines, namely in the sub-disciplines of molecular biology using computational analysis, clinical epidemiology and epidemiology, and in only ten cases. Testing the data-reuse mechanism in other disciplines and in other health sub-disciplines will not only be useful in identifying and, when possible, theorizing, other potential configurations of the data-reuse mechanism, but also in potentially building middle-range theory regarding the phenomenon of researchers’ decision-making when using primary data or secondary data. Comparison of research traditions –or epistemic practices– can be also a useful departure point for testing the data-reuse mechanism in future research.

Furthermore, I suggest that theories and concepts from other disciplines, for instance, psychology, could be helpful in order to answer these type of questions, and to delve into the association between high efforts and the research goal of making a scientific contribution and researchers’ persistence in achieving that goal. For example, the goal-setting theory of motivation (Locke & Latham, 2002; Locke, Latham, Locke, & Latham, 2015; Tosi, Locke, & Latham, 1991), which highlights a relevant relationship between goals and performance, could be useful in explaining researchers’ behavior in achieving challenges goals. Empirical research based on this theory predicts that

“the most effective performance seems to result when goals are specific and challenging, when they are used to evaluate performance and linked to feedback on results, and create commitment and acceptance. The motivational impact of goals may be affected by moderators such as ability and self-efficacy. Deadlines improve the effectiveness of goals. A learning goal orientation leads to higher performance than a performance goal orientation, and group goal-setting is as important as individual goal-setting” (Lunenburg, 2011, p.1)

A multidisciplinary lens for studying researchers’ decisions when reusing data and when performing other types of tasks would be more fruitful than a disciplinary lens.

Regarding my approach to consider all types of data

Some scholars have suggested that it is more difficult to reuse “research data” than data that have been collected or produced in other types of settings because of the difficulties in understanding the context in which the data have been collected or produced in a scientific inquiry context where theories, concepts, and ontologies play a relevant role in the production or collection of the data. However, one of the case studies in this dissertation (#5) has provided us the opportunity to compare the reuse of data collected by two non-research organizations. Data from one institution (BORN Ontario) was easy to reuse, also according to the other two participants that reused BORN Ontario data (case studies #6 and #7). Data from the other institution (ICES) was very difficult to reuse as my interviewee from ICES confirmed¹⁸³. The difference did not rely on the context in which they were produced, but on what type and what level of data curation was applied to the data.

I suggest that studies on data reuse should not be only circumscribed to data collected produced in scholarly or research settings – research data–, since some disciplines in both social and health sciences rely on secondary data collected and/or used in other types of professional settings. I also suggest that it is mistaken to think that data collected and/or used in non-academic settings are easier to understand and to reuse than data collected and/or used in academic settings.

¹⁸³ “Data are extremely complex. We find it takes a typical analyst - so someone who we have hired to work with a researcher to analyze the data - it takes them about a year before they get comfortable with putting the data together.”

Chapter 9

Conclusion

I may be wrong and you may be right,

and by an effort we may get nearer to the truth. (Popper 1902-1994, 1966, p. 420?)

This study makes several contributions. First, it makes two theoretical contributions. One is the *bounded individual horizon (BIH) model*, which is a heuristic model to explain researchers' decisions and behavior when working. The other one, which is underpinned by the BIH model is the *data-reuse mechanism* that explains causally the reuse of data –why and how data reuse happens despite challenges–. Both the model and the mechanism may prove useful for developing middle range theories that explain researchers' decisions when using resources for their research activities. I have theorized the *bounded individual horizon (BIH) model* from critical approaches to rational choice theory, particularly bounded rationality. The concept of *satisficing* has been particularly important here.

Second, this study makes an empirical contribution by testing a causal data-reuse mechanism in ten case studies of data reuse in health sciences, a field in which data reuse has been rarely studied. Results of the analysis have led to a refining of the data-reuse mechanism with regard to some conditions. These conditions can admit ordinal values and not necessarily dichotomous ones as initially theorized.

Third, methodologically, this study, unlike previous ones which have also used a case study approach, uses a diachronic data collection and data analysis based on dual assumptions that the values of conditions or factors affecting the process of reusing data change over time, and that the sequence of these conditions may also affect the outcome. The diachronic data collection has also proved useful in tracking researchers' actions and decisions as they are contextualized in broader embedding decisions, and how changes in a condition can affect other conditions and, thus, the outcome of the process. Longitudinal or diachronic data collection methods may be more appropriate when studying lengthy data reuse processes. Future studies on data reuse should consider the changing value of the conditions of the data-reuse mechanism over time, and the interdependence of these conditions.

Last, but not least, this study may be useful for science policy directions regarding investments in Open Data initiatives, for instance, research data infrastructures and data repositories. Skills and knowledge, acquired by either training or experience, are a necessary condition for not only using secondary data, but for conceiving a research project and for setting a challenging and ambitious, yet attainable, research goal with secondary data. Therefore, funding should be also allocated on the data-reuser side, complementing and creating value from the focus on the data-sharer side. Researchers involved in secondary data analysis, research funders, and Open Data advocates may perceive that *released data* and better research data infrastructures and data repositories are the solution for the success in reusing secondary data. I argue that in isolation these are inappropriate expectations. This study has shown that success in secondary analysis depends importantly on the reuser, particularly where data are not openly released.

Bibliography

Nanos gigantum humeris insidentes, Bernard of Chartres, 12th century

If I have seen further it is by standing on the shoulders of giants, Isaac Newton, 1675

- Abbott, A. D. (2001). *Time matters : on theory and method*. Chicago: Chicago: University of Chicago Press.
- Abbott, A. D. (2004). *Methods of discovery : heuristics for the social sciences*. New York: New York : Norton, 2004.
- Akmon, D., Zimmerman, A., Daniels, M., & Hedstrom, M. (2011). The application of archival concepts to a data-intensive environment: Working with scientists to understand data management and preservation needs. *Archival Science*, *11*(3–4), 329–348. <https://doi.org/10.1007/s10502-011-9151-4>
- Ansbacher, H. L. (1950). The Problem of Interpreting Attitude Survey Data. *Public Opinion Quarterly*, *14*(1), 126. <https://doi.org/10.1086/266155>
- Baker, S. E., & Edwards, R. (2012). How many qualitative interviews is enough ? *National Centre for Research Methods Review Paper*, 1–42. <https://doi.org/10.1177/1525822X05279903>
- Barros, G. (2010). Herbert A. Simon and the concept of rationality: boundaries and procedures. *Revista de Economia Política*, *30*(3), 455–472. <https://doi.org/10.1590/S0101-31572010000300006>
- Baxter, P., & Jack, S. (2008). The Qualitative Report Qualitative Case Study Methodology: Study Design and Implementation for Novice Researchers. *The Qualitative Report Qualitative Report*, *13*(2), 544–559. <https://doi.org/citeulike-article-id:6670384>
- Beach, D. (2017). *Process-Tracing Methods in Social Science*. Oxford University Press. <https://doi.org/10.1093/acrefore/9780190228637.013.176>
- Beach, D., & Pedersen, R. B. (2013). *Process-tracing methods: foundations and guidelines*. (R. B. Pedersen 1978-, M. P. (University of M. Publisher, & R. B. Pedersen 1978- author, Eds.). Ann Arbor : University of Michigan Press, c2013.

- Becker, H. S. (2007). *Telling about society*. Chicago: Chicago : University of Chicago Press, 2007.
- Bell, D. (1976). The coming of the post-industrial society. *The Educational Forum*, 40(4), 575–579. <https://doi.org/10.1080/00131727609336501>
- Borgman, C. L. (2007). *Scholarship in the digital age: Information, infrastructure, and the Internet*.
- Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, 63(6), 1059–1078. <https://doi.org/10.1002/asi.22634>
- Borgman, C. L. (2015). *Big data, little data, no data: scholarship in the networked world*.
- Boulton, G., Campbell, P., Collins, B., Elias, P., Hall, W., Graeme, L., ... Walport, M. (2012). *Science as an open enterprise*. *Science*. Retrieved from http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf
- Bourdieu, P., & Wacquant, L. J. D. (1992). *An invitation to reflexive sociology*. (L. J. D. Wacquant, Ed.). Chicago: University of Chicago Press.
- Bowker, G. C. (2013). Data Flakes: An Afterword to “Raw Data” Is an Oxymoron. In L. Gitelman (Ed.), *“Raw Data” Is an Oxymoron* (pp. 167–171).
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/http://dx.doi.org/10.1191/1478088706qp063oa>
- Brignardello-Petersen, R., Rochweg, B., & Guyatt, G. H. (2014). What is a network meta-analysis and how can we use it to inform clinical practice? *Polskie Archiwum Medycyny Wewnętrznej*. <https://doi.org/10.14219/jada.archive.2013.0035>
- Bril-Mascarenhas, T., Maillet, A., & Mayaux, P.-L. (2017). Process tracing. Inducción, deducción e inferencia causal. *Revista de Ciencia Política*, 37(3), 659–684.
- Brinkmann, S. (2018). The interview. In N. K. Denzin & Y. S. Lincoln (Eds.), *The Sage handbook of qualitative research* (Fifth edit). Los Angeles : Sage, 2018.
- Brown, C. (2003). The changing face of scientific discourse: Analysis of genomic and proteomic database usage and acceptance. *Journal of the American Society for Information Science and Technology*, 54(10), 926–938. <https://doi.org/10.1002/asi.10289>
- Butler, L. (2003). Explaining Australia’s increased share of ISI publications - The effects of a funding formula based on publication counts. *Research Policy*, 32(1), 143–155. [https://doi.org/10.1016/S0048-7333\(02\)00007-0](https://doi.org/10.1016/S0048-7333(02)00007-0)
- Case, D. (2007). *Looking for information : a survey of research on information seeking, needs, and behavior*. Amsterdam: Elsevier/Academic Press.
- Castle, J. (2003). Maximizing research opportunities: Secondary data analysis. *Journal of Neuroscience Nursing*, 35(5), 287–290.
- Cheng, H. G., & Phillips, M. R. (2014). Secondary analysis of existing data: opportunities and implementation. *Shanghai Archives of Psychiatry*, 26(6), 371–375. <https://doi.org/10.11919/j.issn.1002-0829.214171>
- Clarke, S. P., & Cossette, S. (2000). Secondary Analysis: Theoretical, Methodological, and Practical Considerations. *Canadian Journal of Nursing Research*, 32(3), 109–129.
- Connaway, L. S., & Powell, R. R. (2010). *Basic research methods for librarians*. Santa Barbara, Calif.: Libraries Unlimited. Retrieved from <http://mirlyn.lib.umich.edu/Record/009835113>
- Cook, D. P., & Vanderhyden, B. C. (2019). Comparing transcriptional dynamics of the epithelial-

- mesenchymal transition. *BioRxiv*. <https://doi.org/10.1101/732412>
- Costello, M. J. (2009). Motivating online publication of data. *BioScience*, *59*(5), 418–427. Retrieved from <http://bioscience.oxfordjournals.org/content/59/5/418.short>
- Creswell, J. W. (2014). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*.
- Curty, R. G. (2015). *Beyond “data thrifting”: an investigation of factors influencing research data reuse in the social sciences (2015)*. *Dissertations-ALL.266*. Retrieved from <http://surface.syr.edu/etd/266>
- Curty, R. G., Crowston, K., Specht, A., Grant, B. W., & Dalton, E. D. (2017). Attitudes and norms affecting scientists’ data reuse. *PLoS ONE*, *12*(12), 1–22. <https://doi.org/10.1371/journal.pone.0189288>
- Curty, R. G., & Qin, J. (2014). Towards a model for research data reuse behavior. *Proceedings of the American Society for Information Science and Technology*, *51*(1), 1–4. <https://doi.org/10.1002/meet.2014.14505101072>
- Curty, R. G., Yoon, A., Jeng, W., & Qin, J. (2016). Untangling data sharing and reuse in social sciences. *Proceedings of the Association for Information Science and Technology*, *53*(1), 1–5. <https://doi.org/10.1002/pr2.2016.14505301025>
- Dale, A., Gilbert, G. N., & Arber, S. (1983). The General Household Survey as a source for secondary analysis. *Sociology*, *17*, 255–259. <https://doi.org/10.1177/0038038583017002006>
- Daniels, M. G. (2014). Data Reuse in Museum Contexts: Experiences of Archaeologists and Botanists.
- de Rijcke, S., Wouters, P. F., Rushforth, A. D., Franssen, T. P., & Hammarfelt, B. (2016). Evaluation practices and effects of indicator use—a literature review. *Research Evaluation*, *25*(2), 161–169. <https://doi.org/10.1093/reseval/rvv038>
- de Vries, J. I. P., van Pampus, M. G., Hague, W. M., Bezemer, P. D., & Joosten, J. H. (2012). Low-molecular-weight heparin added to aspirin in the prevention of recurrent early-onset pre-eclampsia in women with inheritable thrombophilia: The FRUIT-RCT. *Journal of Thrombosis and Haemostasis*, *10*(1), 64–72. <https://doi.org/10.1111/j.1538-7836.2011.04553.x>
- Dodge, M., & Kitchin, R. (2003). Charting Movement : Mapping. *Network*, (Couclelis 1998), 1–31.
- Donaldson, D. R., & Conway, P. (2015). User conceptions of trustworthiness for digital archival documents. *Journal of the Association for Information Science and Technology*, *66*(12), 2427–2444. <https://doi.org/10.1002/asi.23330>
- Doolan, D. M., & Froelicher, E. S. (2009). Using an Existing Data Set to Answer New Research Questions: A Methodological Review. *Research and Theory for Nursing Practice*, *23*(3), 203–215. <https://doi.org/10.1891/1541-6577.23.3.203>
- Dunn, S. L., Arslanian-Engoren, C., DeKoekkoek, T., Jadack, R., & Scott, L. D. (2015). Secondary Data Analysis as an Efficient and Effective Approach to Nursing Research. *Western Journal of Nursing Research*, *37*(10), 1295–1307. <https://doi.org/10.1177/0193945915570042>
- Easton, G. (2010). Critical realism in case study research. *Industrial Marketing Management*, *39*(1), 118–128. <https://doi.org/10.1016/j.indmarman.2008.06.004>
- Eisenhardt, K. M. (1989). Building Theories from Case Study Research Published by : Academy of Management Stable URL : <http://www.jstor.org/stable/258557> Linked references are available on JSTOR for this article : Building Theories from Case Study Research, *14*(4), 532–550.
- Elder-Vass, D. (2010). *The causal power of social structures. Emergence, structure and agency*. New York: Cambridge University Press.

- Elster, J. (1989). *Nuts and bolts for the social sciences*. Cambridge: Cambridge University Press.
- Faniel, I. M., Barrera-Gomez, J., Kriesberg, A., & Yakel, E. (2013). A Comparative Study of Data Reuse Among Quantitative Social Scientists and Archaeologists. *IConference*, 797–800. <https://doi.org/10.9776/13391>
- Faniel, I. M., Frank, R. D., & Yakel, E. (2019). Context from the data reuser's point of view. *Journal of Documentation, ahead-of-p*(ahead-of-print), 1–31. <https://doi.org/10.1108/JD-08-2018-0133>
- Faniel, I. M., & Jacobsen, T. E. (2010). Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data. *Computer Supported Cooperative Work*, 19, 355–375. <https://doi.org/10.1007/s10606-010-9117-8>
- Faniel, I. M., Kansa, E., Kansa, S. W., Barrera-Gomez, J., & Yakel, E. (2013). The challenges of digging data: A study of context in archaeological data reuse. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*. <https://doi.org/10.1145/2467696.2467712>
- Faniel, I. M., Kriesberg, A., & Yakel, E. (2012). Data reuse and sensemaking among novice social scientists. *Proceedings of the ASIST Annual Meeting*, 49(1). <https://doi.org/10.1002/meet.14504901068>
- Faniel, I. M., Kriesberg, A., & Yakel, E. (2016). Social scientists' satisfaction with data reuse. *Journal of the Association for Information Science and Technology*, 67(6), 1404–1416. <https://doi.org/10.1002/asi.23480>
- Faniel, I. M., & Zimmerman, A. (2011). Beyond the Data Deluge: A Research Agenda for Large-Scale Data Sharing and Reuse. *International Journal of Digital Curation*, 6(1), 58–69. <https://doi.org/10.2218/ijdc.v6i1.172>
- Fear, K. M. (2013). *Measuring and anticipating the impact of data reuse*.
- Fecher, B., & Friesike, S. (2014). Open Science: one term, five schools of thought. In S. Bartling & S. Friesike (Eds.), *Opening Science. The evolving guide on how the internet is changing research, collaboration and scholarly publishing* (pp. 17–47). Springer.
- Federer, L. M., Lu, Y.-L., Joubert, D. J., Welsh, J., & Brandys, B. (2015). Biomedical Data Sharing and Reuse: Attitudes and Practices of Clinical and Scientific Research Staff. *Plos One*, 10(6), e0129506. <https://doi.org/10.1371/journal.pone.0129506>
- Fell, D. B. (2015). *Influenza illness and influenza vaccination during pregnancy and risk of preterm birth and fetal death*.
- Fell, D. B., Platt, R. W., Basso, O., Wilson, K., Kaufman, J. S., Buckeridge, D. L., & Kwong, J. C. (2018). The relationship between 2009 pandemic H1N1 influenza during pregnancy and preterm birth: a population-based cohort study. *Epidemiology*, 29(1), 107–116. <https://doi.org/10.1097/EDE.0000000000000753>
- Feltovich, P. J., Spiro, R. J., & Coulson, R. L. (1997). Issues of expert flexibility in contexts characterized by complexity and change. In K. M. Ford, P. J. Feltovich, & R. R. Hoffman (Eds.), *Expertise in context: Human and machine*. Menlo Park [etc.]: Menlo Park etc. : AAAI Press, cop. 1997.
- Fienberg, S. E., Martin, M. E., & Straf, M. L. (Eds.). (1985). *Sharing research data*. Washington, D.C.: National Academy Press. <https://doi.org/10.1126/science.229.4714.632>
- Flick, U. (2009). *An introduction to qualitative research. Qualitative research* (4th ed.). London: SAGE Publications Ltd.
- Flyvbjerg, B. (2006). Five misunderstandings about case-study research. *Qualitative Inquiry*, 12(2), 219–245. <https://doi.org/10.1177/1077800405284363>

- Flyvbjerg, B. (2013). Case study. In N. K. Denzin & Y. S. Lincoln (Eds.), *Strategies of qualitative inquiry* (4th ed.). Los Angeles: SAGE.
- Forés-Martos, J., Catalá-López, F., Sánchez-Valle, J., Ibáñez, K., Tejero, H., Palma-Gudiel, H., ... Tabarés-Seisdedos, R. (2019). Transcriptomic metaanalyses of autistic brains reveals shared gene expression and biological pathway abnormalities with cancer. *Molecular Autism*, *10*(1), 1–16. <https://doi.org/10.1186/s13229-019-0262-8>
- Frank, R. D., Chen, Z., Crawford, E., Suzuka, K., & Yakel, E. (2017). Trust in qualitative data repositories. *Proceedings of the Association for Information Science and Technology*, *54*(1), 102–111. <https://doi.org/10.1002/pr2.2017.14505401012>
- Frank, R. D., Yakel, E., & Faniel, I. M. (2015). Destruction/reconstruction: preservation of archaeological and zoological research data. *Archival Science*, *15*(2), 141–167. <https://doi.org/10.1007/s10502-014-9238-9>
- García-Sancho, M. (2012a). *Biology, computing and the history of molecular sequencing: from proteins to DNA, 1945-2000*. Palgrave Macmillan. <https://doi.org/DOI.10.1057/9780230370937>
- García-Sancho, M. (2012b). From the genetic to the computer program: the historicity of “data” and “computation” in the investigations on the nematode worm *C. elegans* (1963-1998). *Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(1), 16–28. <https://doi.org/10.1016/j.shpsc.2011.10.002>
- Garmon Bibb, S. C. (2007). Issues associated with secondary analysis of population health data. *Applied Nursing Research*, *20*(2), 94–99. <https://doi.org/10.1016/j.apnr.2006.02.003>
- Gerring, J. (2004). What is a case study and what is it good for? *American Political Science Review*, *98*(2), 341–354. Retrieved from <http://www.jstor.org/stable/4145316>
- Gerring, J. (2010). Causal mechanisms: Yes, but... *Comparative Political Studies*, *43*(11), 1499–1526. <https://doi.org/10.1177/0010414010376911>
- Gilbert, G. N., & Mulkay, M. (1984). *Opening Pandora's box a sociological analysis of scientists' discourse*. Cambridge [UK]: Cambridge University Press.
- Gläser, J. (2012). How does Governance change research content? On the possibility of a sociological middle-range theory linking science policy studies to the sociology of scientific knowledge. *The Technical University Technology Studies Working Papers*. Retrieved from <https://www.ts.tu-berlin.de/fileadmin/fg226/TUTS/TUTS-WP-1-2012.pdf>
- Gläser, J., & Laudel, G. (2013). Life With and Without Coding: Two Methods for Early-Stage Data Analysis in Qualitative Research Aiming at Causal Explanations. *Forum : Qualitative Social Research*, *14*(2). Retrieved from <http://search.proquest.com/docview/1356976111>
- Gläser, J., & Laudel, G. (2015). A Bibliometric Reconstruction of Research Trails for Qualitative Investigations of Scientific Innovations The Need for Innovation Research to Analyse Abrupt Change in Research Content, *40*, 299–330. <https://doi.org/10.12759/hsr.40.2015.3.299-330>
- Goertz, G. (2012). *A tale of two cultures : qualitative and quantitative research in the social sciences*. (J. Mahoney 1968-, Ed.). Princeton, N.J.: Princeton, N.J. : Princeton University Press, 2012.
- Gregory, K., Groth, P., Scharnhorst, A., & Wyatt, S. (2019). Lost or found? Discovering data needed for research. *Preprint of an Article under Review*, (August). Retrieved from <https://arxiv.org/ftp/arxiv/papers/1909/1909.00464.pdf>
- Gris, J. C., Chauleur, C., Faillie, J. L., Baer, G., Marès, P., Fabbro-Peray, P., ... Dauzat, M. (2010). Enoxaparin for the secondary prevention of placental vascular complications in women with abruptio placentae: The pilot randomised controlled NOH-AP trial. *Thrombosis and Haemostasis*, *104*(4), 771–779. <https://doi.org/10.1160/TH10-03-0167>

- Gris, J. C., Chauleur, C., Molinari, N., Marès, P., Fabbro-Peray, P., Quéré, I., ... Dautat, M. (2011). Addition of enoxaparin to aspirin for the secondary prevention of placental vascular complications in women with severe pre-eclampsia: The pilot randomised controlled NOH-PE trial. *Thrombosis and Haemostasis*, 106(6), 1053–1061. <https://doi.org/10.1160/TH11-05-0340>
- Guba, E. G., & Lincoln, Y. S. (2005). Paradigmatic Controversies, Contradictions, and Emerging Confluences. In *The Sage handbook of qualitative research, 3rd ed.* (pp. 191–215). Thousand Oaks, CA: Sage Publications Ltd.
- Hakim, C. (2013). Secondary analysis and the relationship between official and academic social research. In J. Goodwin (Ed.), *Sage Secondary Data Analysis* (pp. 27-44 (print)). London: SAGE Publications Ltd. <https://doi.org/10.4135/9781473963702>
- Hammarfelt, B., & de Rijcke, S. (2015). Accountability in context: Effects of research evaluation systems on publication practices, disciplinary norms, and individual working routines in the faculty of Arts at Uppsala University. *Research Evaluation*, 24(1), 63–77. <https://doi.org/10.1093/reseval/rvu029>
- Hartwig, M. (2007). *Dictionary of critical realism*. London.
- Heaton, J. (2008). Secondary analysis of qualitative data: An overview. *Historical Social Research / Historische Sozialforschung*, 33(3 (125)), 33–45. Retrieved from <http://www.jstor.org/stable/20762299>
- Hedström, P., & Swedberg, R. (Eds.). (1998). *Social mechanisms: an analytical approach to social theory*. New York: Cambridge University Press.
- Hedström, P., & Ylikoski, P. (2010). Causal mechanisms in the social sciences. *Annual Review of Sociology*, 36(1), 49–67. <https://doi.org/10.1146/annurev.soc.012809.102632>
- Heeney, C., Hawkins, N., de Vries, J., Boddington, P., & Kaye, J. (2010). Assessing the privacy risks of data sharing in genomics. *Public Health Genomics*, 14(1), 17–25. <https://doi.org/10.1159/000294150>
- Hilgartner, S. (1995). Biomolecular Databases - New Communication Regimes for Biology. *Science Communication*. <https://doi.org/10.1177/1075547095017002009>
- Hilgartner, S. (1998). Data access policy in genome research. In A. Thackray (Ed.), *Private science: biotechnology and the rise of the molecular sciences*. Philadelphia: University of Pennsylvania Press.
- Hilgartner, S. (2017). *Reordering life : knowledge and control in the genomics revolution*. Cambridge, Massachusetts : The MIT Press, 2017.
- Hilgartner, S., & Brandt-Rauf, S. I. (1994). Data access, ownership, and control. *Knowledge: Creation, Diffusion, Utilization*, 15(4), 355–372.
- Hinds, P. P., Vogel, R. R., & Clarke-Steffen, L. (1997). The possibilities and pitfalls of doing a secondary analysis of a qualitative data set. *Qualitative Health Research*, 7(3), 408–424. <https://doi.org/10.1177/104973239700700306>
- Hodge, J. G. J., & Gostin, L. O. (2004). *Public health practice vs research: A report for Public Health Practitioners Including Cases and Guidance for Making Distinctions. Report funded by Council of State and Territorial Epidemiologists in Atlanta, Georgia*. <https://doi.org/10.1097/DMP.0b013e318187310c>
- Hsu, L., Martin, R. L., McElroy, B., Litwin-Miller, K., & Kim, W. (2015). Data management, sharing, and reuse in experimental geomorphology: Challenges, strategies, and scientific opportunities. *Geomorphology*, 244, 180–189. <https://doi.org/10.1016/j.geomorph.2015.03.039>

- Hummel, N., Debray, T. P. A., Didden, E., Egger, M., Fletcher, C., Moons, K. G. M., ... Valkenhoef, G. Van. (n.d.). *Work Package 4 Methodological guidance, recommendations and illustrative case studies for (network) meta-analysis and modelling to predict real-world effectiveness using individual participant and/or aggregate data. Get Real.*
- Huvila, I. (2009). Analytical information horizon maps. *Library and Information Science Research*, 31(1), 18–28. <https://doi.org/10.1016/j.lisr.2008.06.005>
- Hyman, H. H. (1972). *Secondary analysis of sample surveys*. New York: Wiley.
- Ioannidis, J. P. A., Allison, D. B., Ball, C. A., Coulibaly, I., Cui, X., Culhane, A. C., ... van Noort, V. (2009). Repeatability of published microarray gene expression analyses. *Nature Genetics*, 41(2), 149–155. <https://doi.org/10.1038/ng.295>
- Irwin, S., & Winterton, M. (2011). Debates in Qualitative Secondary Analysis : Critical Reflections Paper, 3349(4), 1–23. <https://doi.org/ISSN: 1758 3349>
- Jacobson, A. F., Hamilton, P., & Galloway, J. (1993). Obtaining and evaluating data sets for secondary analysis in nursing research. *Western Journal of Nursing Research*, 15(4), 483.
- Jasanoff, S. (2005). *Designs on nature science and democracy in Europe and the United States*. Princeton, N.J: Princeton University Press.
- Jaspers, G. J., & Degraeuwe, P. L. J. (2014). A failed attempt to conduct an individual patient data meta-analysis. *Systematic Reviews*, 3(1), 1–2. <https://doi.org/10.1186/2046-4053-3-97>
- Jones, B. D. (1999). Bounded Rationality. *Annual Review of Political Science*, 2, 297–321. https://doi.org/10.1007/978-1-349-20568-4_5
- Jørgensen, U. (2012). Mapping and navigating transitions. *Research Policy*, 41(6), 996. <https://doi.org/10.1016/j.respol.2012.03.001>
- Kaandorp, S. P., Goddijn, M., Van Der Post, J. A. M., Hutten, B. A., Verhoeve, H. R., Hamulyák, K., ... Middeldorp, S. (2010). Aspirin plus heparin or aspirin alone in women with recurrent miscarriage. *Obstetrical and Gynecological Survey*, 65(10), 621–622. <https://doi.org/10.1097/OGX.0b013e3182021f71>
- Kahlem, P., & Birney, E. (2006). Dry work in a wet world: Computation in systems biology. *Molecular Systems Biology*, 2. <https://doi.org/10.1038/msb4100080>
- Kane, H., Lewis, M. A., Williams, P. A., & Kahwati, L. C. (2014). Using qualitative comparative analysis to understand and quantify translation and implementation. *Translational Behavioral Medicine*, 4(2), 201–208. <https://doi.org/10.1007/s13142-014-0251-6>
- Kaye, J., Heeney, C., Hawkins, N., de Vries, J., & Boddington, P. (2009). Data sharing in genomics—re-shaping scientific practice. *Nature Reviews Genetics*, 10(5), 331–335. Retrieved from <http://www.nature.com/nrg/journal/v10/n5/abs/nrg2573.html>
- Kiecolt, K J, & Nathan, L. E. (1985). Secondary Analysis of Survey Data, 10–14.
- Kiecolt, K Jill, & Nathan, L. E. (2012). Introduction to Secondary Analysis of Survey Data. *SAGE Secondary Data Analysis*.
- Kilgour, D. M. ed., & Eden, C. ed. (2010). *Handbook of group decision and negotiation*. (D. M. Kilgour & C. Eden, Eds.) (Vol. 4). Dordrecht: Springer Netherlands. <https://doi.org/10.1007/978-90-481-9097-3>
- Kim, Y., & Yoon, A. (2017). Scientists' data reuse behaviors: A multilevel analysis. *Journal of the Association for Information Science and Technology*, 68(12), 2709–2719. <https://doi.org/10.1002/asi.23892>

- King, G., Keohane, R. O., & Verba, S. (1994). *Designing social inquiry scientific inference in qualitative research*. (R. O. (Robert O. Keohane 1941- & S. Verba, Eds.). Princeton, N.J.: Princeton, N.J. : Princeton University Press, c1994.
- Kitchin, R. (2014). The Data Revolution: Big data, Open data, Data Infrastructures and their consequences.
- Klein, G., Phillips, J. K., Rall, E. L., & Peluso, D. A. (2007). A data-frame theory of sensemaking. In R. R. Hoffman (Ed.), *Expertise out of context : proceedings of the Sixth International Conference on Naturalistic Decision Making*. New York : Lawrence Erlbaum Associates, 2007.
- Kohler, R. E. (1994). *Lords of the fly : Drosophila genetics and the experimental life*. Chicago: Chicago : University of Chicago Press, 1994.
- Latour, B. (1987). *Science in action: how to follow scientists and engineers through society*. Cambridge, Mass.: Harvard University Press.
- Laudel, G. (2002). Collaboration and reward. *Beaver*, 11(1), 3–15. <https://doi.org/10.1007/978-1-4419-7082-4>
- Laudel, G., & Gläser, J. (2007). Interviewing Scientists. *Science, Technology and Innovation Studies*, 3(2), 91–111. <https://doi.org/10.4324/9781315671338-4>
- Law, M. (2005). Reduce , Reuse , Recycle : Issues in the Secondary Use of Research Data. *IASSIST Quarterly*, 29(1), 5–10. Retrieved from <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Reduce+,+Reuse+,+Recycle+:++Issues+in+the+Secondary+Use+of+Research+Data#0>
- Leonelli, S. (2015). What Counts as Scientific Data? A Relational Framework. *Philosophy of Science*, 82(5), 810–821. <https://doi.org/10.1086/684083>
- Leonelli, S. (2016). *Data-centric biology: a philosophical study*. (T. U. of C. Press, Ed.). The University of Chicago Press. <https://doi.org/10.7208/chicago/980226416502.001.0001>
- Lesch, C. W. C., & Hazeltine, J. E. (2012). Secondary Research , New Product Screening , and the Marketing Research Course : An Experiment in Structured Decision Making. *Sage Secondary Data Analysis*, (April 1988), 1–13. <https://doi.org/10.1177/027347539001200105>
- Leydesdorff, L., & Rafols, I. (2012). Interactive overlays: A new method for generating global journal maps from Web-of-Science data. *Journal of Informetrics*, 6(2), 318–332. <https://doi.org/10.1016/j.joi.2011.11.003>
- Little, D. (1991). *Varieties of social explanation: an introduction to the philosophy of social science*. Boulder, Colorado: Westview Press.
- Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist*, 57(9), 705–717. <https://doi.org/10.1037/0003-066X.57.9.705>
- Locke, E. A., Latham, G. P., Locke, E. A., & Latham, G. P. (2015). New Directions in Goal-Setting Theory New Directions in Goal-Setting Theory. *Psychological Science*, 15(October), 265–268. <https://doi.org/10.1111/j.1467-8721.2006.00449.x>
- Lopez de Vallejo, I., Scerri, S., & Tuikka, T. (2019). Towards a European Data Sharing Space, (April). Retrieved from <http://bdva.eu/AIPPP-Vision-paper-PressRelease>
- López, J. (2004). How sociology can save bioethics . . . maybe. *Sociology of Health & Illness*, 26(7), 875–896. <https://doi.org/10.1111/j.0141-9889.2004.00421.x>
- Lunenburg, F. C. (2011). Goal-Setting Theory of Motivation. *International Journal of Management*,

- Business, and Administration*, 15(1), 1–6. Retrieved from <http://www.nationalforum.com/Electronic Journal Volumes/Lunenburger, Fred C. Goal-Setting Theory of Motivation IJMBA V15 N1 2011.pdf>
- Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, 67(1), 1–25. <https://doi.org/10.1086/392759>
- Marshall, C., & Rossman, G. B. (2011). *Designing qualitative research* (5th ed.). Los Angeles: SAGE Publications, Inc.
- Martin, S. N. V. (2012). *Research sans frontières? The effects of funding schemes on international research collaboration*. <https://doi.org/10.25911/5d5151831323c>
- Martinelli, I., Ruggerenti, P., Cetin, I., Pardi, G., Perna, A., Vergani, P., ... Mannucci, P. M. (2012). Heparin in pregnant women with previous placenta-mediated pregnancy complications: A prospective, randomized, multicenter, controlled clinical trial. *Blood*, 119(14), 3269–3275. <https://doi.org/10.1182/blood-2011-11-391383>
- Maxwell, J. A. (2004). Causal Explanation, Qualitative Research,. *Educational Researcher*, 33(2), 3–11.
- Mayntz, R. (2004). Mechanisms in the Analysis of Social Macro-Phenomena. *Philosophy of the Social Sciences*, 34(2), 237–259. <https://doi.org/10.1177/0048393103262552>
- McAllister, J. W. (2018). Scientists' reuse of old empirical data: Epistemological aspects. *Philosophy of Science*, 85(5), 755–766. <https://doi.org/10.1086/699695>
- Meltzoff, J. (1998). *Critical thinking about research: psychology and related fields*. Washington, DC: American Psychological Association.
- Miles, Mathew B, & Huberman, A. M. (1994). *Qualitative Data Analysis: An expanded sourcebook* (2nd ed.). California: SAGE Publications, Inc.
- Miles, Matthew B, & Huberman, A. M. (1996). *Qualitative data analysis: An expanded sourcebook*.
- Miles, Matthew B, Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis : a methods sourcebook* (3rd ed.). Los Angeles: SAGE Publications, Inc.
- Mingus, M. S. (2007). Bounded rationality and organizational influence: Herbert Simon and the behavioral revolution. In G. Mor. . I (Ed.), *Handbook of decision making* (pp. 61–79). Boca Raton, FL: Boca Raton, FL : Taylor & Francis, c2007.
- Moore, N. (2006). The Contexts of Context: Broadening Perspectives in the (Re)use of Qualitative Data. *Methodological Innovations Online*, 1(2), 21–32. <https://doi.org/10.4256/mio.2006.0009>
- Moore, W. J., Newman, R. J., Sloane, P. J., Steely, J. D., & Corp, A. (2002). Productivity Effects of Research Assessment Exercises. *Departmental Working Papers, Department of Economics, Louisiana State University*, (January 2002). Retrieved from http://www.bus.lsu.edu/economics/papers/pap02_15.pdf
- Morçöl, G. (2007a). Decision making: an overview of theories, contexts, and methods. In G. Morçöl (Ed.), *Handbook of decision making* (pp. 3–18). Boca Raton, FL: Taylor & Francis Group.
- Morçöl, G. (Ed.). (2007b). *Handbook of decision making*. Boca Raton, FL: Taylor & Francis Group.
- Murillo, A. P. (2016). *Data sharing and data reuse: an investigation of descriptive information facilitators and inhibitors*.
- Nahar, V., & He, L. (2016). Reuse of scientific data in academic publications: An investigation of Dryad Digital Repository. *Aslib Journal of Information Management*, 68(4), 478–495.

- Neuendorf, K. A. (2002). *The content analysis guidebook*. Thousand Oaks, Calif.: Thousand Oaks, Calif. : Sage Publications, c2002.
- Niu, J. (2009a). Overcoming inadequate documentation. *Proceedings of the American Society for Information Science and Technology*, 46(1), 1–14. <https://doi.org/10.1002/meet.2009.145046024>
- Niu, J. (2009b). *Perceived documentation quality of social science data*.
- Niu, J., & Hedstrom, M. (2009). Documentation evaluation model for social science data. *Proceedings of the American Society for Information Science and Technology*, 45(1), 11–11. <https://doi.org/10.1002/meet.2008.1450450223>
- Nutt, P. C., & Wilson, D. C. (2010a). Discussion and implications: toward creating a unified theory of decision making. In P. C. Nutt & D. C. Winson (Eds.), *Handbook of decision making* (pp. 645–677). John Wiley & Sons, Ltd.
- Nutt, P. C., & Wilson, D. C. (Eds.). (2010b). *Handbook of decision making*. John Wiley & Sons, Ltd.
- Oneal, J. R. (1988). The rationality of decision daking during international crises. *Polity*, 20(4), 598–622. <https://doi.org/10.2307/3234897>
- Orsi, A. J., Grey, M., Mahon, M. M., Moriarty, H. J., Shepard, M. P., & Carroll, R. M. (1999). Conceptual and Technical Considerations when Combining Large Data Sets. *Western Journal of Nursing Research*, 21(2), 130–142. <https://doi.org/10.1177/01939459922043785>
- Ortega y Gasset, J. (1914). *Meditaciones del Quijote: meditación preliminar, meditación primera*. Madrid: Imprenta Clásica Española.
- Palmer, C. L., Weber, N. M., & Cragin, M. H. (2011). The analytic potential of scientific data: Understanding re-use value. *Proceedings of the American Society for Information Science and Technology*, 48(1), 1–10. <https://doi.org/10.1002/meet.2011.14504801174>
- Parker, I. (2005). *Qualitative psychology: introducing radical research. Qualitative psychology introducing radical research*. Maidenhead, England ; New York : Open University Press. Retrieved from <https://ebookcentral.proquest.com/lib/umichigan/detail.action?docID=295430>
- Parsons, M. a., Godoy, O., LeDrew, E., de Bruin, T. F., Danis, B., Tomlinson, S., & Carlson, D. (2011). A conceptual framework for managing very diverse data for complex, interdisciplinary science. *Journal of Information Science*, 37(6), 555–569. <https://doi.org/10.1177/0165551511412705>
- Pasquetto, I. V. (2018). *From Open Data to knowledge rodution: biomedical data sharing and unpredictable data reuses*. Retrieved from <https://escholarship.org/uc/item/1s1814ej>
- Pasquetto, I. V, Randles, B. M., & Borgman, C. L. (2017). On the Reuse of Scientific Data. *Data Science Journal*, 16(8), 1–9. <https://doi.org/https://doi.org/10.5334/dsj-2017-008>
- Pasquetto, I. V, Sands, A. E., Darch, P. T., & Borgman, C. L. (2016). Open Data in scientific settings. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, 1585–1596. <https://doi.org/10.1145/2858036.2858543>
- Penders, B., Horstman, K., & Vos, R. (2008). Walking the Line between Lab and Computation: The “Moist” Zone. *BioScience*, 58(8), 747. <https://doi.org/10.1641/B580811>
- Pettigrew, A. M. (1990). Longitudinal field research on change: theory and practice. *Organization Science. Special Issue: Longitudinal Field Research Methods for Studying Processes of Organizational Change*, 1(3), 267–292. Retrieved from <https://www.jstor.org/stable/2635006>
- Pinch, T. (1990). Reviewed work (s): The reflexive thesis: wrighting sociology of scientific knowledge by Malcolm Pinch. *Contemporary Sociology*, 19(6), 882–883. Retrieved from

<http://www.jstor.org/stable/2073241>

- Piwovar, Heather A, Street, W. M., & Suite, A. (2011). A method to track dataset reuse in biomedicine : filtered GEO accession numbers in PubMed Central. *October*, 1–2. <https://doi.org/10.1002/meet.14504701450>
- Piwovar, Heather Alyce. (2010). *Foundational studies for measuring the impact, prevalence, and patterns of publicly sharing biomedical research data*. <https://doi.org/10.1080/04580631003677038>
- Poole, M. S., & Van de Ven, A. (2010). Empirical methods for research on organizational decision-making processes. In P. C. Nutt & D. C. Wilson (Eds.), *Handbook of decision making* (pp. 543–580). John Wiley & Sons, Ltd.
- Popper 1902-1994, K. R. (1966). *The open society and its enemies*. (5th ed.). Princeton, N.J.: Princeton University Press, 1966.
- Porter, A. L., & Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics*, *81*(3), 719–745. <https://doi.org/10.1007/s11192-008-2197-2>
- Ragin, C. C. (1987). *The comparative method : moving beyond qualitative and quantitative strategies*. Berkeley: University of California Press.
- Ragin, C. C. (1999). Using Qualitative Comparative Analysis to Study Causal Complexity. *Health Services Research*, *34*(5 Part II), 1225–1239. Retrieved from <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1089061&tool=pmcentrez&rendertype=abstract>
- Ragin, C. C. (2012). Qualitative comparative analysis using fuzzy sets (fsQCA). In B. Rihoux & C. C. Ragin (Eds.), *Configurational Comparative Methods: qualitative comparative analysis and related techniques*. Thousand Oaks: SAGE Publications, Inc. <https://doi.org/10.4135/9781452226569>
- Reichenbach, H. (1938). *Experience and Prediction: An Analysis of the Foundations and the Structure of Knowledge*. Chicago: University of Chicago Press.
- Reichertz, J. (2014). *The SAGE Handbook of Qualitative Data Analysis*. London: SAGE Publications Ltd. <https://doi.org/10.4135/9781446282243>
- Rew, L., Koniak-Griffin, D., Lewis, M. A., Miles, M., & O’Sullivan, A. (2000). Secondary data analysis: New perspective for adolescent research. *Nursing Outlook*, *48*(5), 223–229. <https://doi.org/http://dx.doi.org/10.1067/mno.2000.104901>
- Rey, E., Garneau, P., David, M., Gauthier, R., Leduc, L., Michon, N., ... Rodger, M. (2009). Dalteparin for the prevention of recurrence of placental-mediated complications of pregnancy in women without thrombophilia: A pilot randomized controlled trial. *Journal of Thrombosis and Haemostasis*, *7*(1), 58–64. <https://doi.org/10.1111/j.1538-7836.2008.03230.x>
- Ribes, D., & Jackson, S. J. (2013). Data bite man: The work of sustaining a long-term study. In L. Gitelman (Ed.), *Raw Data is an Oxymoron* (pp. 147–166).
- Riedel, M. (2000). *Research strategies for secondary data : a perspective for criminology and criminal justice*. Thousand Oaks, Calif.: Thousand Oaks, Calif. : Sage Publications Inc., c2000.
- Riet, G. Ter, Bachmann, L. M., Kessels, A. G. H., & Khan, K. S. (2013). Individual patient data meta-analysis of diagnostic studies: Opportunities and challenges. *Evidence-Based Medicine*, *18*(5), 165–169. <https://doi.org/10.1136/eb-2012-101145>
- Rihoux, B., & Meur, G. De. (2012). Crisp-set qualitative comparative analysis (csQCA). In B. Rihoux

- & C. C. Ragin (Eds.), *Configurational comparative methods: qualitative comparative analysis (QCA) and related techniques* (pp. 33–68). <https://doi.org/10.4135/9781452226569>
- Rodger, M. A., Carrier, M., Le Gal, G., Martinelli, I., Perna, A., Rey, E., ... Gris, J. C. (2014). Meta-analysis of low-molecular-weight heparin to prevent recurrent placenta-mediated pregnancy complications. *Blood*, *123*(6), 822–828. <https://doi.org/10.1182/blood-2013-01-478958>
- Rodger, M. A., Gris, J. C., de Vries, J. I. P., Martinelli, I., Rey, É., Schleussner, E., ... Mayhew, A. D. (2016). Low-molecular-weight heparin and recurrent placenta-mediated pregnancy complications: a meta-analysis of individual patient data from randomised controlled trials. *The Lancet*, *388*(10060), 2629–2641. [https://doi.org/10.1016/S0140-6736\(16\)31139-4](https://doi.org/10.1016/S0140-6736(16)31139-4)
- Rodger, M. A., Hague, W. M., Kingdom, J., Kahn, S. R., Karovitch, A., Sermer, M., ... Wells, P. S. (2014). Antepartum dalteparin versus no antepartum dalteparin for the prevention of pregnancy complications in pregnant women with thrombophilia (TIPPS): A multinational open-label randomised trial. *The Lancet*, *384*(9955), 1673–1683. [https://doi.org/10.1016/S0140-6736\(14\)60793-5](https://doi.org/10.1016/S0140-6736(14)60793-5)
- Rodger, M. A., Langlois, N. J., de Vries, J. I. P., Rey, É., Gris, J. C., Martinelli, I., ... Kaaja, R. (2015). Low-molecular-weight heparin for prevention of placenta-mediated pregnancy complications: Protocol for a systematic review and individual patient data meta-analysis (AFFIRM). *Systematic Reviews*, *3*(1), 1–11. <https://doi.org/10.1186/2046-4053-3-69>
- Rojas, F. (2017). *Theory for the working sociologist*. New York : Columbia University Press, 2017.
- Rolland, B., & Lee, C. (2013). Beyond trust and reliability: reusing data in collaborative cancer epidemiology research. <https://doi.org/10.1145/2441776.2441826>
- Saldaña, J. (2003). *Longitudinal qualitative research : analyzing change through time*. Walnut Creek, CA : Walnut Creek, CA : AltaMira Press, 2003.
- Sayer, R. A. (2000). *Realism and social science*. London: Sage Publications.
- Sayer, R. A. (2010). *Method in social science: a realist approach* (Revised 2n). Oxon: Routledge.
- Secrist, H. (1920). *An introduction to statistical methods : a textbook for college students*. New York: Macmillan.
- Sidlauskas, B., Ganapathy, G., Hazkani-Covo, E., Jenkins, K. P., Lapp, H., McCall, L. W., ... Kidd, D. M. (2010). Linking big: The continuing promise of evolutionary synthesis. *Evolution*, *64*(4), 871–880. <https://doi.org/10.1111/j.1558-5646.2009.00892.x>
- Silverman, D. (Ed.). (2004). *Qualitative research: theory, method and practice* (2nd ed.). London, UK; Thousand Oaks, CA; New Delhi: SAGE Publications.
- Simmonds, M., Stewart, G., & Stewart, L. (2015). A decade of individual participant data meta-analyses: A review of current practice. *Contemporary Clinical Trials*, *45*, 76–83. <https://doi.org/10.1016/j.cct.2015.06.012>
- Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, *69*(1), 99–118. <https://doi.org/10.2307/1884852>
- Simon, H. A. (1957). *Models of man: social and rational. Mathematical essays on rational human behavior*. New York [u.a.: Wiley [u.a.].
- Simon, H. A. (1976). From substantive to procedural rationality, in method and appraisal in economics, ed. by Latsis S.J., (9), 573–574. Retrieved from <http://digitalcollections.library.cmu.edu/awweb/awarchive?type=file&item=33828>
- Simon, H. A. (2000). Bounded rationality in social science: today and tomorrow. *Mind & Society*,

- 1(1), 25–39.
- Smith, E. (2008). Pitfalls and promises: the use of secondary data analysis in educational research. *British Journal of Educational Studies*, 56(3), 323–339. <https://doi.org/10.1111/j.1467-8527.2008.00405.x>
- Stake, R. E. (2005). Qualitative case studies. In N. K. Denzin & Y. S. Lincoln (Eds.), *The SAGE handbook of qualitative research* (3rd ed.). Thousand Oaks: Thousand Oaks : Sage Publications, c2005.
- Stake, R. E. (2006). *Multiple case study analysis*. New York: New York : The Guilford Press, c2006.
- Stewart, L. A., & Michael, J. C. (1995). Practical methodology of meta-analyses (overviews) using updated individual patient data. *Statistics in Medicine*, 14(19), 2057–2079. <https://doi.org/10.1002/sim.4780141902>
- Strasser, B. J. (2010). Collecting, Comparing, and Computing Sequences: The Making of Margaret O. Dayhoff's Atlas of Protein Sequence and Structure, 1954-1965. *Journal of the History of Biology*, 43(4), 623–660. <https://doi.org/10.1007/s10739-009-9221-0>
- Strasser, B. J., & De Chadarevian, S. (2011). The comparative and the exemplary: Revisiting the early history of molecular biology. *History of Science*, 49(3), 317–336. <https://doi.org/10.1177/007327531104900305>
- Tenopir, C., Dalton, E., Allard, S., Frame, M., Pjesivac, I., Birch, B., ... Dorsett, K. (2015). Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. *PLoS One*, 10(8). <https://doi.org/10.1371/journal.pone.0134826>
- Thanos, C. (2017). Research Data Reusability : Conceptual Foundations , Barriers and Enabling Technologies. <https://doi.org/10.3390/publications5010002>
- Tosi, H. L., Locke, E. A., & Latham, G. P. (1991). A Theory of Goal Setting and Task Performance. *The Academy of Management Review*, 16(2), 480. <https://doi.org/10.2307/258875>
- Townley, B. (2011). *Reason's Neglect: Rationality and Organizing*. New York: Oxford University Press.
- Tricco, A. C., Ashoor, H. M., Soobiah, C., Rios, P., Veroniki, A. A., Hamid, J. S., ... Straus, S. E. (2018). Comparative Effectiveness and Safety of Cognitive Enhancers for Treating Alzheimer's Disease: Systematic Review and Network Metaanalysis. *Journal of the American Geriatrics Society*, 66(1), 170–178. <https://doi.org/10.1111/jgs.15069>
- Tricco, A. C., Soobiah, C., Lillie, E., Perrier, L., Chen, M. H., Hemmelgarn, B., ... Straus, S. E. (2012). Use of cognitive enhancers for mild cognitive impairment: Protocol for a systematic review and network meta-analysis. *Systematic Reviews*, 1(1), 1–6. <https://doi.org/10.1186/2046-4053-1-25>
- Trzesniewski, K. H., Donnellan, M. B., & Lucas, R. E. (Eds.). (2011). *Secondary data analysis an introduction for psychologists* (1st ed.). Washington, D.C.: American Psychological Association.
- Tsoukas, H. (2010). Strategic Decision Making and Knowledge: A Heideggerian Approach. In P. C. Nutt & D. C. Wilson (Eds.), *Handbook of decision making* (pp. 379–402). John Wiley & Sons, Ltd.
- UNESCO. (2005). *UNESCO World Report: Towards Knowledge Societies*.
- Van de Ven, A., & Poole, M. (1990). Methods for studying innovation development in the Minnesota innovation research program. *Organization Science*, 1(3), 313–335.
- van Walraven, C. (2010). Individual patient meta-analysis-rewards and challenges. *Journal of Clinical*

- Epidemiology*, 63(3), 235–237. <https://doi.org/10.1016/j.jclinepi.2009.04.001>
- Veroniki, A. A., Ashoor, H. M., Le, S. P. C., Rios, P., Stewart, L. A., Clarke, M., ... Tricco, A. C. (2019). Retrieval of individual patient data depended on study characteristics: a randomized controlled trial. *Journal of Clinical Epidemiology*, 113(2019), 176–188. <https://doi.org/10.1016/j.jclinepi.2019.05.031>
- Veroniki, A. A., Straus, S. E., Ashoor, H. M., Hamid, J. S., Hemmelgarn, B. R., Holroyd-Leduc, J., ... Tricco, A. C. (2016). Comparative safety and effectiveness of cognitive enhancers for Alzheimer’s dementia: Protocol for a systematic review and individual patient data network meta-analysis. *BMJ Open*, 6(1), 1–8. <https://doi.org/10.1136/bmjopen-2015-010251>
- Veroniki, A. A., Straus, S. E., Ashoor, H., Stewart, L. A., Clarke, M., & Tricco, A. C. (2016). Contacting authors to retrieve individual patient data: Study protocol for a randomized controlled trial. *Trials*, 17(1), 1–8. <https://doi.org/10.1186/s13063-016-1238-z>
- Vidaillet, B. (2009). When “Decision Outcomes” are not the Outcomes of Decisions. In G. P. Hodgkinson & W. H. Starbuck (Eds.), *The Oxford Handbook of Organizational Decision Making* (pp. 418–437). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199290468.003.0022>
- Vinck, D. (2010). *The sociology of scientific work: the fundamental relationship between science and society*. Cheltenham: Edward Elgar.
- Visser, J., Ulander, V. M., Helmerhorst, F. M., Lampinen, K., Morin-Papunen, L., Bloemenkamp, K. W. M., & Kaaja, R. J. (2011). Thromboprophylaxis for recurrent miscarriage in women with or without thrombophilia - HABENOX*: A randomised multicentre trial. *Thrombosis and Haemostasis*, 105(2), 295–301. <https://doi.org/10.1160/TH10-05-0334>
- Wallis, J. C., Wynholds, L. A., Borgman, C. L., Sands, A., & Traweek, S. (2012). Data, Data Use, and Inquiry: A new point of view on data curation, (January).
- Weick, K. E. (1993). The Collapse of Sensemaking in Organizations: The Mann Gulch Disaster. *Administrative Science Quarterly*, 38(4), 628–652. <https://doi.org/10.2307/2393339>
- Weick, K. E. (1995). *Sensemaking in organizations*. Thousand Oaks, Calif.: Sage Publications.
- Welch, V. A., Ghogomu, E., Hossain, A., Awasthi, S., Bhutta, Z. A., Cumberbatch, C., ... Wells, G. A. (2017). Mass deworming to improve developmental health and wellbeing of children in low-income and middle-income countries: a systematic review and network meta-analysis. *The Lancet Global Health*, 5(1), e40–e50. [https://doi.org/10.1016/S2214-109X\(16\)30242-X](https://doi.org/10.1016/S2214-109X(16)30242-X)
- Welch, V. A., Ghogomu, E., Hossain, A., Riddle, A., Gaffey, M., Arora, P., ... Wells, G. (2019). Mass deworming for improving health and cognition of children in endemic helminth areas: A systematic review and individual participant data network meta-analysis. *Campbell Systematic Reviews*, 15(4). <https://doi.org/10.1002/cl2.1058>
- Yakel, E., Faniel, I. M., Kriesberg, A., & Yoon, A. (2013). Trust in Digital Repositories. *International Journal of Digital Curation*, 8(1), 143–156. <https://doi.org/10.2218/ijdc.v8i1.251>
- Yin, R. K. (2003). *Case study research: design and methods* (3rd ed.). Thousand Oaks, Calif.: Sage Publications.
- Yoon, A. (2014a). End users’ trust in data repositories: Definition and influences on trust development. *Archival Science*, 14(1), 17–34. <https://doi.org/10.1007/s10502-013-9207-8>
- Yoon, A. (2014b). “Making a square fit into a circle”: Researchers’ experiences reusing qualitative data. *Proceedings of the ASIST Annual Meeting*, 51(1). <https://doi.org/10.1002/meet.2014.14505101140>

- Yoon, A. (2016a). Data reusers' trust development. *Journal of the Association for Information Science and Technology*, (May). <https://doi.org/10.1002/asi.23730>
- Yoon, A. (2016b). Red flags in data: Learning from failed data reuse experiences. *Proceedings of the Association for Information Science and Technology*, 53(1), 1–6. <https://doi.org/10.1002/pra2.2016.14505301126>
- Yoon, A. (2017). Role of communication in data reuse. *Proceedings of the Association for Information Science and Technology*, 54(1), 463–471. <https://doi.org/10.1002/pra2.2017.14505401050>
- Yoon, A., & Kim, Y. (2017). Social scientists' data reuse behaviors: Exploring the roles of attitudinal beliefs, attitudes, norms, and data repositories. *Library and Information Science Research*, 39(3), 224–233. <https://doi.org/10.1016/j.lisr.2017.07.008>
- Young, W. B., & Ryu, H. (2000). Secondary Data for Policy Studies: Benefits and Challenges. *Policy, Politics, & Nursing Practice*, 1(4), 302–307. <https://doi.org/10.1177/152715440000100408>
- Zimmerman, A. S. (2003). *Data sharing and secondary use of scientific data: experiences of ecologists*.
- Zimmerman, A. S. (2007). Not by metadata alone: The use of diverse forms of knowledge to locate data for reuse. *International Journal on Digital Libraries*, 7(1–2), 5–16. <https://doi.org/10.1007/s00799-007-0015-8>
- Zimmerman, A. S. (2008). New knowledge from old data: the role of standards in the sharing and reuse of ecological data. *Science, Technology & Human Values*, 33(5), 631–652. <https://doi.org/10.1177/0162243907306704>

Appendix

Annex 1 – Interview script vs3 IAB 26 Nov 2016

INTERVIEW TO [pseudonym]; DATE of the interview:

Case study #/name of the study	
Research project:	
Research data:	
Month and year of birth:	
Role or position in the research project:	
Role in the process of reusing data:	

Introduction and warm up conversation + give the participant the informed consent¹ duly signed by MIAB.

Good morning/afternoon.

First of all, I want to thank you for your time and collaboration. I do really appreciate that you have agreed to participate in this research.

Second, I would like to remind you that this research is part of my PhD dissertation. With this research I want to understand the context and factors that exist in each of the stages of the process of reusing data in order to answer a new research question. More specifically I am interested in knowing about the types of decisions and negotiations that your research team made in that context, and how your research team managed to balance simultaneously challenges and motivations when reusing data.

Here is your consent form duly signed by me (here I will hand in the consent form) As you have agreed to voice-record this interview, I will turn on the voice recorder now if you do not mind / (alternative message: As you prefer this interview not being voice-recorded, I will not turn on the voice recorder and I will take notes of the information you provide to me)

Do you have any question or request with regard to this research so far?

(if yes, I will answer him/her question or please him/her with the request provided that the request)

¹ MIAB would have obtained the informed consent prior to this first interview as explained in the REB application

Well, let me explain you a little bit about this interview. For practical reasons, I have divided the interview in five parts based on what I think is a chronological sequence of the process of reusing data. Although I think that this structure may be useful for you to think about the different stages, it may happen that this structure does not correspond to how things happened. However, it is very important for this study that you tell me about the exact chronological order of how things happened. So, in each part of the interview, you can provide me with this information.

This is the way I have structured the interview: The first part of this interview is about the research question which required the reuse of the data; the second part refers to the moment when you realized that you needed the data and how you looked for them (data awareness and seeking process). The third one is related to the steps that you carried out to access the data. The fourth part is about the process of making sense of the data - of understanding them - to make sure that you can answer the research question with them. And finally, the fifth part refers to the use of the data, meaning the novel application of the data for your research question.

In each of the five parts of the interview I would like you to tell me about how decisions were taken, what type of decision and agreements you had to make with your colleagues, but also with other people outside your research team, for example, the people/institution who originally collected the data. Also, I am very interested in knowing what type of challenges, and motivations you found in each of the steps, and why and how you managed to overcome them. I will be reminding you about these issues in each part of the interview.

Before we start, would you like to ask something or make any comment at this stage?

By the way, most –if not all – my questions refer to “you”, but I do mean “your research team”. It is also important for this study that I know “who” exactly did “what”, so feel free to add as much as detail as you can.

PART 1 – “ABOUT THE RESEARCH QUESTION/PROJECT”

So, I would like you to start telling me about the research project and the specific research question which triggered the reuse of the data. I am interested in the following things, but also in others that you may find interesting to mention:

- 1. What’s –in very general terms– the research project about?*
- 2. What was the initial research question?*

-
3. *What is the relationship between the research project and the research question? I mean, how far does the research question accomplish the goals of the research project?*
 4. *Did the research question change at any point? If so, why? And what were the changes? // If it did not change, why not?*
 5. *Who participates in the research project (=answering the research question)?*
 6. *What is the documentation related to the research question?: any ethics application, any funding application, e-mail messages between the research team members?*
 7. *When you thought of the research question, did you know at that moment that you were going to use these data collected by others?*

PART 2 – “DATA AWARENESS AND SEEKING PROCESS”

Now, let’s talk about the moment when you became aware that you needed these specific data, and how you looked for them.

8. *Why did the team decide to use existing data?*
9. *How you decide to use existing data? (one person’s decision? Team’s decision?)*
10. *How and when did you know that these data existed?*
11. *How did you know that these data would fit your research question?*
12. *When exactly did you know that these data would fit your research question?*
13. *Who of your team (and other stakeholders) was involved in this stage?*
14. *What were the main challenges that your team encounter when seeking the data? And how did you manage to overcome them?*

PART 3 – “THE PROCESS OF ACCESSING THE DATA”

Thanks for the information provided so far. In this part –the third one– I am interested in knowing how your research team gained access of the data. Let’s start then:

15. *How did you get access to the data?*
16. *Who of the research team was involved in this step? What other stakeholders were involved?*
17. *Did the original owners/producers (or anyone else) of the data establish some conditions for accessing and reusing it?*
 - 17.1. *What were the conditions?*

17.2. *Did you accept these conditions or was there some negotiation to change them?*

18. *Was there any type of conflict in the process of accessing the data?*

19. *What were the main challenges that you encounter when accessing the data?*

PART 4 – “UNDERSTANDING THE DATA”

Let's move to the fourth part of the interview. In this part I would like you to tell me about the process of understanding the data. I guess that the process of understanding the data is an on-going process which starts at a very early moment when the research team decides to access the data or even before. Anyway, these are my questions. Feel free to tell me if they do not make sense to you.

20. *Can you describe the process of understanding the data? = How did you gain understanding of the data along the whole process?*

21. *When did you start to understand the data in a way that you could know they would fit your research question?*

22. *What type of understanding you had of the data when you decided to access or request them?*

23. *Who was part of this process?*

23.1. *Who in the research team fully understood the data?*

23.2. *Did you contact any other stakeholders or experts to understand these data?*

PART 5 – “USE OF THE DATA”

We have come to the fifth and last part of the interview. Let's talk a little bit about the use you made of the data so that you could answer your research question.

24. *What in your research question is different to the original research question which motivated the collection of the data? (=what is novel in the use of the data you have made)*

25. *When did you know about the type of use (analysis) you would use with the data?*

26. *Who took part in the analysis/novel application of the data?*

27. *Was there any type of disagreement between the research team about how to proceed with the use and application of the data?*

28. *What types of challenges did you encounter when using the data, and how did you overcome them?*

Wrapping up, thanking, and ending the session

Thank you so much for all this valuable information. I do really appreciate your time and your efforts in remembering all these issues. I will transcribe the interview as soon as I can, and I will also do a preliminary analysis. Option a) As you have requested the transcript of the interview, I will share it with you from my uOttawa drive as soon as I have transcribed it. Option b) You have chosen not to have the transcription of this interview, but if you change your mind, please let me know. I will share it with you without hesitation.

After making the analysis of this information you have provided me I might contact you if I have any doubt or I need further information from you.

If, meanwhile, you remember something which you think it is relevant to what you have told me today I would appreciate if you could let me know by email.

Thank you so much again for your collaboration. I wish you a nice day.

Case study #1: *Chromatin regulators: jacks of all states* (Selection criteria: *released data*)

Data collection instruments in 5 stages and a simplified chronology of interactions with **one participant**

STAGE 1



Preliminary face-to-face meeting and emails with participant to check eligibility criteria

Web site of GTEX portal (<https://gtexportal.org/home/>)

Check participant's online scientific profile

STAGE 2



First interview (51'24")

STAGE 3



Analysis of first interview

Drawing of the workflow diagram of the process of reusing data

STAGE 4



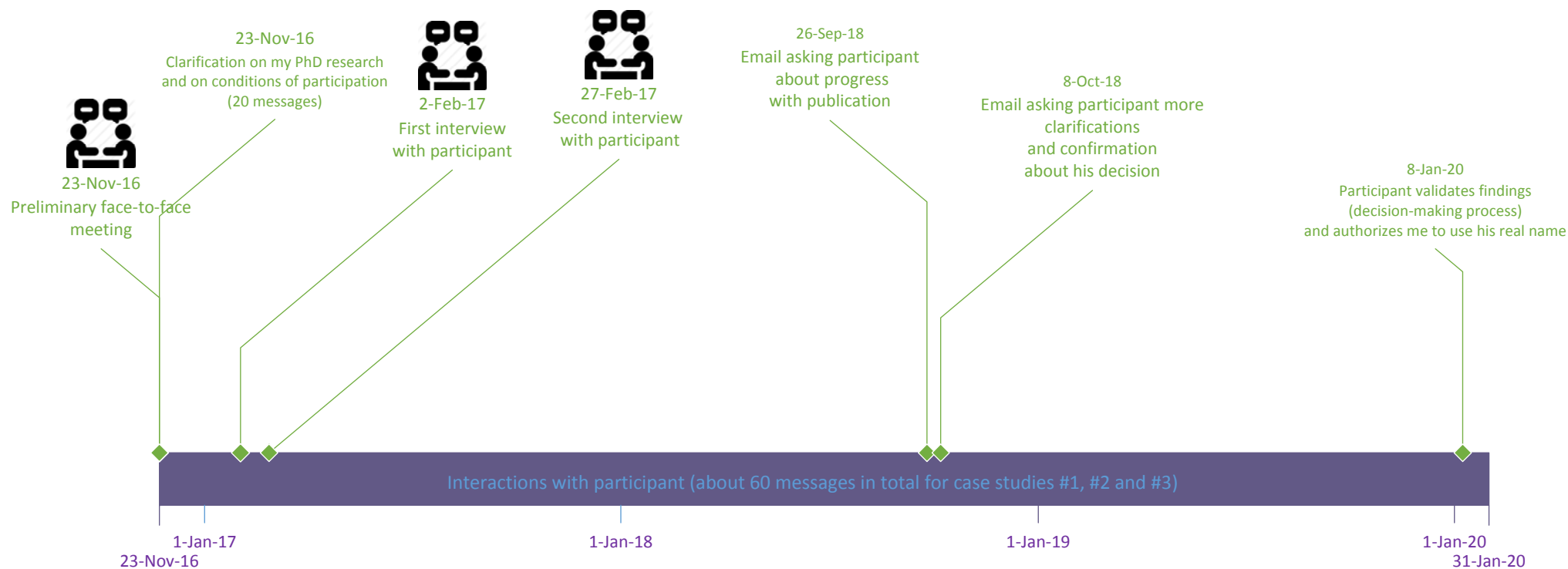
Second interview (36'23") and validation of the workflow diagram of the process

STAGE 5



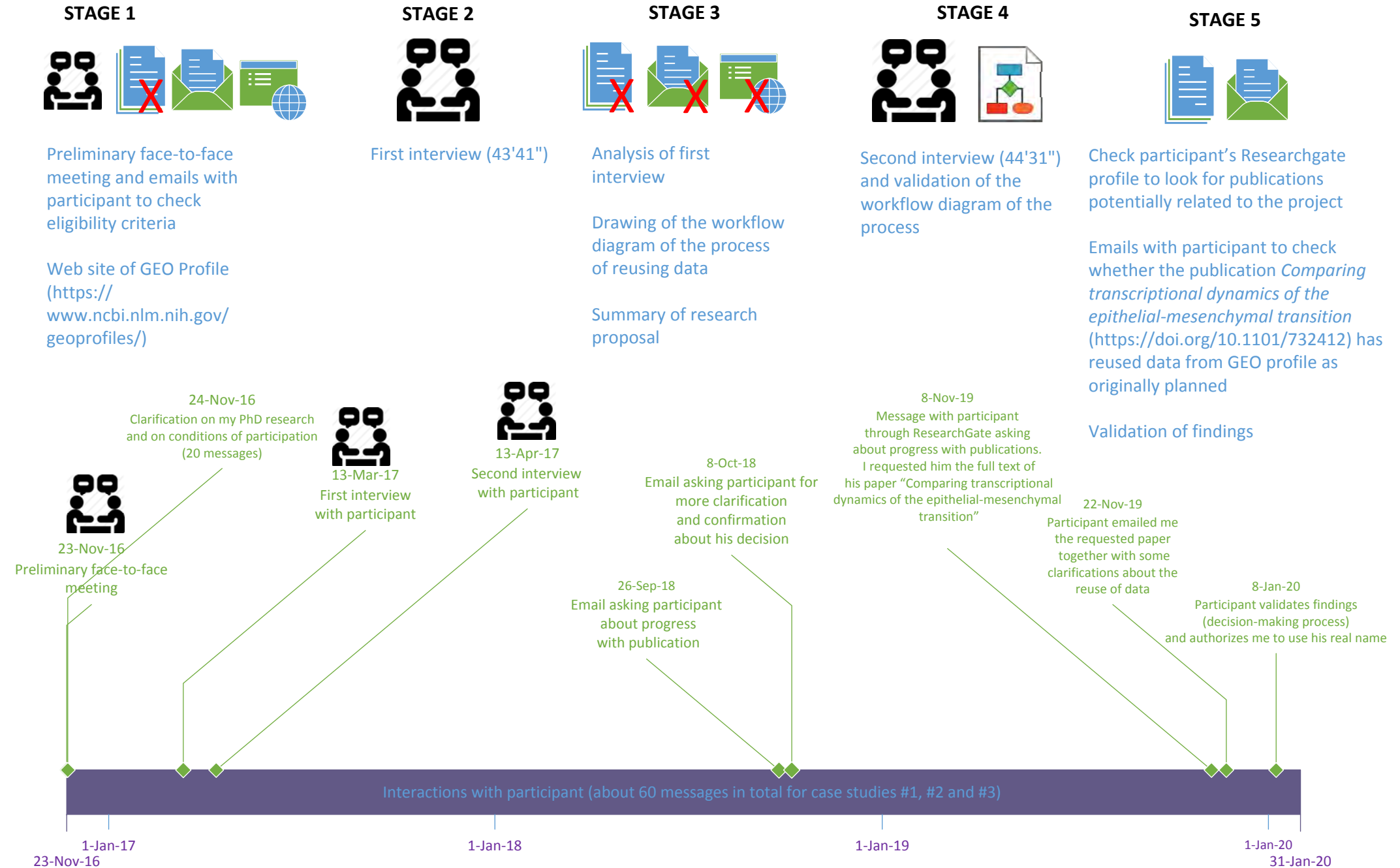
Emails with participant to find out the outcome of his process of reusing data

Validation of findings



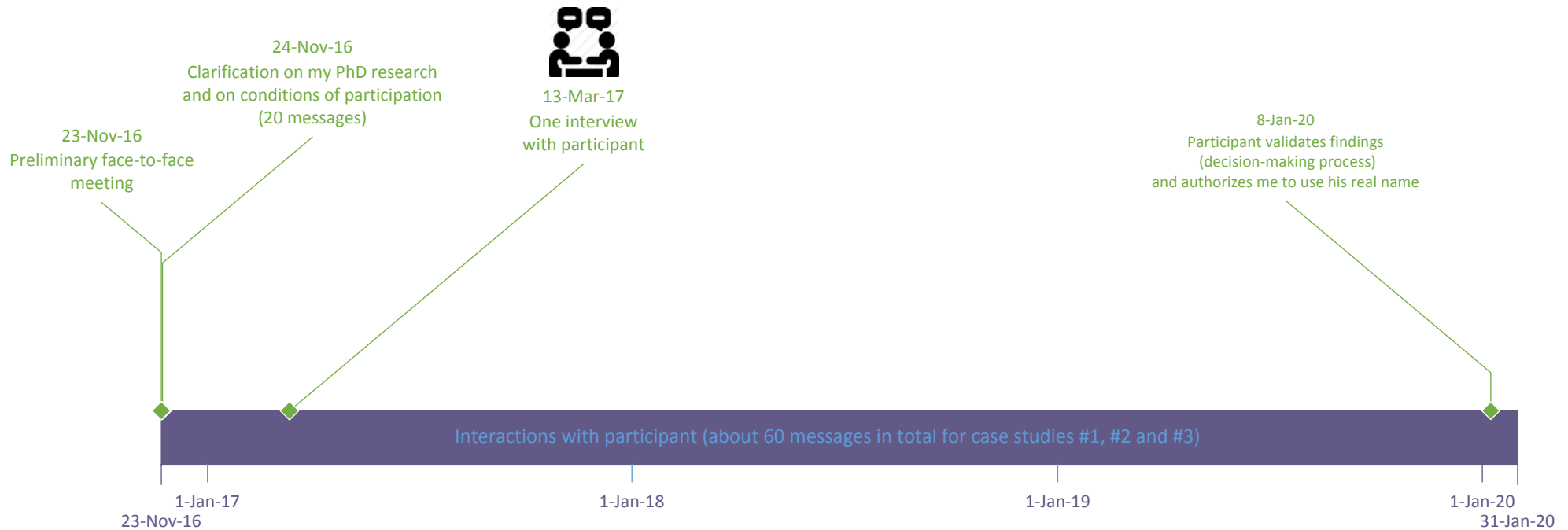
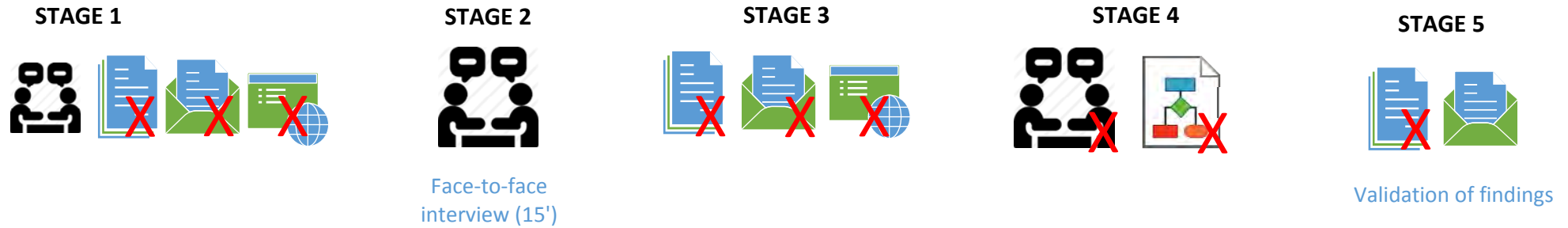
Case study #2: Transcriptional and epigenetic determinants of the epithelial-to-mesenchymal transition in ovarian cancer (Selection criteria: *released data*)

Data collection instruments in 5 stages and a simplified chronology of interactions with **one participant**



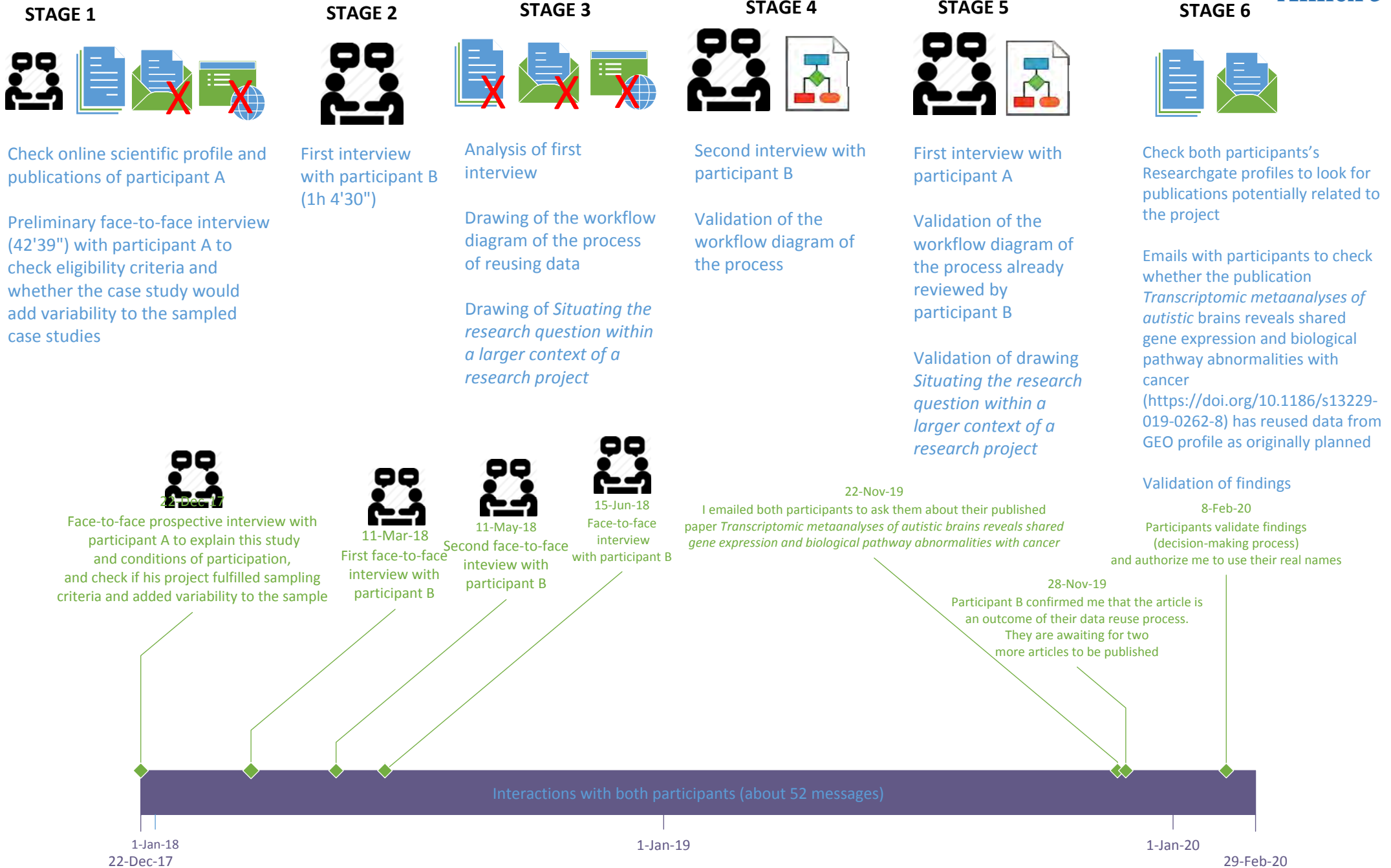
Case study #3: A case of non-reuse after having started the process of reusing data. (Selection criteria: *released data*)

Data collection instruments in 5 stages and a simplified chronology of interactions with **one participant**



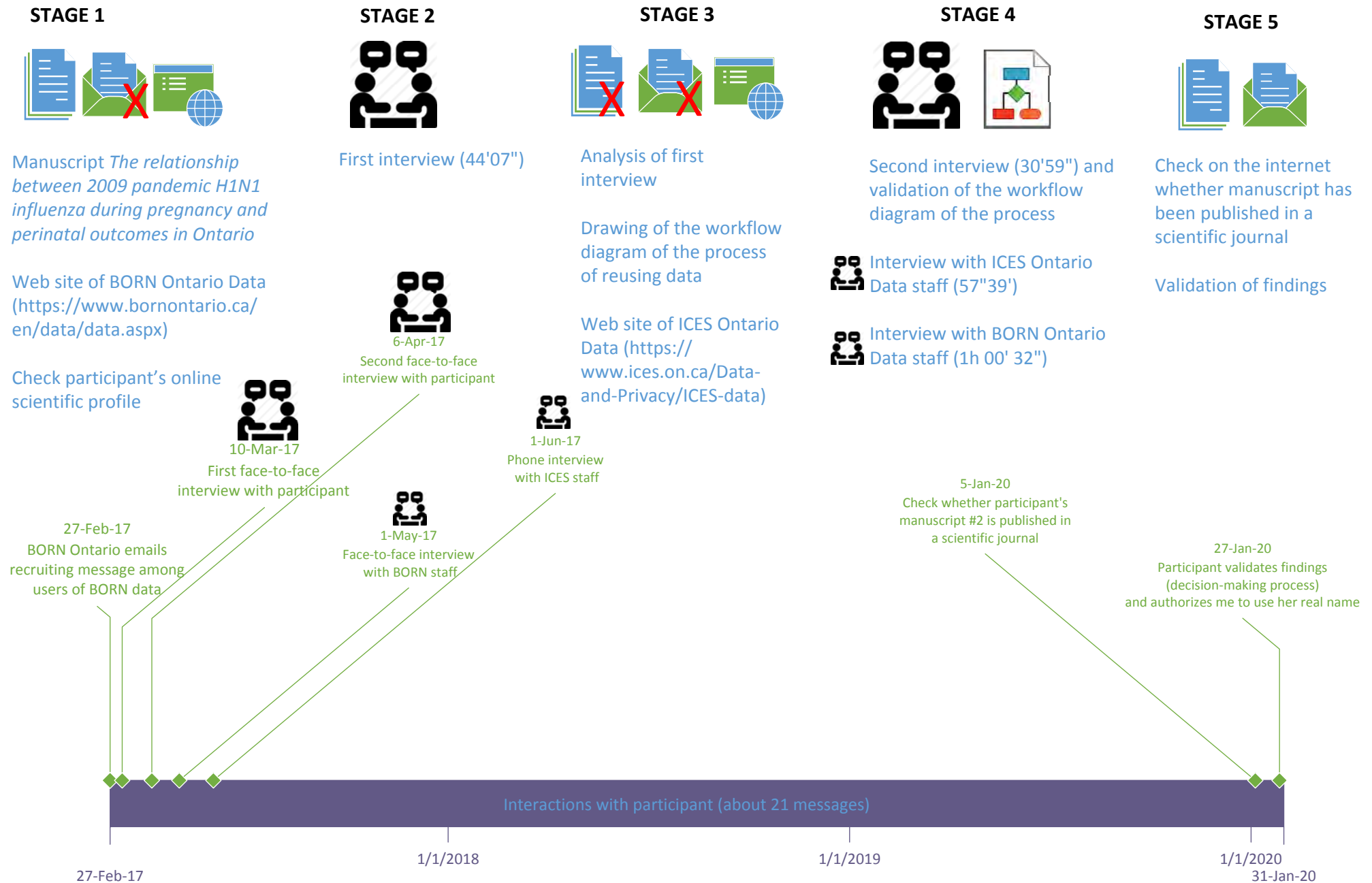
Case study #4: Breast cancer (gene and biological relationship between autism spectrum and cancer; the role of TRIM29) (Selection criteria: *released data*)

Data collection instruments in 6 stages and a simplified chronology of interactions with **two participants**



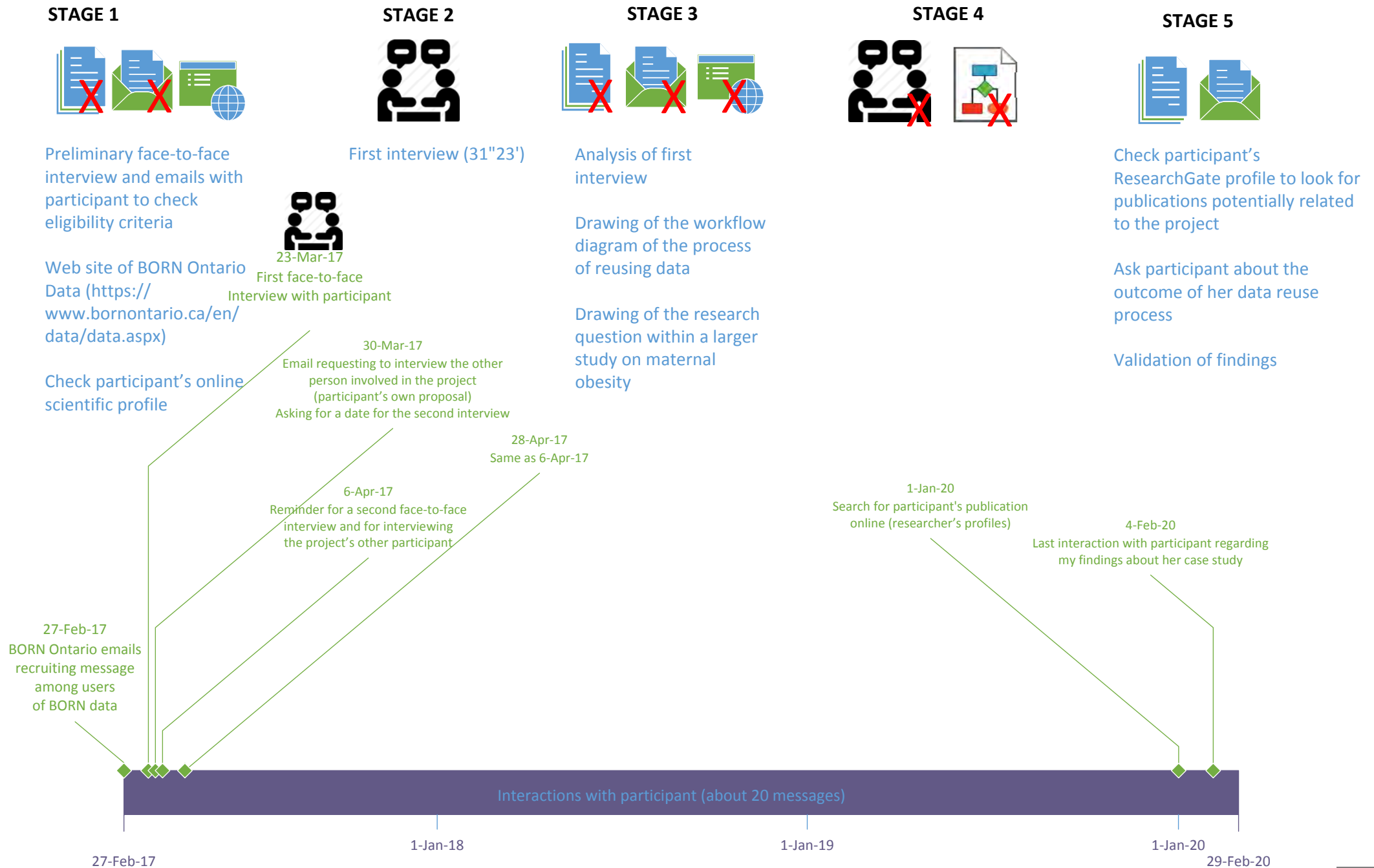
Case study #5: The relationship between 2009 pandemic H1N1 influenza during pregnancy and perinatal outcomes in Ontario, manuscript 2 of PhD thesis titled *Influenza illness and influenza vaccination during pregnancy and risk of preterm birth and fetal death, December 2015* (Selection criteria: *stewarded data*)

Data collection instruments in 5 stages and a simplified chronology of interactions with **one participant**



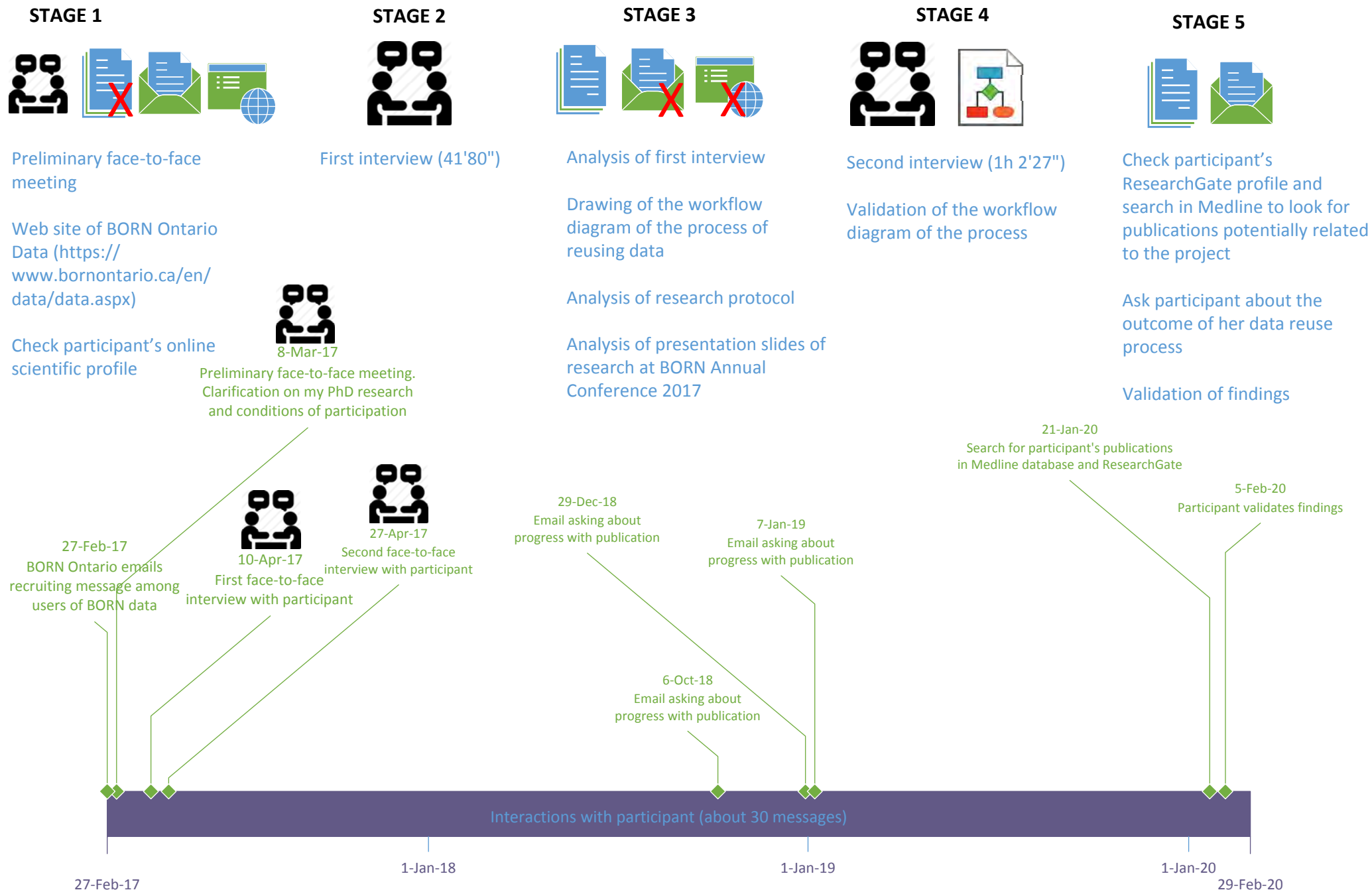
Case study #6: The effect of maternal obesity on stillbirth and neonatal death (Selection criteria: *stewarded data*)

Data collection instruments in 5 stages and a simplified chronology of interactions with 1 participant



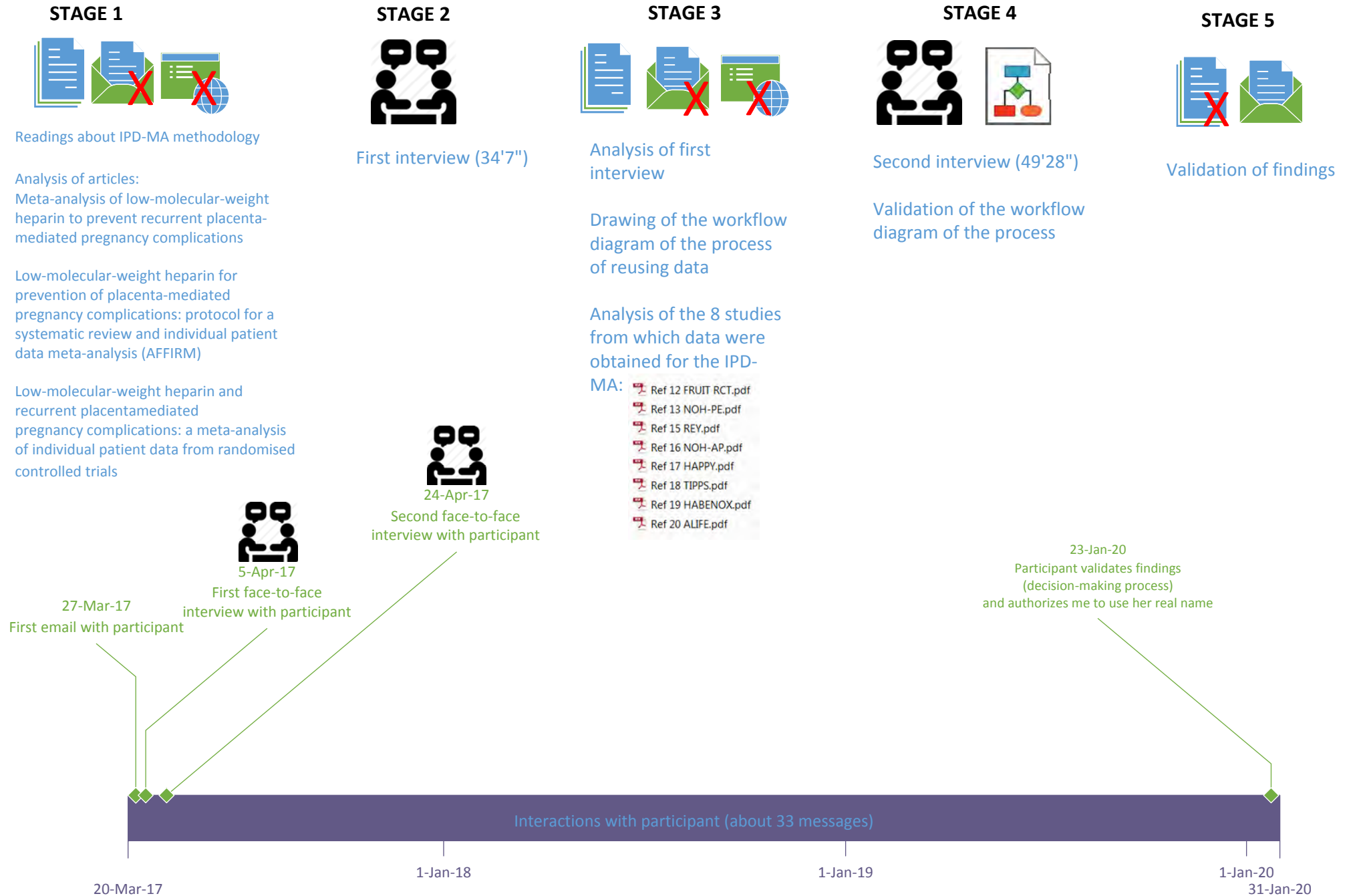
Case study #7: A research study in epidemiology related to maternal and neonatal health (Selection criteria: *stewarded data*)

Data collection instruments in 5 stages and a simplified chronology of interactions with 1 participant



Case study #8: Low-molecular-weight heparin and recurrent placenta-mediated pregnancy complications: a meta-analysis of individual patient data from randomised controlled trials ([http://dx.doi.org/10.1016/S0140-6736\(16\)31139-4](http://dx.doi.org/10.1016/S0140-6736(16)31139-4)) (Selection criteria: *proprietary data*)

Data collection instruments in 5 stages and a simplified chronology of interactions with 1 participant



STAGE 1



Preliminary emails with participant to check eligibility criteria

Check participant's online scientific profile

STAGE 2



First interview (50'80")

STAGE 3



Analysis of first interview

Drawing of the workflow diagram of the process of reusing data

Analysis of article *Mass deworming to improve developmental health and wellbeing of children in low-income and middle-income countries: a systematic review and network meta-analysis* (2017)

Analysis of Data Transfer Agreement and Terms of Reference

STAGE 4



Second interview (37'24")

Validation of the workflow diagram of the process

STAGE 5

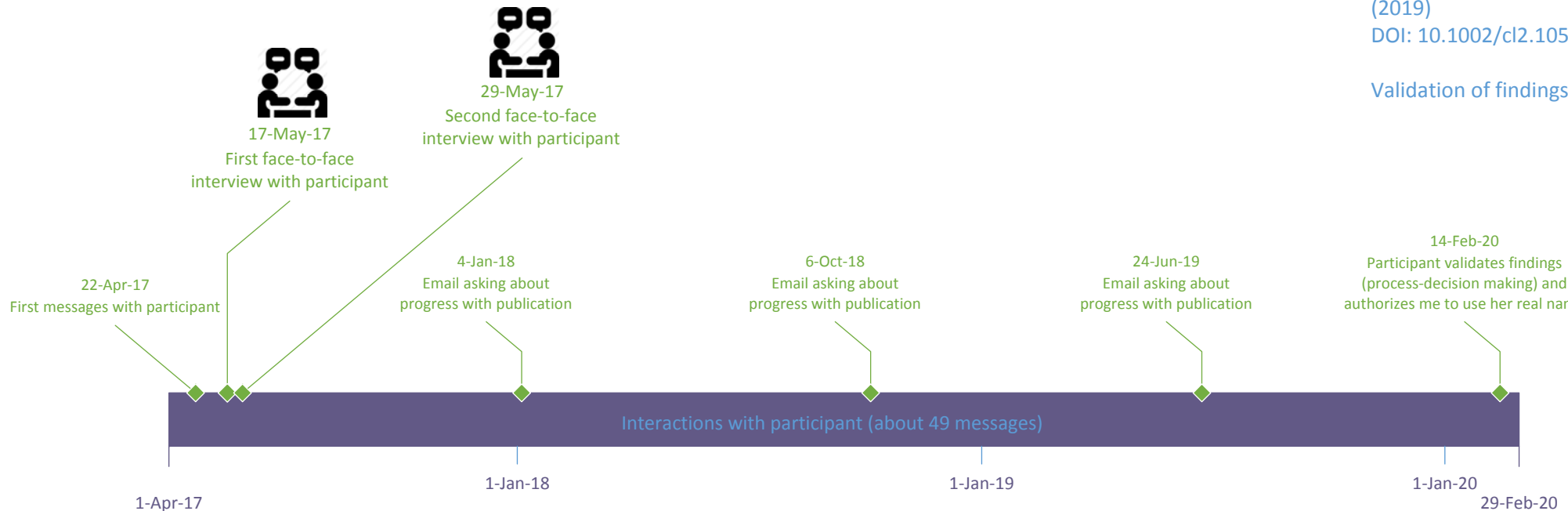


Emails with participant to check the actual reuse of the data and whether the reuse has ended up in publication

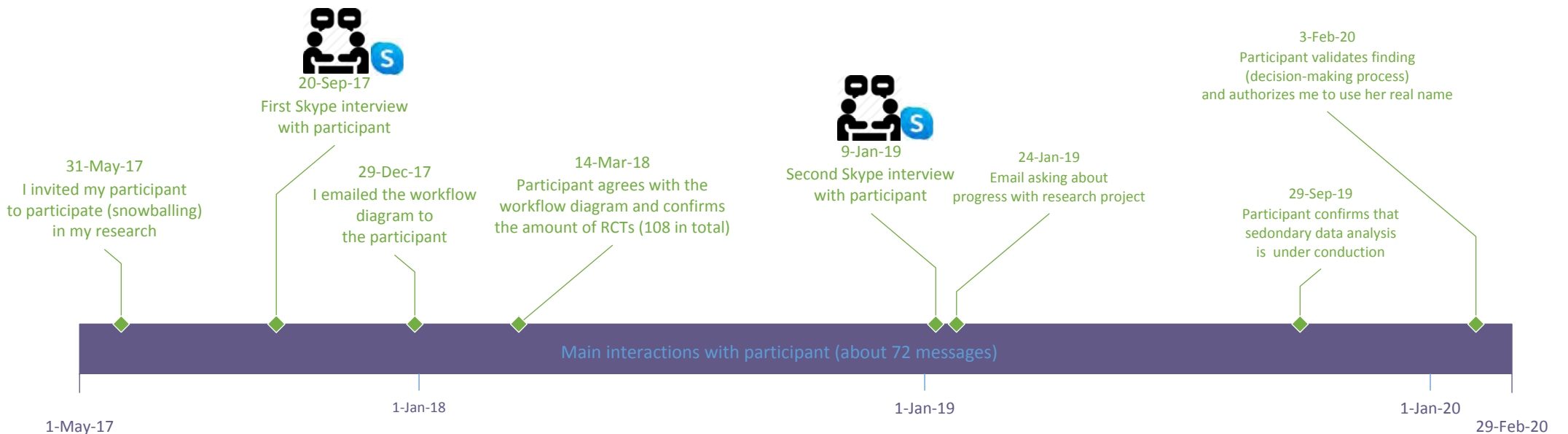
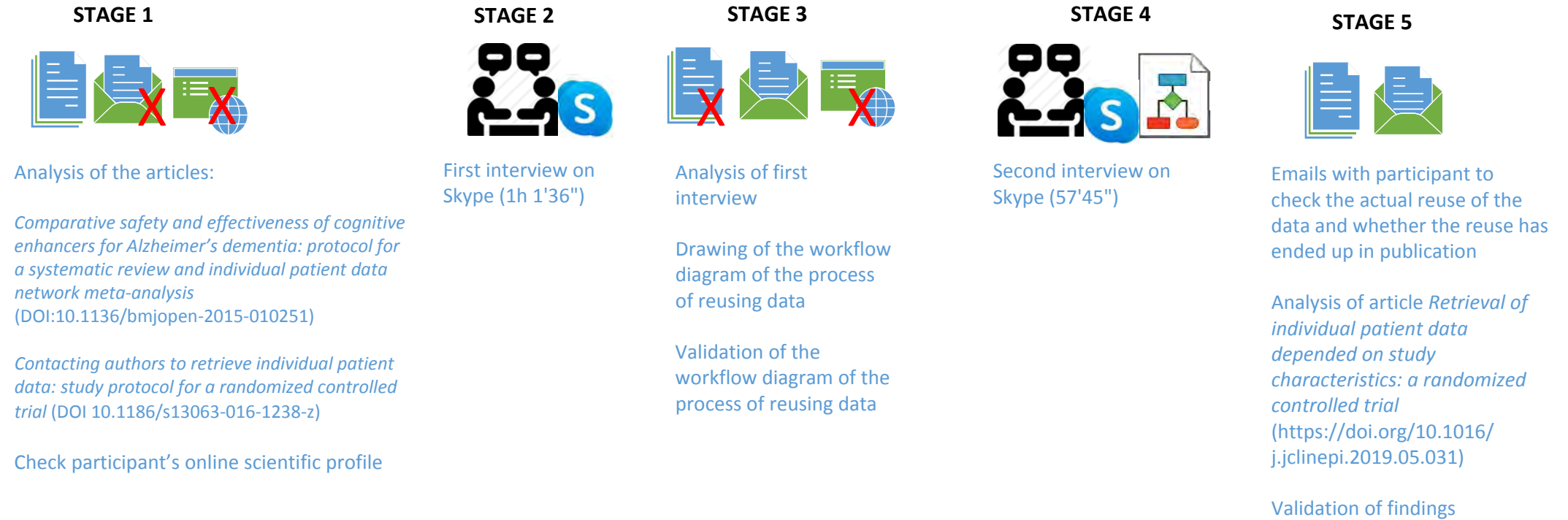
Mass deworming for improving health and cognition of children in endemic helminth areas: A systematic review and individual participant data network meta-analysis (2019)

DOI: 10.1002/cl2.1058

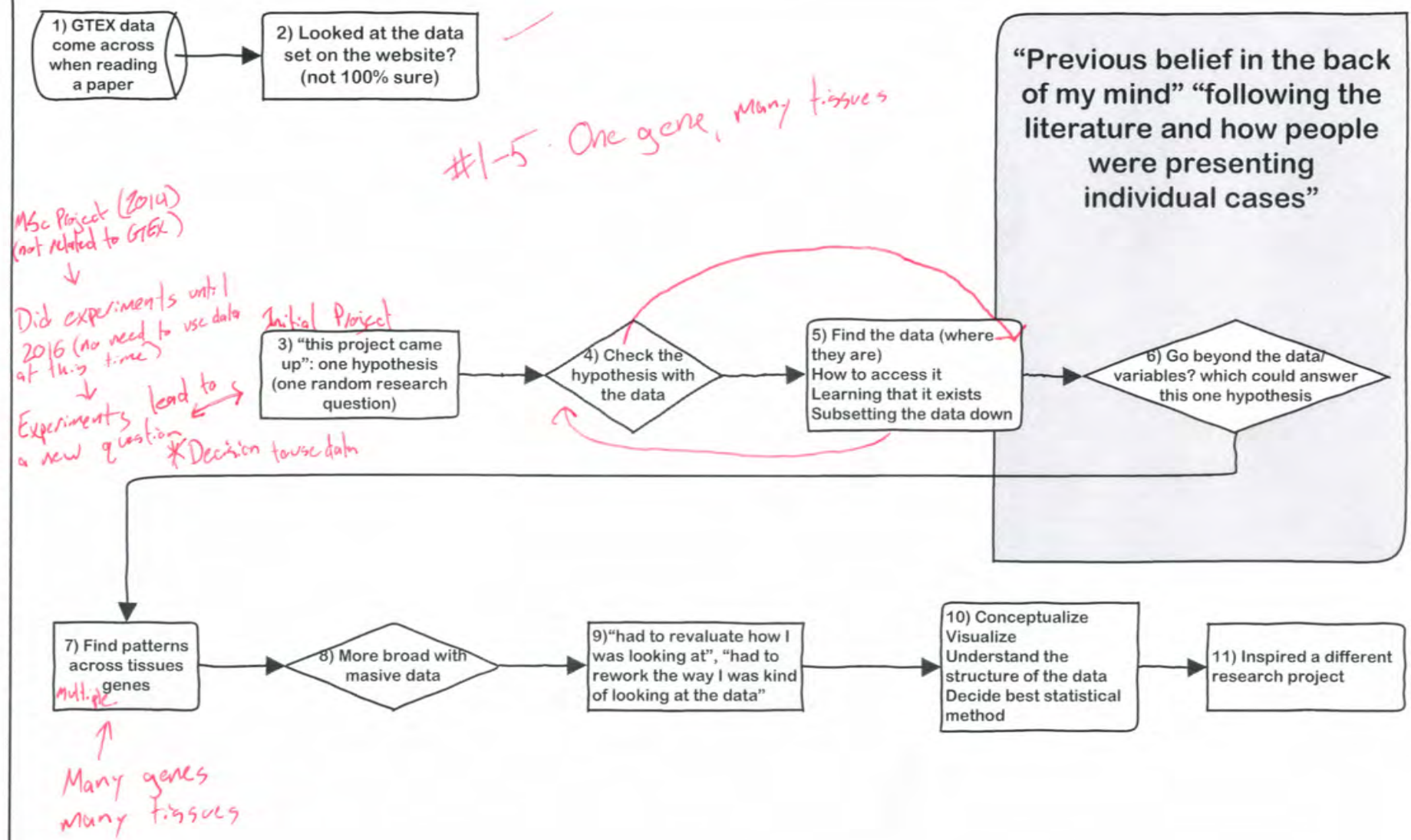
Validation of findings



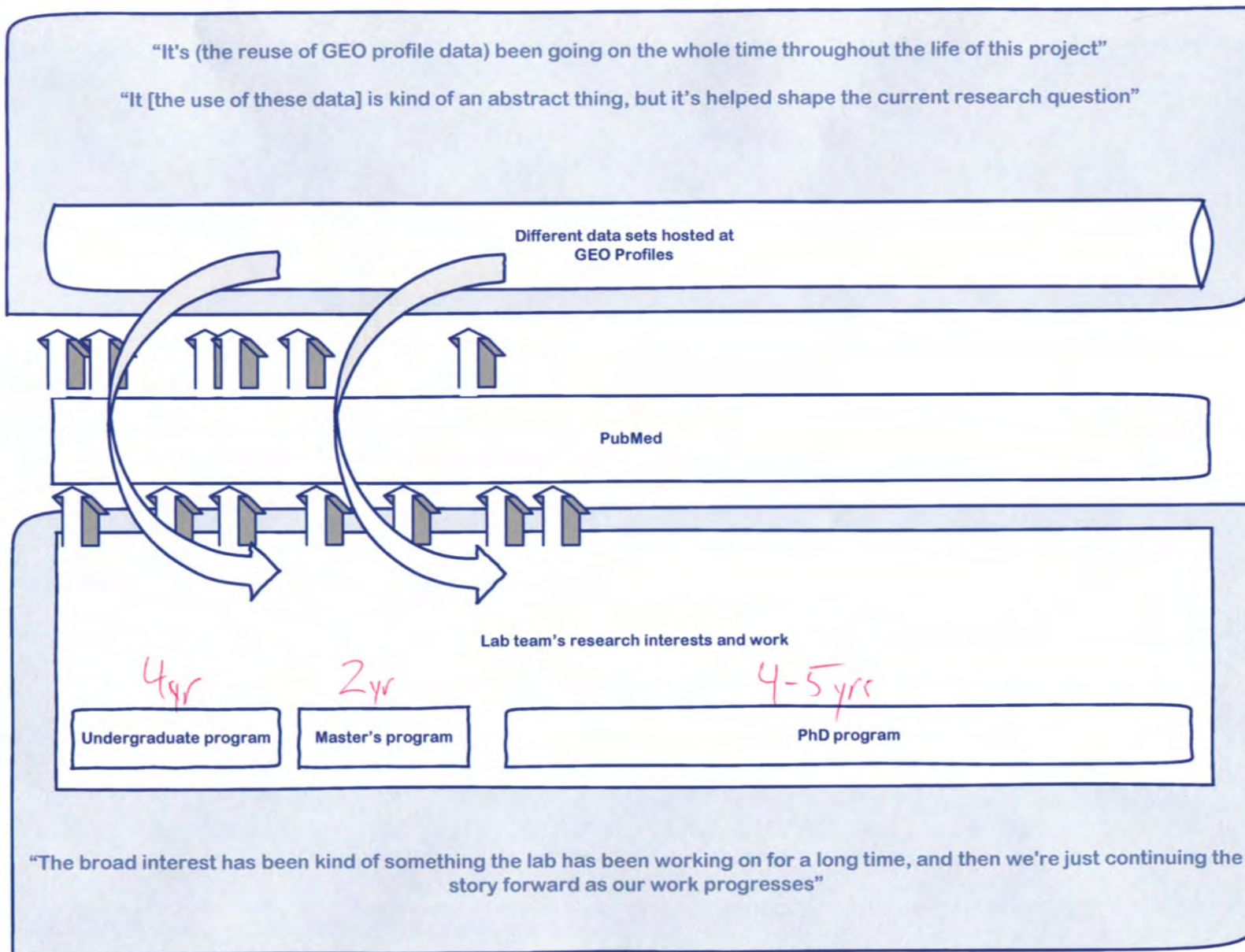
Data collection instruments in 5 stages and a simplified chronology of interactions with 1 participant



Process of reusing GTEX data



Process of reusing data sets of GEO Profiles
<https://www.ncbi.nlm.nih.gov/geoprofiles/>



Process of reusing data in GEO profiles

1) GEO profiles repository
 "I knew this repository from my master's program in Bioinformatics. I worked with it"

"This research was something that J. (his colleague) had already in mind because J. had preliminary results from his own previous research. And so, this (research question) came up" X

(2013-)

TCGA data
 "I also created a data base with MySQL with expression data from TCGA, but finally we decided to use only data from GEO"

2) Visit Geo profiles website

3) Select the commercial platform of arrays that we need

4) Download all selected data sets

5) Decide which of the retrieved studies (data sets) fit our needs

6) Apply an statistical quality control on the arrays of each study or data set

7) Disregard samples within data sets because of lack of quality or because we do not need them

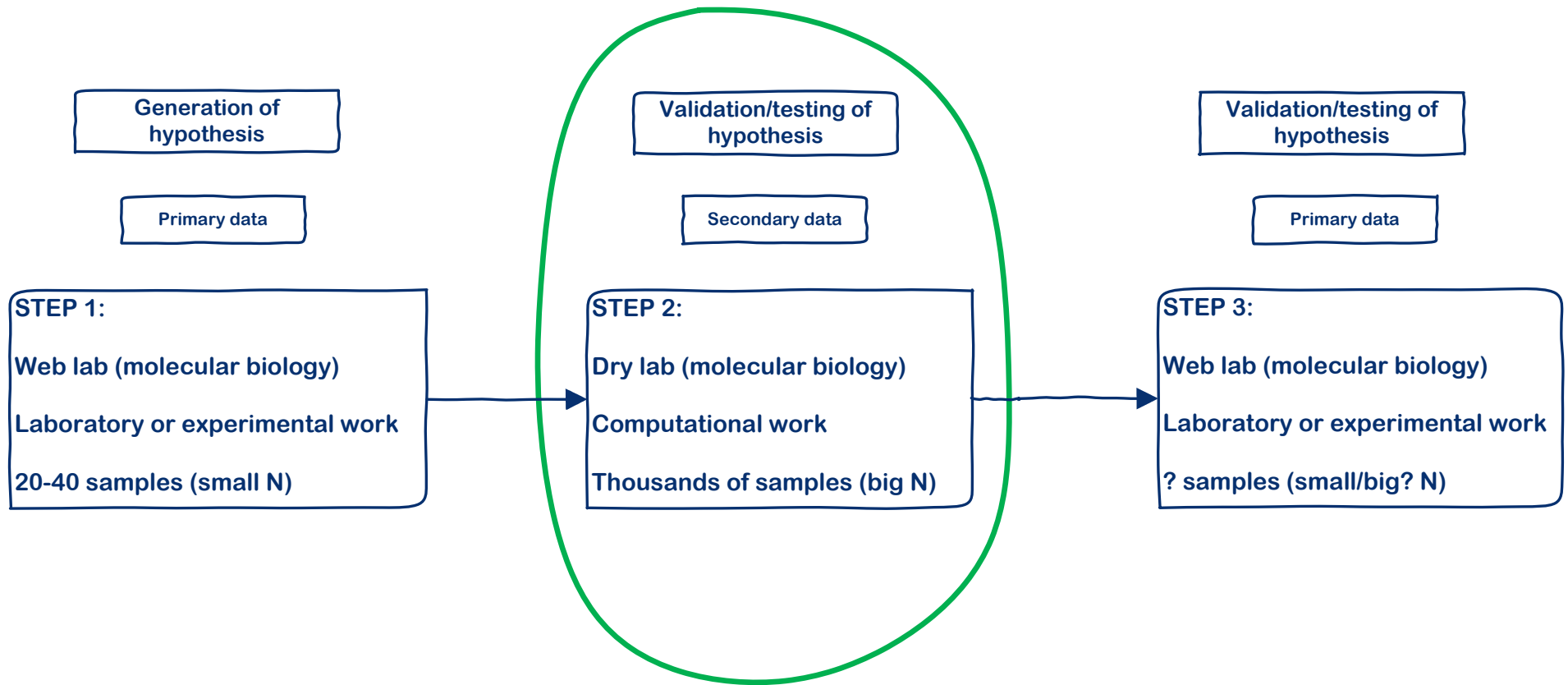
8) Process all data sets with the necessary methodologies for obtaining results

Going on...

- ANALISIS STATISTICO

2013

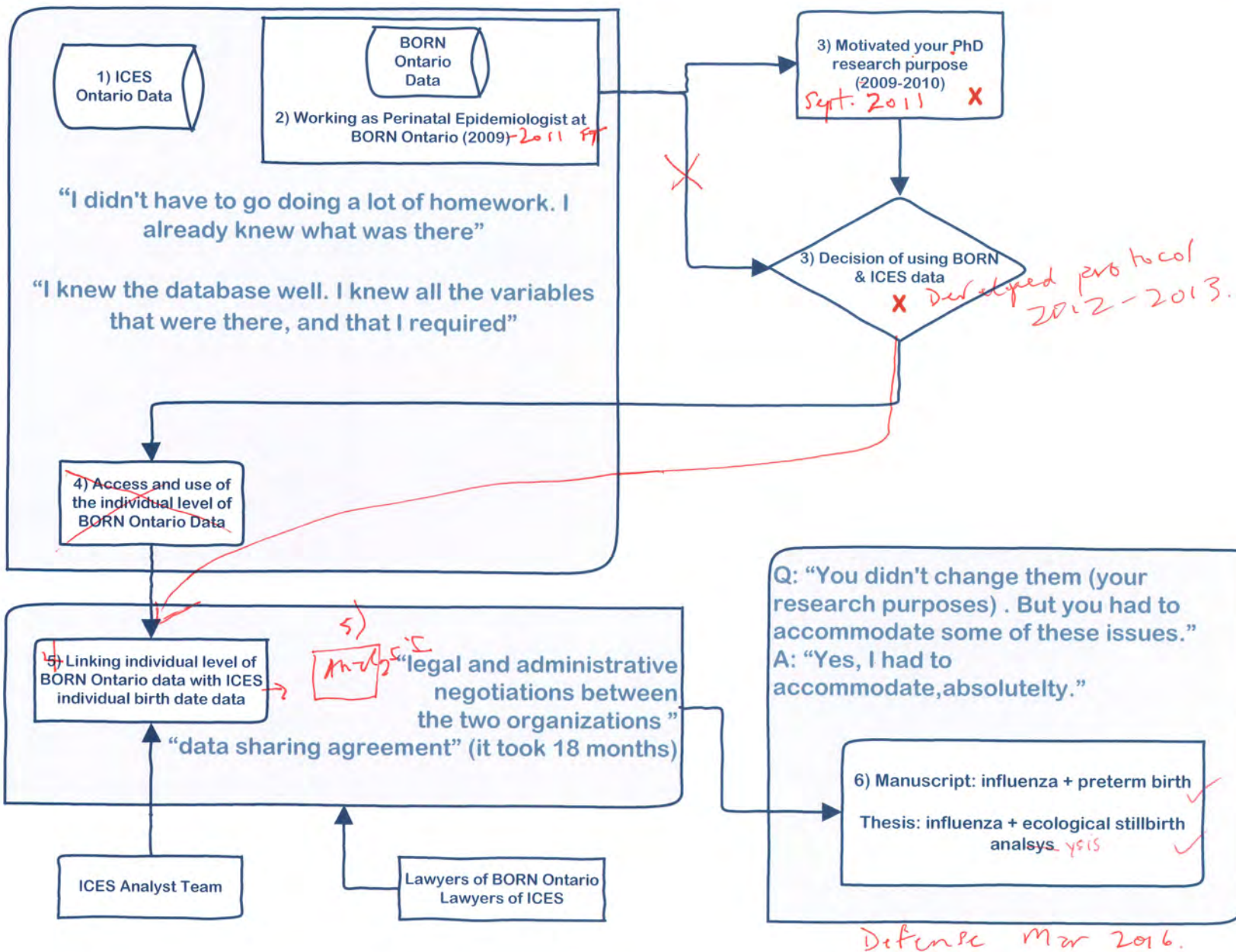
Situating the research question within a larger context or “research project”

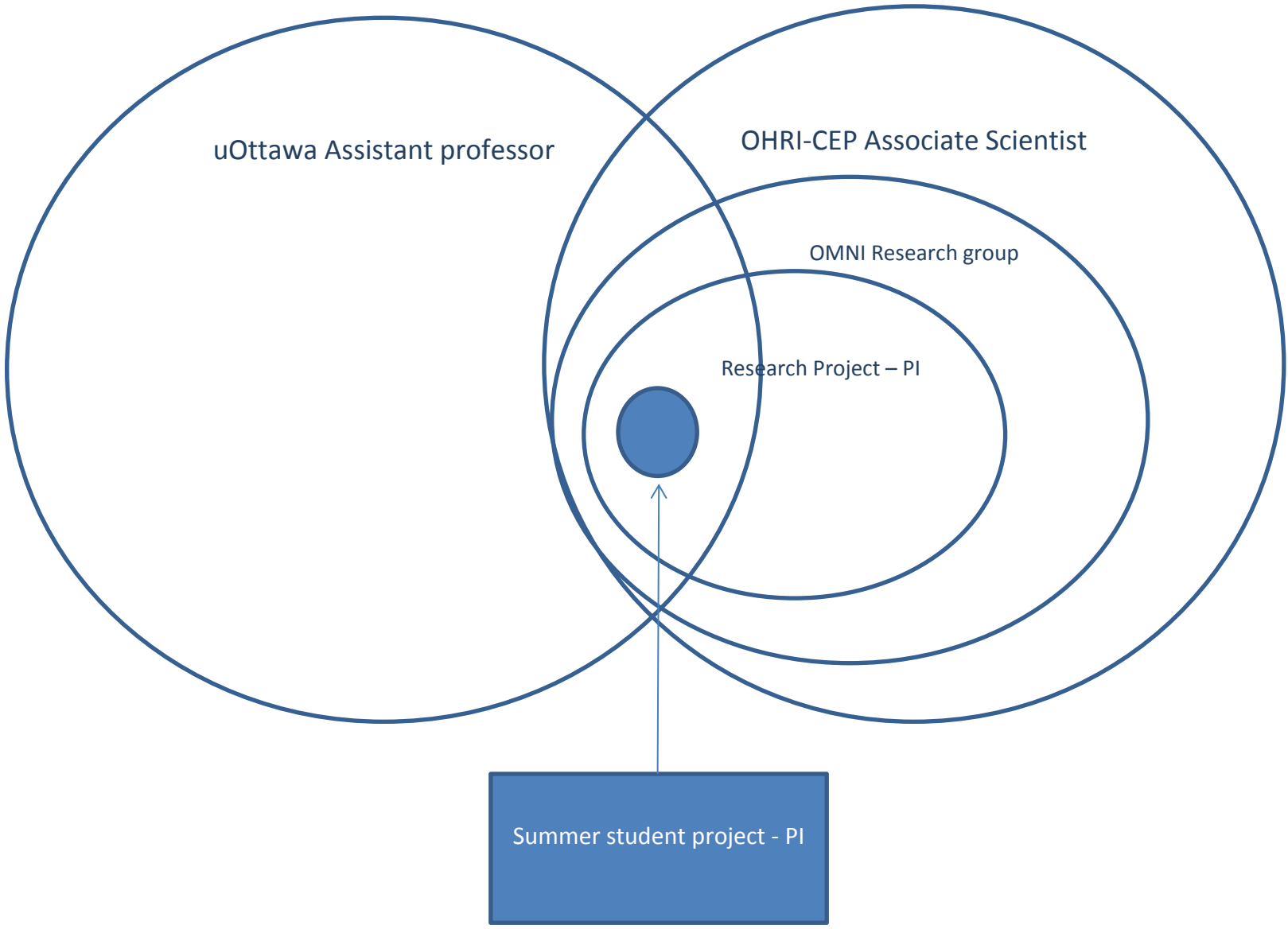


PhD thesis "Influenza illness and influenza vaccination during pregnancy and risk of preterm birth and fetal death"
 Chapter 4 – Manuscript 2 "The relationship between 2009 pandemic H1N1 influenza during pregnancy and perinatal outcomes in Ontario"

Process

Process of using BORN Ontario Data and ICES data



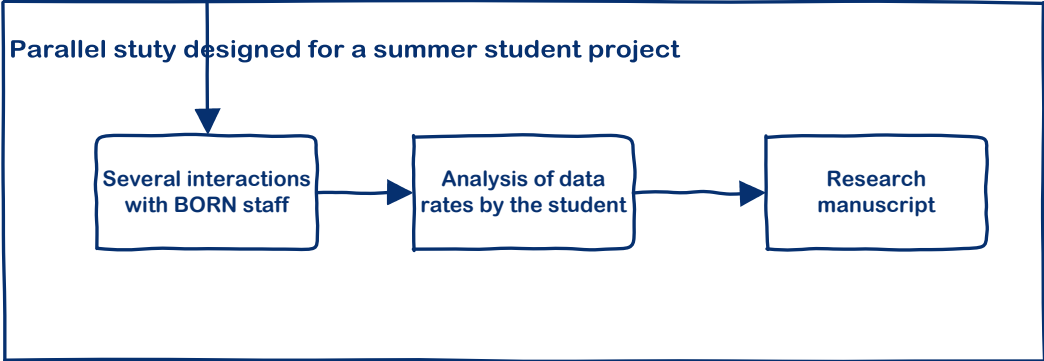
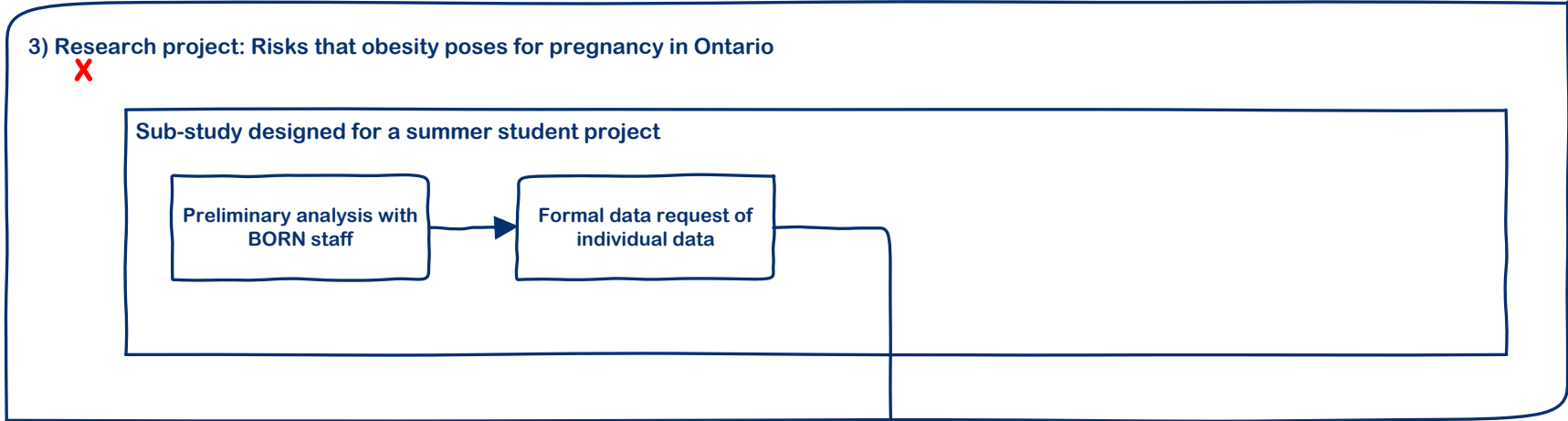


Process of using BORN Ontario Data

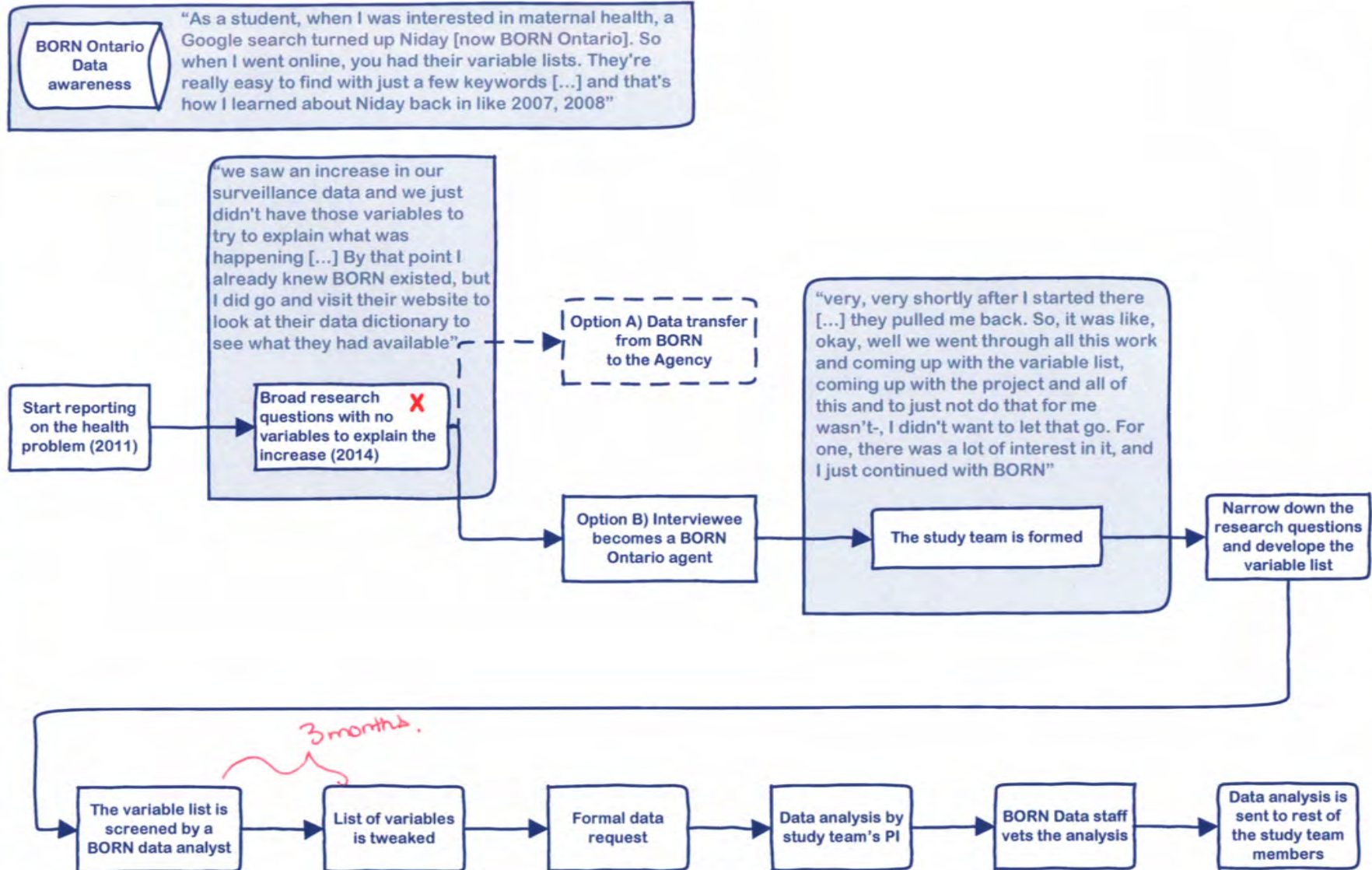
"I guess I first became aware of BORN as in my clinical role. All data on Ontario pregnancy is input into the database, so I knew that it existed. I first used BORN data when I was doing my master's thesis. That was my first contact with them"

1) BORN Ontario Data awareness

2) BORN Ontario Data usage



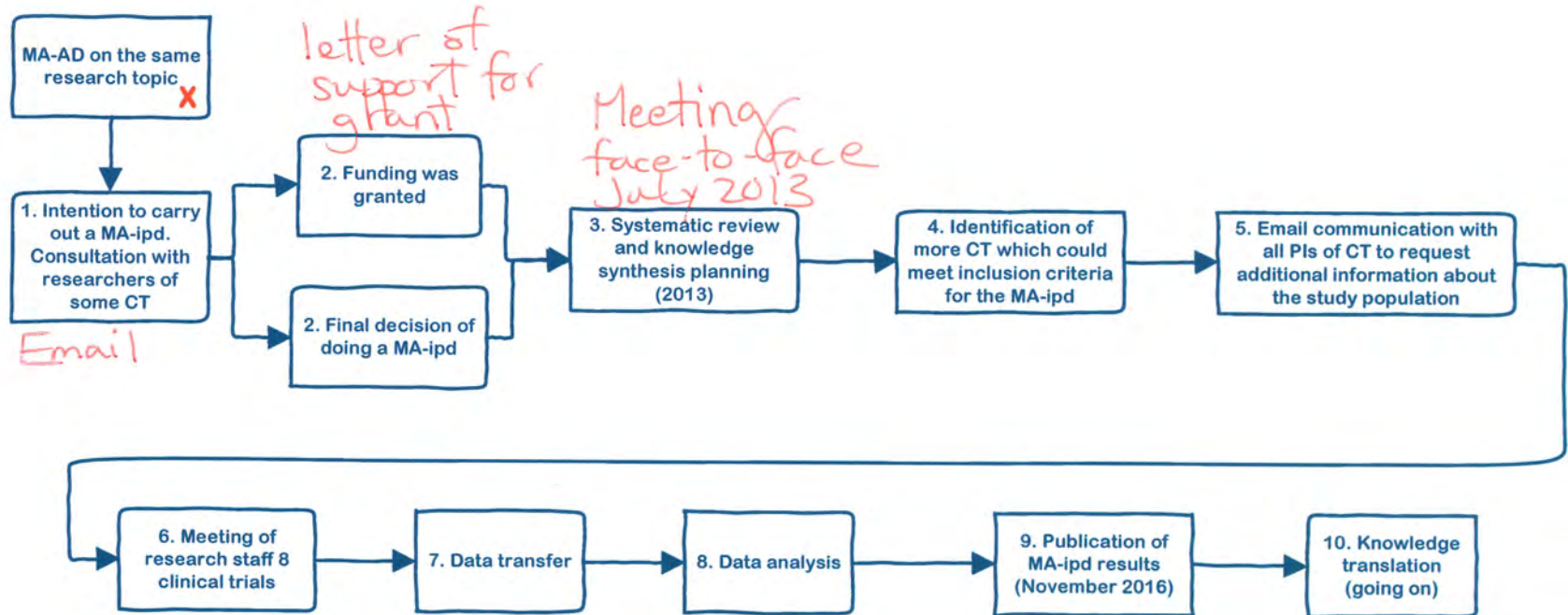
Process of using BORN Ontario Data



"There's a small community who've taken on these studies. [...] So the community itself is fairly small and the researchers know each other [...] Even before doing the systematic review there was an idea of who was going to be participating, which data we would use"



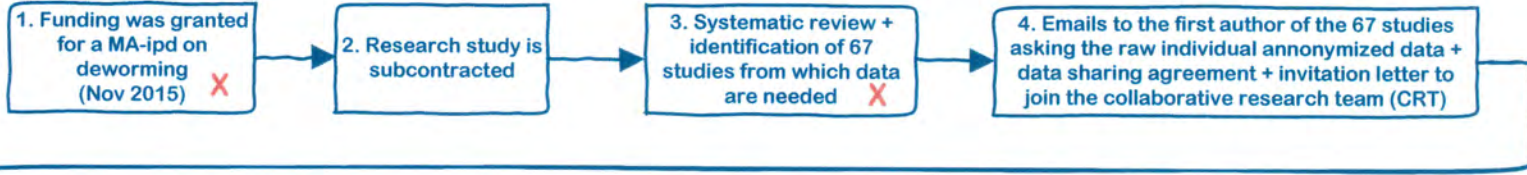
Process of using IPD from 8 clinical trials



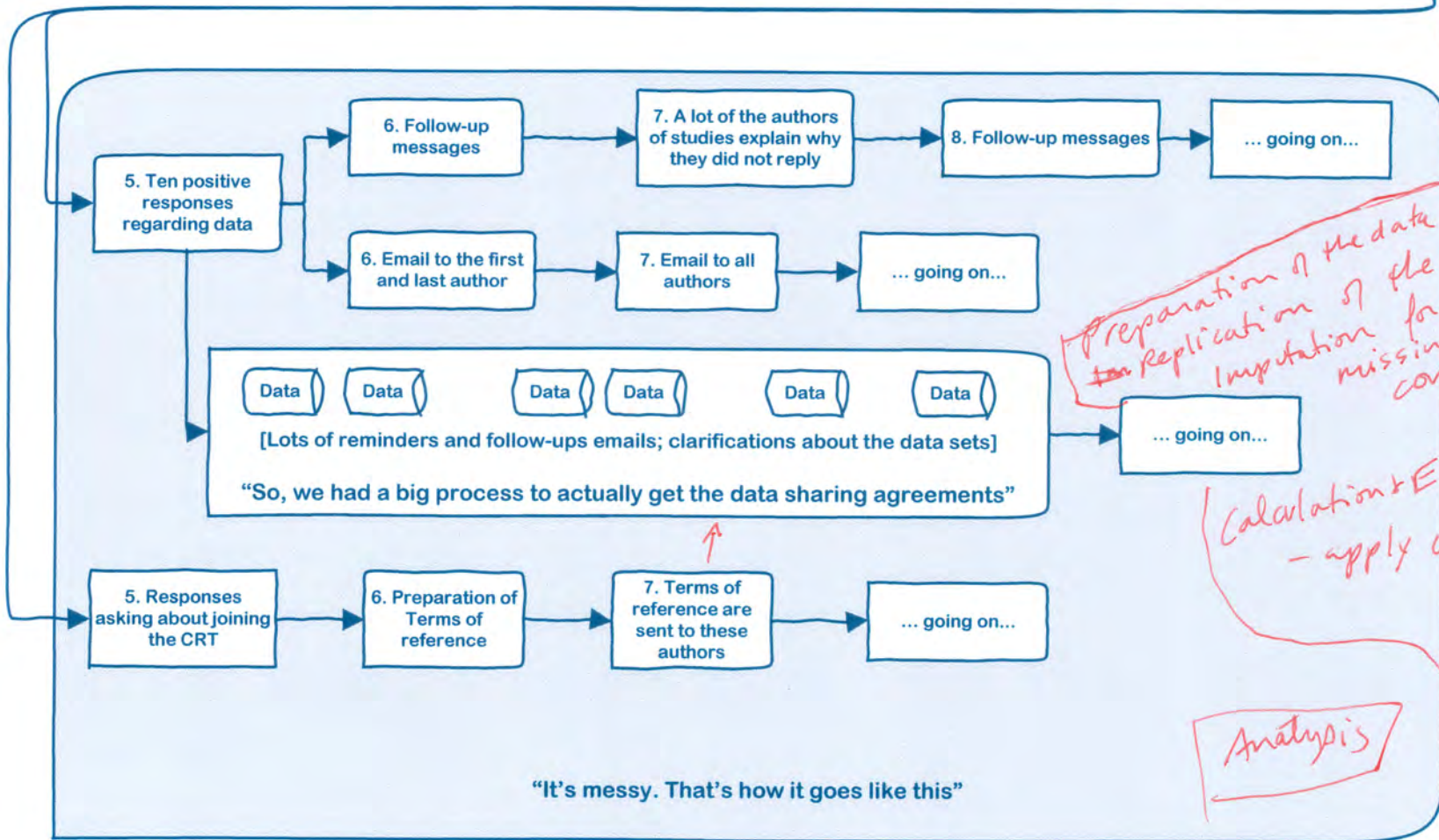
Mass deworming to improve developmental health and wellbeing of children in low-income and middle-income countries: a systematic review and network meta-analysis

Contact of authors + responses from 11/20,

Proposal writing →



Process of using IPD from ...? studies



Authors

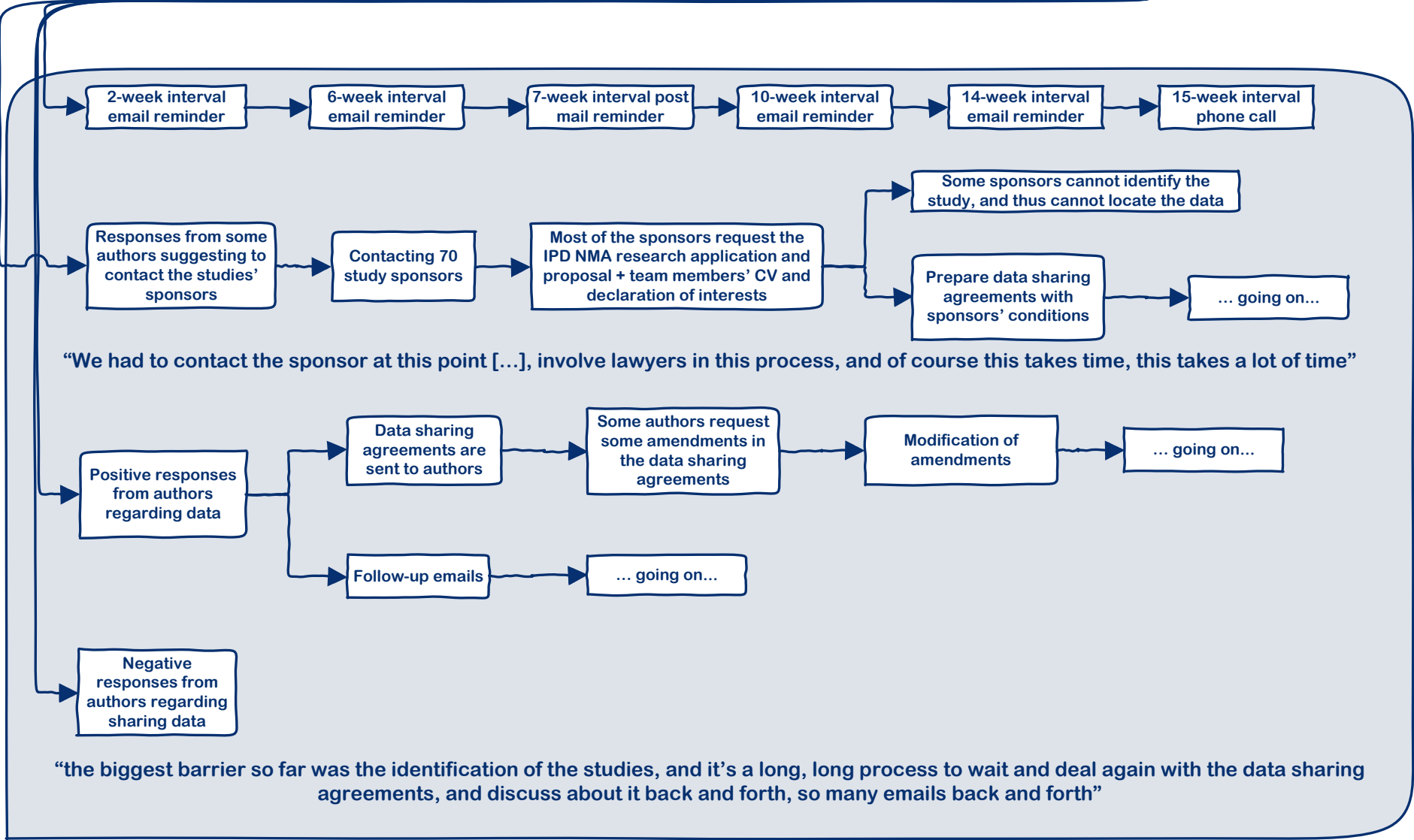
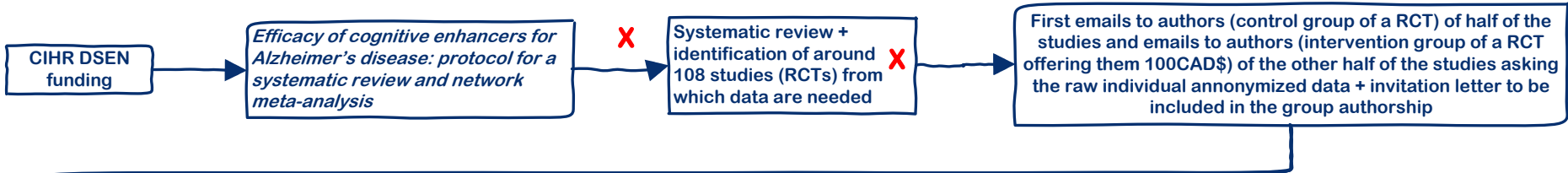
Preparation of the data
 Replication of the manuscript
 Imputation of missing baseline covariates

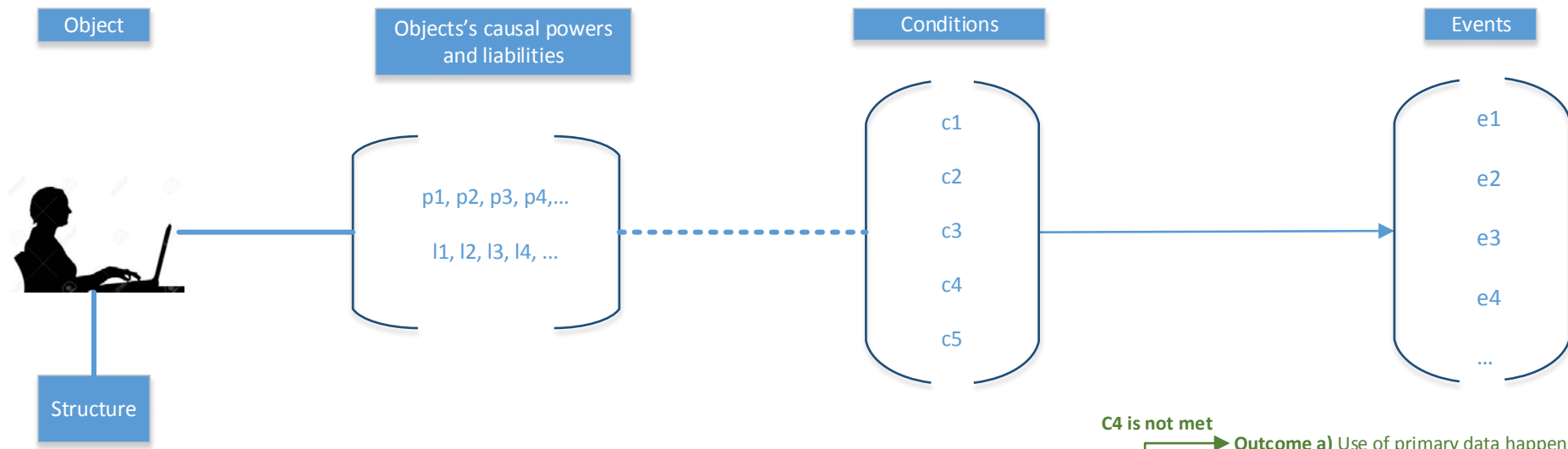
Calculation & Estimates
 - apply cutoffs
 - low
 - med
 - high

Analysis

"It's messy. That's how it goes like this"

Process of using IPD from ...? studies





Researcher X, who has a research activity and a structure S, which at least has these necessary internal relations:

Belongs to a (research) institution

Belongs to a discipline

Belongs to a reward system of science

Necessarily possesses the causal powers and liabilities to:

- Make (*satisficing*) decisions
- Take action
- Learn
- Identify knowledge gaps
- Make scientific contributions
- Act according to personal values
- Set up and pursue goals (whatever the motivation or reason is) and obtain resources (e.g., data, funding, etc.) to achieve them
- Interpret primary and secondary data
- Know epistemic practices of her discipline
- Know the limitations of the use of secondary data
- Know the limitations of collecting primary data
- Be influenced by norms (e.g., epistemic norms, institutional norms, etc.)
- React to unexpected circumstances

Under the following necessary conditions:

- (C1) The researcher knows that secondary data exist
- (C2) Data are obtained
- (C3) Particular secondary data are an initial *satisficing* option
- (C4) The idea of collecting particular primary data is not an initial *satisficing* option
- (C5) An expected scientific contribution exists and the researcher finds its potential rewards *satisficing*

C4 is not met

C4 is met

Outcome a) Use of primary data happen and primary data are used as evidence of scientific claims. Use of secondary data does not happen.

Outcome b) Use of primary data happen and primary data are used as evidence of scientific claims. Use of secondary data happens, but secondary data are not used as evidence of scientific claims. Instead, secondary data are used for the creation of background knowledge, thus, they do not appear in the scientific publication or contribution.

Outcome c) Use of primary data and secondary data happen, and primary data are presented as evidence of scientific claims. Secondary data can be presented in two ways: to support the scientific claim done with primary data or as evidence of scientific claims in combination with primary data.

Outcome 1) Use of secondary data does not happen at all after having tried or considered the option.

Outcome 2) Use of secondary data happens BUT reuse is not shared with the research community and the data do not end up being evidence of scientific claims. Thus, secondary data end up serving as widening the researcher's background knowledge and triggering new research hypotheses.

Outcome 3) Use of secondary data happens AND ONLY secondary data are used as evidence of scientific claims.

The literature included in Table 1 is not exhaustive. Its main purpose is to show the varied aspects that have studied under the topic of “data (re)use”. Inclusion or exclusion of literature in Table 1 - *Summary of some literature on data reuse* is based on a qualitative analysis of the content of each of the studies. The table includes some of the relevant empirical studies in English conducted mainly by IS scholars about researchers’ data reuse practices and factors affecting them when reusing data. I have also included empirical studies by authors in other disciplines, who have studied these or very similar issues. I have excluded publications, which contain the term “reuse” or “use of secondary data” but are unrelated to the aforementioned issues, e.g., Federer et al., 2015, Gregory et al., 2018, Kriesberg, Frank, Faniel, & Yakel, 2013, Nahar & He, 2016, van de Sandt et al., 2019, etc. I have also excluded publications that include both the terms “data sharing” and “data reuse” in their titles, goals or key words, but whose findings and conclusions focus mainly on “data sharing” (e.g., Tenopir et al., 2015). I have also disregarded studies about information reuse or knowledge reuse as per the reasons provided in Chapter 2. Literature review.

Some of the studies presented in the table are part of larger studies that are also include in the table since they fulfill with the above sampling criteria. I present the empirical studies in a descendent chronological order. Most of the text in the table, if not all, is authors’ *verbatim* text, despite not using italics. Rephrasing own authors’ words could distort their original meaning.

Table 1 - Summary of some literature on data reuse

REFERENCE	THEORIES, THEORETICAL, CONCEPTUAL OR ANALYTICAL FRAMEWORKS OR APPROACHES	RESEARCH QUESTIONS and/or RESEARCH GOALS	METHODS AND EMPIRICAL FIELDS	MAIN FINDINGS AND CONCLUSIONS
(Gregory, Groth, Scharnhorst, & Wyatt, 2019)	None	Data-seeking practices. Who are the people seeking data? What data are needed for research and how are those data used? How do people discover data needed for research? How do people evaluate and make sense of data needed for research?	Survey from responses from 105 countries. Disciplines: STEM Social sciences Health sciences Arts and humanities	Data needs are diverse and difficult to pigeonhole. Respondents selected using data as the basis for a new study most often, followed by teaching and preparing for a new project or proposal. Data uses are common, but their enactments are complex. People discover data via academic literature, via social connections, via “mediated” search, via specific searches plus casting a wide net, by building new practices. People make sense of data using varied evaluation criteria and sensemaking strategies, by using social connections in sensemaking, by using different contextual information for different purposes, by establishing trust and data quality.

(Faniel et al., 2019)	Chin and Lansing's four types of context	<p>Examine the types of context information that are needed to preserve data's meaning in ways that support data reuse.</p> <p>What types of context information do researchers need when deciding whether to reuse data?</p> <p>How do researchers' need for different types of context information vary across disciplinary communities?</p> <p>Our</p>	105 interviews and observations with researchers from three disciplinary communities: quantitative social science, archaeology, and zoology	<p>First, we found that social science, zoological, and archeological communities relied on four key context types Chin and Lansing (2004) identified: 1) general data set properties, 2) experimental properties, 3) data provenance, and 4) analysis and interpretation.</p> <p>Second, our findings showed it useful to distinguish and retain context information about the entity responsible for creating the data (i.e. data producer) from the entity that legally owned the data (i.e. data owner). In some cases, data producers were evaluated, but in other cases, their institutions had reputations and history, which served as a proxy for the individuals creating data (e.g. The World Bank, U.S. Government, etc.).</p> <p>Third, findings identified three new types of context: 1) information about specimens, 2) artifacts, and 3) missing data, which were specific to a particular discipline.</p>
(Pasquetto, 2018)	Knowledge infrastructures	What motivates the design of policies and infrastructures for open research data? How do researchers reuse open research data for knowledge production? What are the societal implications of making available and reusing open biomedical data across contexts of production?	<p>Ethnographic fieldwork: observations, interviews, document analysis (a case study)</p> <p>Biomedicine</p>	Impossible to predict how open research data will be reused, by whom and to what purposes. Meta-information might not be sufficient to enable reuse when data are accessed at a low level of processing, to run novel statistical analyses. Scientists (of the case study) seem to be mostly in favor of reusing data in collaborative settings.
(Curty, Crowston, Specht, Grant, & Dalton, 2017)	Theory of reasoned action	<p>Hypotheses:</p> <p>H1: Perceptions that data reuse has benefits will positively correlate with data reuse (H1a: Perceived efficiency of data reuse will positively correlate with data reuse. H1b: Perceived efficacy of data reuse will positively correlate with data reuse.)</p> <p>H2: Concerns about the trustworthiness of data will negatively correlate with data reuse.</p> <p>H3: Perceived norms against data reuse will negatively correlate with data reuse.</p> <p>H4: Perceived importance of data reuse will positively correlate with data reuse.</p>	Authors reused data from a worldwide survey of scientists developed and administered by the DataONE Usability and Assessment Working Group.	Results show that the perceived efficacy and efficiency of data reuse are strong predictors of reuse behavior, and that the perceived importance of data reuse corresponds to greater reuse. Expressed lack of trust in existing data and perceived norms against data reuse were not found to be major impediments for reuse contrary to our expectations.
(Kim & Yoon, 2017)	<p>Institutional theory and theory of planned behavior.</p> <p>A research model was developed to explain how disciplinary (or institutional) and individual factors influenced the data reuse behaviors of scientists</p>	<p>Hypotheses:</p> <p>H1: An open, collaborative research climate in a scientific discipline positively influences a scientist's intention to reuse other scientists' data.</p> <p>H2: The availability of data repositories in a scientific discipline positively influences the intention of a scientist within that discipline to reuse other scientists' data.</p>	<p>Survey</p> <p>STEM (Science, Technology, Engineering, and Mathematics) disciplines at academic institutions in the U.S.</p>	There are significant between-discipline variances as well as within-discipline variances in the impacts of both individual and disciplinary factors on data reuse intentions. At the individual level, perceived usefulness, perceived concern, and organizational resource were found to have significant relationships with data reuse intention. At the disciplinary level, availability of a data repository was found to have a significant positive relationship with data reuse intention. Perceived usefulness was found to be the most important factor influencing data reuse intentions,

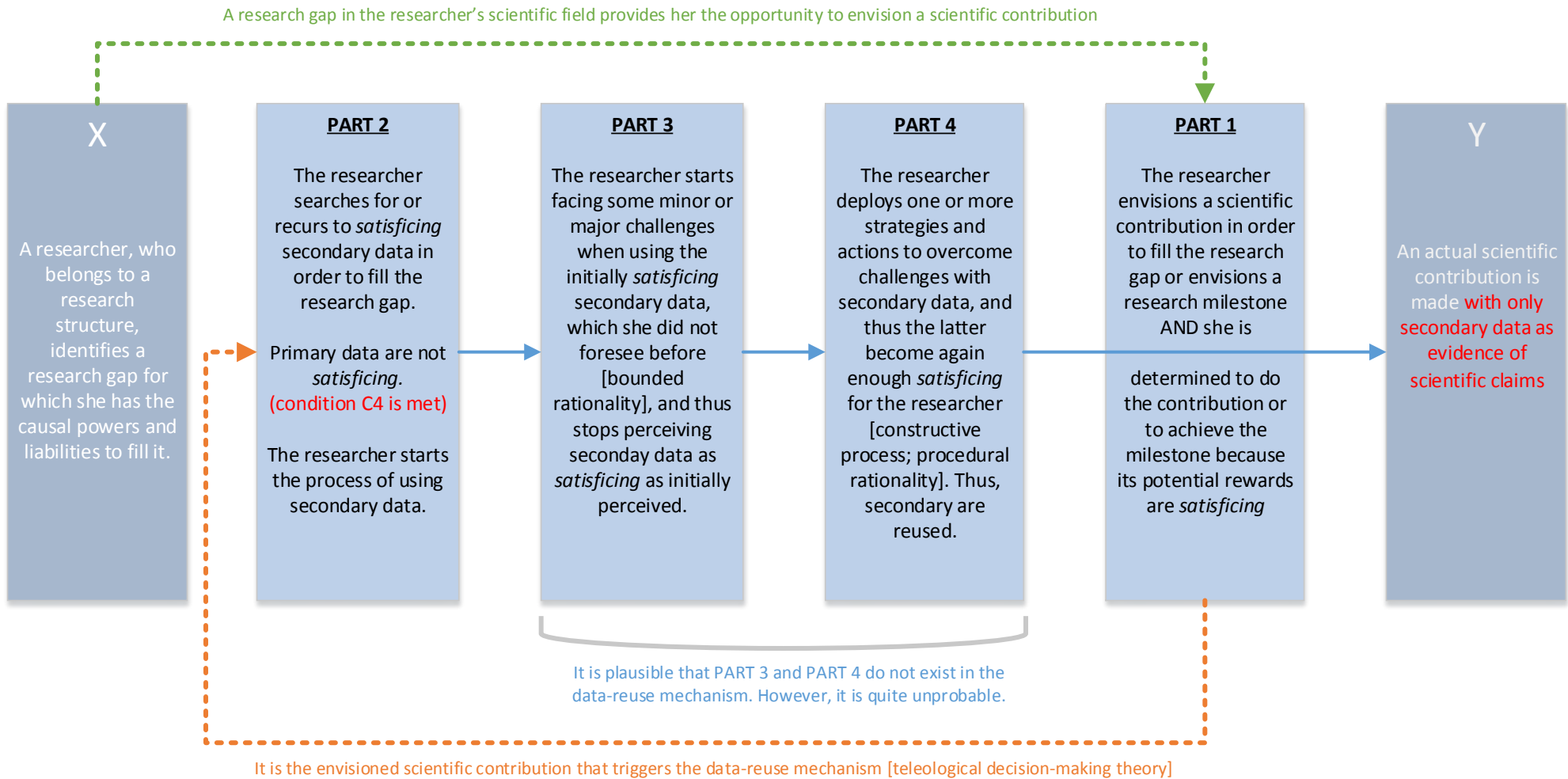
		<p>H3: The perceived usefulness of data reuse positively influences a scientist's intention to reuse other scientists' data.</p> <p>H4: The perceived concern involved in data reuse negatively influences a scientist's intention to reuse other scientists' data.</p> <p>H5: The perceived effort involved in data reuse negatively influences a scientist's data reuse intention.</p> <p>H6: The availability of internal resources supporting data reuse positively influences a scientist's intention to reuse other scientists' data.</p>		<p>whereas the perceived effort of data reuse is not associated with scientists' intentions to reuse data.</p>
(Murillo, 2016)	Unable to find out	<p>RQ1: What types of descriptive information are being made discoverable through DataONE? (How robust is the descriptive information made available regarding that data?; How is information being provided about the data, such as information regarding metadata standards, provenance information, research methods, instrumentation?; How is the provision of this descriptive information impacting the data-sharing infrastructure?)</p> <p>RQ2: What types of descriptive information could inhibit or facilitate data reuse? (How is information about the data such as information regarding metadata standards, provenance information, research methods, and instrumentation, influencing scientists' ability to determine if that data is reusable?; How does this information assist scientists in their ability to reuse this data?)</p>	<p>A data profiling assessment in the form of a quantitative and qualitative content analysis and a quasi-experiment think-aloud in DataONE.</p> <p>Earth and environmental disciplines</p>	<p>The quasi-experiment think-aloud indicates that scientists found pieces of descriptive information particularly useful for their ability to determine data reusability. These include: (a) the data description, (b) the attribute table, and (c) the research methods.</p> <p>Metadata schema, member node standards, and community standards, impact what types of descriptive information are provided through the shared data.</p> <p>Attribute and unit lists, research methods information, and succinctly written abstracts facilitate data reuse. However long abstracts and having the same information in multiple places, and the exclusion of data descriptions inhibit data reuse.</p>
(Faniel, Kriesberg, & Yakel, 2016)	Consumer's fulfillment response	<p>What data quality attributes influence data reusers' satisfaction after controlling for journal rank?</p> <p>Hypotheses:</p> <p>H1: Data relevancy is positively related to data reusers' satisfaction.</p> <p>H2: Data completeness is positively related to data reusers' satisfaction.</p> <p>H3: Data accessibility is positively related to data reusers' satisfaction.</p> <p>H4: Data ease of operation is positively related to data reusers' satisfaction.</p> <p>H5: Data credibility is positively related to data reusers' satisfaction.</p>	<p>Citation analysis and a survey</p> <p>Social sciences</p>	<p>We found that data completeness (H2), data accessibility (H3), data ease of operation (H4), and data credibility (H5) were significant, as predicted. Support for these hypotheses suggests that data reusers' satisfaction corresponded with reusing data that were comprehensive, easy to obtain, easy to manipulate, and believable. We also found that documentation quality (H7) was significant. Higher levels of documentation quality corresponded with higher levels of data reusers' satisfaction. Data accessibility had the strongest relationship with data reusers' satisfaction. Social scientists were more satisfied when the data they reused were easily obtainable.</p> <p>Not all of the hypotheses were supported. Surprisingly,</p>

		H6: Data producer reputation is positively related to data reusers' satisfaction. H7: Data documentation quality is positively related to data reusers' satisfaction. H8: Journal rank is positively related to data reusers' satisfaction.		data relevancy (H1) and data producer reputation (H6) were not significant in the multiple regression model.
Renata Gonçalves's dissertation (Curty, 2015)	Unified Theory of Acceptance and Use of Technology (UTAUT)	What are the factors that influence scientists' research data reuse? To what degree do these factors influence scientists' research data reuse? To what extent do scientists reuse research data?	Interviews and a survey Social sciences	25 factors that were found to influence their perceptions and experiences, including both their unsuccessful and successful attempts to reuse data.
(Curty & Qin, 2014)	Unified Theory of Acceptance and Use of Technology (UTAUT)	How do scientists assess the reusability of research data? What are the factors that influence scientists' research data reuse?	Interviews and a survey Social sciences	8 factors were identified to affect researchers' when they reuse data.
(Daniels, 2014)	Social worlds; community of practice; trading zones; infrastructure; boundary objects.	<i>What is the relationship between museum objects, their representations, and research use?</i> This question deals with the practices of researchers using museum collections, seeking to understand how they use museum data to make new contributions to their own field. <i>What factors influence the practices of staff members as they describe and manage museum data?</i> Through this question, I address the transformation of museum objects into data that is performed by museum staff through the description of collections, asking what norms structure that activity.	Comparative case study: Semi-structured interviews, non-participant observation, archival research Researchers in botany and archaeology	Researchers use complex accumulations of museum objects and their representations, including metadata, to address different types of research goals, applying the evidential norms of their research communities to their approach to data. The need for access to objects, metadata, and documentation differs among both botanists and archaeologists, corresponding to the kind of analysis (type- or provenance-based) researchers intended to use. Data collectors reputation is a major factor in the decision to reuse data.
(Yoon, 2014)	None	Explore qualitative researchers' experiences reusing data in the field of social science in US, which have not been empirically addressed yet. The	In-depth interviews Social sciences	Qualitative data reusers have a strong preference and needs for personal interaction with original investigators.
(Fear, 2013)	No theoretical framework – Prediction of reuse using data citation	What is the scholarly impact of data reuse? How can stakeholders anticipate the impact the data they fund, create or curate will have? How and why do social scientists cite data? Which datasets held by ICPSR are high-impact according to different measures of reuse impact? What characteristics of data predict whether they will be reused?	Document analysis for bibliometric behavior, and literature review on factors affecting data reuse Social sciences	Regarding size of data set: broad datasets, those with many variables, are more likely to be reused than deep datasets, or those with a large number of cases. Regarding the discipline of the data producer: Whether a dataset's producers represented a single or multiple disciplines was never a significant predictor of reuse or downloaders, and it was not significantly related to any of the citation-based impact metrics. Regarding data collection process information: Processed studies were three to four times more likely to be reused than unprocessed studies and had more than five times as many downloaders. But whether the importance of the

				<p>processing status predictor was related to documentation quality, as posited earlier in this study, is unclear.</p> <p>Data producer reputation: the reputation of the data producer was not a substantial predictor of reuse. Regarding Connection with the data producer (co-authorship network size): co-authorship network size was not a significant predictor of reuse or any of the reuse impact measures, nor did it ever have a meaningful effect on any of the outcome variables.</p> <p>Regarding Prominence of data (presence in research literature): the amount of literature about a dataset seems to be a key indicator of reuse. An increase of one primary publication prior to reuse increased the odds of reuse by between 13% and 20%, with a tipping point at three publications, at which point datasets were more than three times as likely to be reused.</p>
(I. Faniel, Kansa, Kansa, Barrera-Gomez, & Yakel, 2013)	None	<p>1) How does contextual information serve to preserve the meaning of and trust in archaeological field research over time?</p> <p>2) How can existing cultural heritage standards be extended to incorporate these contextual elements? More</p>	<p>Semi-structure interviews</p> <p>Archaeology</p>	Context surrounding the research methods, people, and repository processes were particularly important.
(Yakel, Faniel, Kriesberg, & Yoon, 2013)	Dimensions of trust from management and information systems literatures	<p>1. How do data reusers construct/conceive of trust in repositories? 2. How do data reusers associate repository actions with trustworthiness?</p> <p>Our</p>	<p>Semi-structured interviews</p> <p>Social scientists</p> <p>Archaeologists</p> <p>Our</p>	There are similarities and differences across the two disciplines. Both disciplinary communities associated trust with a repository's transparency. However, archaeologists mentioned guarantees of preservation and sustainability more frequently than social scientists who talked about the influence of colleagues and institutional reputation.
(I. M. Faniel, Kriesberg, & Yakel, 2012)	Communities of practice literature	How do novice social science researchers make sense of social science data?	Semi-structure interviews with novice social science researchers reusing data from ICPSR repository	Novice social science researchers were particularly interested in making sense of how data 1) were transformed from qualitative to quantitative, 2) captured concepts not well-established in literature, and 3) could be matched and merged across multiple datasets
(I. M. Faniel & Jacobsen, 2010)	None	Examine how earthquake engineering (EE) researchers assess the reusability of colleagues' experimental data for model validation	<p>Interviews</p> <p>EE researchers</p>	EE researchers' strategy when assessing the relevance of colleagues' data is to match key parameters from their models to key parameters from colleagues' experiments. Using such a strategy, EE researchers prefer high, rather than low-level context information about how the experiment was set up and how the specimens were tested. The fine-grained understanding of the context of data production is based on how competently primary data producers document the artifacts and processes used and created during their experiments.
(Niu, 2009a)	None	How do researchers overcome inadequate documentation to understand (secondary) data for reuse?	Unstructured and exploratory interviews	Researchers use documents (previously written articles using the data, websites of data producers, websites of data

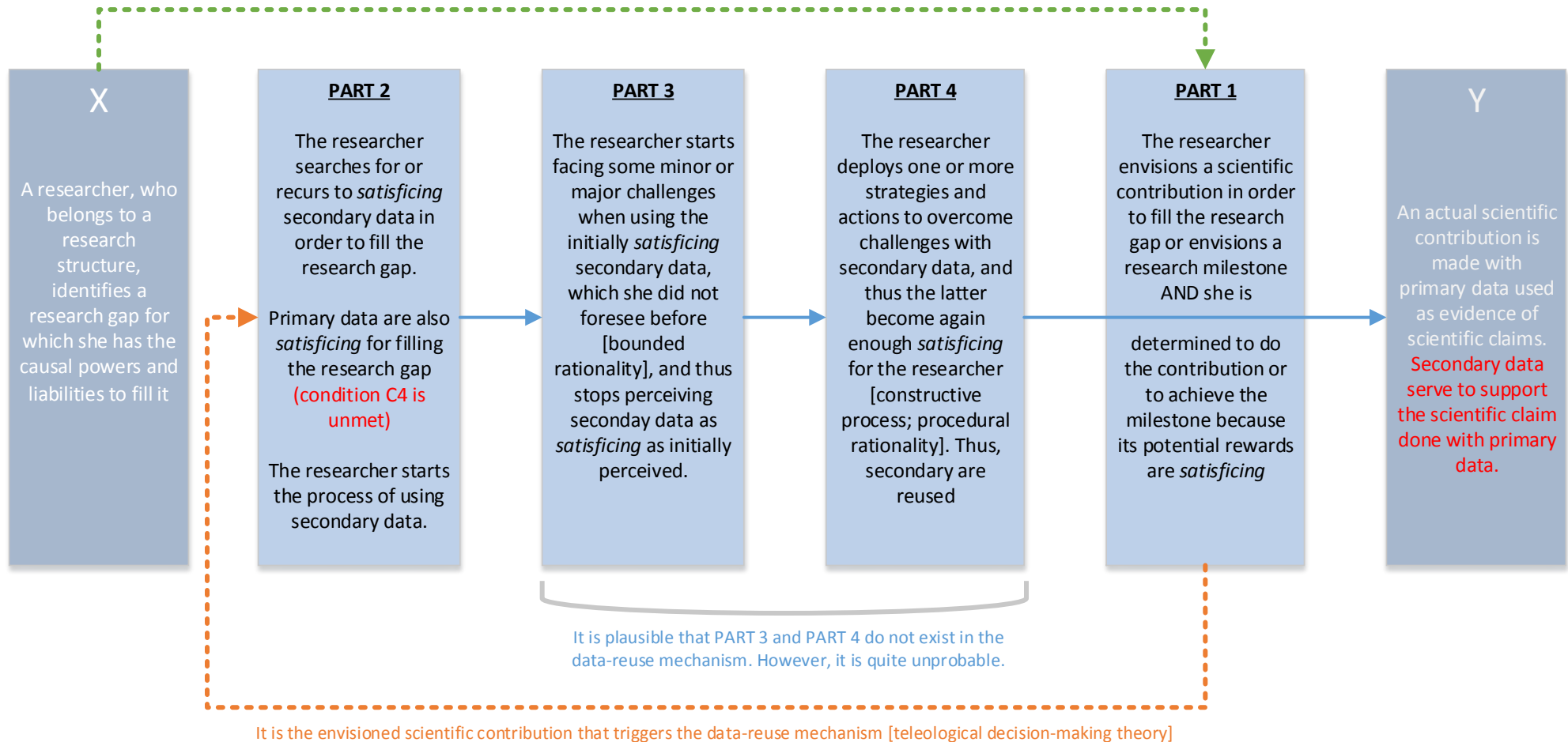
			Social sciences	archives) and people (data producers, other secondary users, data archivists, and in workshops)
(Niu, 2009b)	Knowledge transfer and absorptive capacity	What are the impacting factors of perceived documentation quality? How does perceived documentation quality affect users' incentives to use secondary data? How do secondary data users overcome inadequate documentation?	Interviews and one survey Social sciences	Inadequate documentation increases use cost and may turn users away in some situations. However, users' incentives to use secondary data mostly depend on how well the data fit their information needs rather than documentation quality. Users with stronger absorptive capacity tend to perceive the documentation they use as slightly better than users with weaker absorptive capacity.
(Heaton, 2008)	None	It clarifies what secondary analysis is and how the methodology relates to other similar approaches used in qualitative research. It looks at the development of secondary analysis in qualitative research, and some of the factors that have shaped this. It examines the ways in which researchers have re-used qualitative data.	Publications based on reused social science data	86% of the reusers were involved in the primary research. 14% were not. Key findings about factors that reuse: data fitness, verification of original studies, understanding/interpreting the data, legal and ethical issues.
(Zimmerman, 2008)	Distance spanners, circulating reference	What are the experiences of ecologists who use shared data? Goal of the study: understand the role that the presence or absence of standards has in knowledge transfer	Interviews Ecologists	The knowledge that ecologists acquire through fieldwork enables them to recover the local details that are so critical to their comprehension of data collected by others. Social processes also play a role in ecologists' efforts to judge the quality of data they reuse.
(Zimmerman, 2007)	None	Investigates the processes by which ecologists locate data that were initially collected by others.	Interviews Ecologists	Ecologists use formal and informal knowledge that they have gained through disciplinary training and through their own data-gathering experiences to help them overcome hurdles related to finding, acquiring, and validating data collected by others. Ecologists rely on formal notions of scientific practice that emphasize objectivity to justify the methods they use to collect data for reuse.
(Zimmerman, 2003)	Communities of practice; theory and concepts of measurement as a social technology; circulating reference; inscriptions; boundary objects; standards.	What are the experiences of ecologists who use shared data? How do ecologists locate data? What are the characteristics of the data received? What information about the data do ecologists receive and/or depend on to use the data? How do ecologists assess the quality of the data they receive? What challenges do secondary data users face, and how do they overcome them?	Semi-structured in-depth Interviews (case studies) Ecologists	Results show that while personal interaction and cultural factors play a role in nearly all experiences, neither changes the overall approach that ecologists take throughout the process. Ecologists choose methods to gather data for reuse and to make decisions about data acceptance that meet community and individual standards and that can be defended publicly. Ecologists' decisions regarding what data to reuse are influenced by a combination of domain knowledge, personal tolerance for uncertainty, and individual knowledge.
(Hyman, 1972)	None	Discover and describe the distinctive styles of work and thought of scholars who are "successful secondary analysts".	Survey and "interviews" (in writing)	To teach secondary analysis and increase the number of successful practitioners.

The data-reuse mechanism, which explains causally the use of secondary data as evidence of scientific claims



The data-reuse mechanism, which explains causally the use of secondary data as evidence of scientific claims

A research gap in the researcher's scientific field provides her the opportunity to envision a scientific contribution



The data-reuse mechanism, which explains causally the use of secondary data

A research gap in the researcher's scientific field provides her the opportunity to envision a scientific contribution

