Routledge
Taylor & Francis Group

Economics of Innovation and New Technology

ISSN: 1043-8599 (Print) 1476-8364 (Online) Journal homepage: www.tandfonline.com/journals/gein20

Opening the black box: an explanatory model of data reuse in health research and beyond

Inma Aleixos-Borrás , Pablo D'Este & José J. López

To cite this article: Inma Aleixos-Borrás , Pablo D'Este & José J. López (30 Apr 2026): Opening the black box: an explanatory model of data reuse in health research and beyond, Economics of Innovation and New Technology, DOI: 10.1080/10438599.2026.2660308

To link to this article: https://doi.org/10.1080/10438599.2026.2660308

ECONOMICS OF INNOVATION AND NEW TECHNOLOGY
https://doi.org/10.1080/10438599.2026.2660308

Opening the black box: an explanatory model of data reuse in health research and beyond

Inma Aleixos-Borrás[a,b], Pablo D'Este [c] and José J. López[d]

[a]Institute of Entrepreneurship and Innovation Science (ENI), Universität Stuttgart, Stuttgart, Germany; [b]Department of Management, Universitat de València, Valencia, Spain; [c]INGENIO (CSIC-Universitat Politècnica de València), Valencia, Spain; [d]School of Sociological and Anthropological Studies, University of Ottawa, Ottawa, Canada

ABSTRACT

Previous studies on the use of secondary data – or data reuse – found that researchers confront considerable (un)foreseen challenges related to finding, accessing, decoding, and recontextualizing secondary data. However, why some researchers persist in this endeavor while others do not is unclear. We develop a plausible theoretical explanatory model – the data-reuse mechanism of this phenomenon. This model captures the relational aspects of the different elements involved in the process, i.e. scientific horizon and reward system, researcher's cognitive frameworks and relational structures, the properties and relational structure of data, specific conditions, and time. We argue that it is the interaction among these elements that ultimately allows the researcher to decide about data reuse. Guided by this model, we analyze the data reuse process in ten molecular/computational biology and epidemiology case studies. The results of our analysis show how the intertwined synchronic and diachronic relations among these elements add considerable uncertainty to the reuse process. It is precisely this uncertainty combined with researchers' tenacity, scientific horizon, and flexibility in changing this horizon that enable data reuse. Our findings demonstrate also that while some data properties such as open access may facilitate data reuse, they do not fully determine whether data are reused.

ARTICLE HISTORY
Received 2 October 2024
Accepted 24 March 2026

KEYWORDS
Data reuse; social mechanism; use of secondary data; bounded individual horizon; health sciences

> Knowledge that has been codified […] can be reconstituted at a later time, in a different place, or by a different group of individuals with varying degrees of effectiveness depending upon the 'cognitive framework' of those attempting to use this information. (Cohendet and Steinmueller 2000, 197)

CONTACT Inma Aleixos-Borrás maria.aleixos-borras@eni.uni-stuttgart.de Institute of Entrepreneurship and Innovation Science (ENI), Pfaffenwaldring 19, c/o ARENA2036 70569, Stuttgart-Vaihingen, Germany

Supplemental data for this article can be accessed online at https://doi.org/10.1080/10438599.2026.2660308.
This article has been corrected with minor changes in authors' affiliations. These changes do not impact the academic content of the article.

1. Introduction

More than two decades ago, Ed Steinmueller, together with other colleagues, anticipated some of the effects of knowledge codification and the influence in science of information and communication technologies (ICTs) (Cohendet and Steinmueller 2000; David and Steinmueller 2003). According to these authors, ICTs promised to 'boost research productivity, particularly in the processing, storage and exchange of information' (David and Steinmueller 2003, 2), which, in turn, would facilitate path-breaking research advancements. These authors were visionaries since 'the explosion of computer processing power and storage capacity has transformed the sciences' ability to find, use, coordinate, and reuse […] data' (Edwards et al. 2011, 668). Indeed, recent science policy requirements related to data sharing and open data under the auspices of Open Science (Huang and Soete 2025) have contributed to the increasing codification of knowledge and growing requirements for investment in the design, development, and maintenance of modern complex socio-technical 'collaboratories' (Finholt 2003) such as data infrastructures, data repositories, and codification and inscripting efforts of those who share data (Borgman 2012).

These investments and efforts are justified only if data are actually reused and if the results of this reuse are beneficial to science in particular and to society in general (Borgman 2015; Cohendet and Steinmueller 2000; Pasquetto, Randles, and Borgman 2017). In this context, Steinmueller and co-authors have long called for research into the social processes and mechanisms that contribute to 'reproducing "tacit" capabilities to make codification or decoding [of knowledge] feasible' (Cohendet and Steinmueller 2000, 206). They have called also for 'empirical work to discover the heuristics that individuals and groups use to manage the growing complexity of the information flux associated with knowledge codification' (Steinmueller 2000, 373–374) which often requires increased division of labor (Caldas 2003; Foray and Steinmueller 2003).

In this inductive qualitative study of the socio-technical processes involved in reuse of research data in two health disciplines, we take up the baton and focus on the 'level of the individual' (Steinmueller 2000, 374) by examining the problems related to decoding and using information (Cohendet and Steinmueller 2000). We conceive of data as 'knowledge reduced to messages that can be transmitted to decision agents' (Ancori, Bureth, and Cohendet 2000, 256) that involve 'a large tacit component of skill[s] and capabilities that is embodied in people, products, and procedures' (Wolfe and Lucas 2001, 174–175). We define *reuse of data* or *use of secondary data* as the application of data to answer a novel research question by a researcher working within the institutional norms of science who did not participate in the original collection or generation of the data. This inductive study is guided by findings from recent research that shows that use of encoded knowledge – in the form of research data – entails a range of challenges whose 'costs and complexities' could outweigh codification efforts (Cohendet and Steinmueller 2000, 206). Cohendet and Steinmueller (2000, 197) argue that

> [k]nowledge that has been codified […] can be reconstituted at a later time, in a different place, or by a different group of individuals with varying degrees of effectiveness depending upon the 'cognitive framework' of those attempting to use this information,

suggesting that the cognitive framework influences how reusers assign different meanings to data (Cowan, David, and Foray 2000).

However, the different reinterpretations depend not only on the researcher's cognitive framework (Ansbacher 1950; Borgman 2015; Leonelli 2016) but also on their training, skills, and tenacity (Hyman 1972; Irwin and Winterton 2011). Recent research shows that some of the difficulties (i.e. understanding, analyzing, or repurposing data) associated with decoding research data are cognitive (see e.g. Curty 2015; Curty et al. 2017; Faniel, Frank, and Yakel 2019; Gregory et al. 2020; Hemphill et al. 2022; Kim and Yoon 2017; Murillo 2016; Pienta et al. 2018; Yoon 2016b) and contingent on the situated and contextualized nature of data collection or creation (Borgman 2007; Foray and Steinmueller 2003; Leonelli 2016). The situatedness of the data creation and data collection processes may be better captured in synchronous collaboration between data producer and data reuser (Olson and Olson 2003). However, if such collaboration is impossible, context can be transmitted by means of metadata (Pasquetto 2018), codebooks (David and Steinmueller 1991; Garmon Bibb 2007; Niu 2009a), standards or dictionaries (Cowan, David, and Foray 2000; Zimmerman 2008), or by asynchronously contacting the original data collectors (Yoon 2017). Additionally, as Zimmerman (2007; 2008) showed in her study of data reuse by ecologists, researchers can use their field experience to compensate for lack of contextual detail related to data collected by others.

Researchers reusing data can also face some more practical and technical issues, including accessing, retrieving, storing, manipulating, and cleaning the data, which in turn may depend on legal and ethical considerations. Most previous studies found that these challenges can be anticipated at any stage in the data reuse process and may deter reuse of the data (Yoon 2016b). However, in many cases, the expected problems and the significant perceived effort required to overcome them do not prevent the reuse of data by researchers. For instance, Niu (2009b) showed that if the data meet their particular information needs, social scientists will make considerable efforts to overcome inadequate documentation. Similarly, Zuiderwijk and Spiers (2019) found that if 'the data is important for [their] analysis,' astrophysics researchers will expend great efforts to overcome data reuse problems (238). Kim and Yoon's (2017) statistical analysis of data reuse behaviour in several disciplines indicates a statistically significant relationship between researchers' intentions to reuse data and the perceived efforts connected to their reuse. They conclude that '[p]erhaps the needs of the scientist and the usefulness of the data are more important than any effort' (Kim and Yoon 2017, 2716) and suggest further exploration of this finding.

These inconclusive findings inform the present study's research question: What factors underpin continuation of the data reuse process to completion in the face of the possible challenges the researcher might anticipate and encounter?

To address this question, we theorize an explanatory model, the *data-reuse mechanism* where use of secondary data is a socially embedded process that incorporates a broad range of behavioural and social elements such as researcher's goal-oriented work and researcher's reward system, and where artifacts and practices strongly influence the social-technical process (Latour 1987, 2005; Latour and Woolgar 1986). The data-reuse mechanism serves as a tool for our analysis of ten case studies of the use of secondary data: four cases in molecular/computational biology and six cases in epidemiology. Our findings show that it is mainly the uncertainty originating in the relational-dependent (a)synchronous conditions combined with researchers' tenacity and flexibility in sacrificing their scientific horizon which sustain the data reuse process to its conclusion.

2. The theorized data-reuse mechanism

In this study, we draw on 'social mechanisms' as useful and pragmatic intellectual practices to cope with the demonstrated limitations of 'grand theory' for capturing the complexity of the social world (Rojas 2017). Social mechanisms can be understood as theoretical models for aspects of social life that cannot be explained by grand theories (Beach and Pedersen 2013; Gläser 2012; Mayntz 2004; Rojas 2017). A social mechanism can be defined empirically as '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 and Laudel 2013, 5).

Social mechanisms can be theorized inductively or on the basis of existing theory and plausible theoretical hypotheses (Beach and Pedersen 2013). We adopt the latter approach to theorizing the *data-reuse mechanism* in an attempt to provide an account of the conditions, additional to the researcher's cognitive capacity, under which researchers reuse research data, and why researchers reuse data under unfavorable conditions. The data-reuse mechanism rests on four presumptions. First, data reuse is embedded in a wider situated epistemic inquiry process in which the researcher works towards a future rewardable goal or a scientific contribution horizon. Aleixos-Borrás and López (2025, 4) consider that 'a scientific contribution emerges as a clear orientation towards the future, a horizon or intention in a web of social, material, and institutional relations', i.e. a researcher's *bounded individual horizon* (BIH). Researchers' contributions are conditioned by both their scientific and social circumstances that govern the resources available to them, and their personal and professional circumstances, values, beliefs, and relations with others. Second, although acknowledged by most previous studies that data reuse is not the goal *per se* (Borgman 2015; Pasquetto, Randles, and Borgman 2017) but rather the means enabling the production of knowledge, their empirical methods neglect the fact that the ultimate and intrinsic goal of the data reuse process might be the creation of knowledge. We take this fact into account and theorize the data-reuse mechanism as a process that mediates between the need or willingness to address a research gap and the solution to that gap that is 'worthy of scholarly recognition' (Borgman 2015, xviii). Third, data reuse depends heavily on both myriad legal, ethical, and technical – digital and analogical – aspects of the data as well as the reuser's available infrastructure (Mansell and Edward Steinmueller 2000). Fourth, data reusers cannot fully foresee the challenges that they will encounter when reusing the data (Faniel, Kriesberg, and Yakel 2012; Hyman 1972; Yoon 2016a; Zimmerman 2007) and thus, cannot calculate in advance the effort that will be required to overcome these challenges. According to Simon (1955, 1957; 2000) individual capacity to make decisions is bounded, and therefore rather than the best choice, the individual will select the one that is *satisficing* or good enough.

Although in what follows the data-reuse mechanism is presented as involving four sharp, linear, and orderly steps, in reality and empirically it is likely to be cyclical, overlapping, and messy (Marshall and Rossman 2011; Reichenbach 1938). Moreover, the researcher's agency in terms of fitting the data to the research gap(s) is continuous throughout the entire process (Aleixos-Borrás and López 2025).

Step 1. The researcher identifies a research gap and envisions a scientific contribution, to be transmitted usually in the form of '[a] journal article, book, conference paper, or

other products worthy of scholarly recognition' (Borgman 2015, xviii), which triggers the data reuse process. The data themselves might be the source of new research questions, setting in motion a process of data reuse.

Step 2. The researcher searches for and obtains *satisficing* data whose codification may or may not be standardized within the epistemic community that collected or created the data (Cowan 2001; Cowan, David, and Foray 2000). Satisficing data are those data that *a priori* are perceived by the researcher as fitting the envisioned scientific contribution (SC), and are reusable according to an initial quality check of their integrity, reliability, validity, trustworthiness, and/or provenance. In step 2, researchers may encounter unexpected technical, ethical, and legal issues related to accessing or analyzing the data which may make secondary data less satisficing than initially foreseen. That is, the subjective fitness between the data and the envisioned SC may change throughout the process.

Step 3. The researcher can adopt different strategies to overcome the problems that arise when accessing, analyzing, interpreting, and using the data to support their scientific claims. Aleixos-Borrás and López (2025, 12) maintain that the researcher's agency plays an essential role in the 'fitting process' that involves interaction with the properties and attributes of the data and the 'fus[ing] [of] concepts, constructs, theoretical elements, variables, etc'. These authors' claims concur with Hyman's (1972) argument that researchers must have not just a cognitive framework but a specific 'style of work' and 'distinctive qualities' (Hyman 1972, 78) in order to 'convert' imperfect data into valuable material (Ansbacher 1950; Cohendet and Steinmueller 2000; Hyman 1972) suited to repurposing the data.

Step 4. The data reuse process concludes, and the scientific contribution is achieved by employing secondary data.

In what follows, we propose an operational data-reuse mechanism model based on Sayer's explanatory structure of social phenomena (Sayer 2000; 2010).

Sayer (2000, 2010) proposed five constituent elements of a mechanism: an *object* or *subject* that belongs to a relational *structure* and has *causal powers and liabilities* or 'capacities to behave in particular ways, and […] specific susceptibilities to certain kinds of change' (Sayer 2000, 11), *conditions* which activate these capacities and susceptibilities, and *events* which are the mechanism's outcomes. Thus, in Sayer's terminology, the data-reuse mechanism (Figure 1) is one that provides an individual with the necessary causal powers and liabilities, the ability to make a decision about the use of secondary data – an object with its own causal powers and liabilities – in the context of all the individual's structural relationships (*relational structure*), and under certain specific conditions (*conditions*) in order to make a scientific contribution (event) (Sayer 2010, 70–79).

Based on Sayer's structure of explanations, we theorize five *sine qua non* conditions of the data-reuse mechanism. The researcher knows that secondary data exist (condition C1); the data have to be accessed or obtained by the researcher (condition C2); secondary data are a satisficing option for the researcher's objective of making a scientific contribution (condition C3); the idea of collecting particular primary data is not a satisficing option (condition C4). Although it is well known that researchers may combine both primary and secondary data in order to produce new knowledge, we have theorized this mechanism to explain the use of secondary data, allowing a simpler operationalization. Moreover, primary data do not require decodification and recontextualization processes by the researcher who collects or produces them. Condition C5 is that an

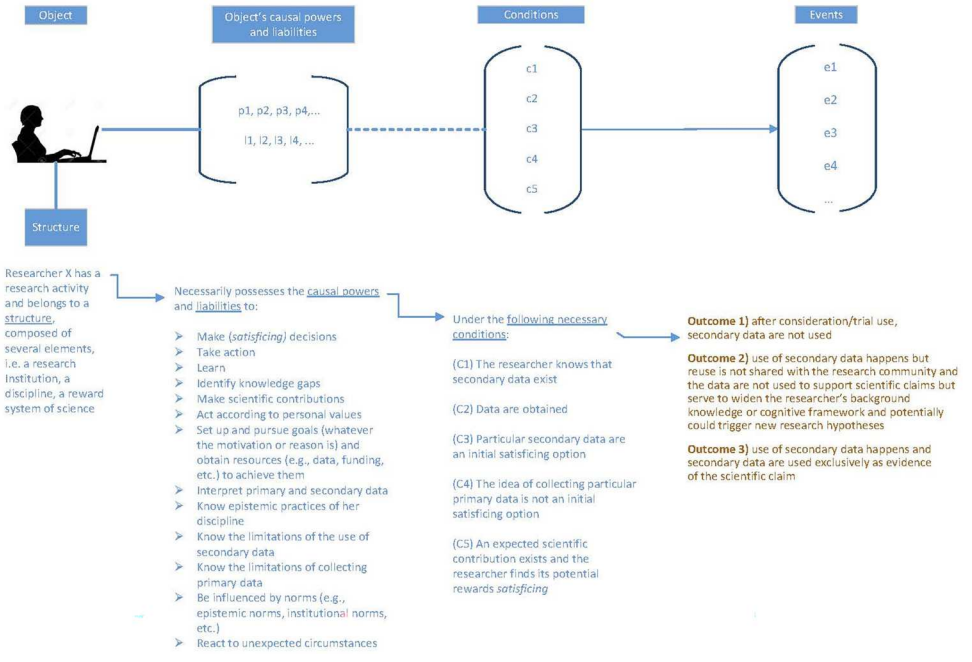

Figure 1. The theorized data-reuse mechanism based on Sayer's structure of (causal) explanations (2000; 2010, 74).

expected scientific contribution exists and the researcher finds its potential rewards satisficing for their scientific career. Researchers' theorized causal powers and liabilities consist mainly in having the cognitive capacity, training, and skills to interpret and thus, repurpose data which in turn, allow them to decode and reuse data. Also, researchers belong to a relational structure with specific organizational and epistemic norms and a reward system that may condition data reuse.

Also, research data are data-reuse mechanism objects with a relational structure and causal powers and liabilities. Data may have several interpretations and serve as evidence of scientific claims; data curation can change the 'state' of data, e.g. from unprocessed to processed state; data can become conceptually or technologically obsolete; and data availability (e.g. open, closed-but-available-for-use, proprietary) and data ownership can change.

All parts of the mechanism (Figure 1) have an *initial* and *final* value because the values of the elements, relational structure, and any of the mechanism's conditions can change across time (Sayer 2000, 2010). For the data-reuse mechanism, we have theorized three outcomes. Outcome (1): after consideration/trial use, secondary data are not used; Outcome (2): use of secondary data happens, but reuse is not shared with the research community, and the data are not used to support scientific claims but serve to widen the researcher's background knowledge or cognitive framework and potentially could trigger new research hypotheses; Outcome (3): use of secondary data happens, and secondary data are used exclusively as evidence of the scientific claim.

In this study, *data* represent phenomena in any kind of organizational setting, whether scholarly or not, that the individual or group of researchers regard as potential or

definitive evidence of the scientific claims in the context of a concrete scientific inquiry (Leonelli 2016, 2015). *Reuse of data* or *use of secondary data* is defined as the process of employing data to answer a novel research question by a researcher working within the institutional norms of science, who did not participate in the collection or generation of the particular data. Since data reuse is a contextually embedded phenomenon, case study is an appropriate study approach (Baxter and Jack 2008; Yin 2003). In the present paper, the case study is an attempted use of secondary data as 'evidence of phenomena' (Borgman 2015, xviii) and support for scientific claims to be shared with the scientific community via publication or other means, or exploited as background knowledge. Background knowledge or cognitive framework can be either tacit or explicit knowledge created by the researcher because it is 'important to [the researcher's] research activities but … not […] reported in publications' (Wallis, Rolando, and Borgman 2013, 4). Although the mechanism is theorized to work only if the research horizon is a scientific contribution, the inclusion of background knowledge as a potential output allows comparison of the mechanism conditions for each of the outcomes.

3. Methods

Our choice to focus on bio-health sciences is justified by data reuse in these fields becoming increasingly relevant for the production of scientific knowledge as integration and combination of data from different repositories and scientific disciplines has the potential to improve human health in previously unforeseen ways (Leonelli 2016). We chose a multiple-case approach since multiple cases enable theory confirmation (Gerring 2004; Stake 2006). In the context of our analysis, the case studies rather than being relevant *per se* are instrumental based on their support and contribution to answering our research question (Stake 2005).

3.1. Sampling

We chose 'purposive, rather than random' sampling (Miles and Huberman 1994, 27) based mainly on data access regimes and the data curation services which are conducted mainly by 'professionals familiar with the datafile structures and with the population of current users' (David and Steinmueller 1991, 53). Our rationale for comparing data reuse with different access regimes and different levels of curation (Figure 2) is that reusers may abandon reuse of poorly curated or uncurated data and data with more restrictive access. The data access regimes considered are: open data, closed-but-available-for-use (CAU) data, and proprietary data. Despite definitional disagreements related to open data (Borgman 2015), we consider open data as publicly released data in repositories with no restrictive access by pecuniary, confidentiality, or technical barriers. CAU refers to data deposited in a repository and available for reuse on request but not publicly released or shared. These data can be mediated also by ethical, economic, or technical walls. Finally, proprietary data are not publicly released data whose availability for reuse is thus *a priori* uncertain since they remain private to the data collector or producer, and are curated and packaged only on request by a potential reuser.

Following Hyman's (1972) approach, we also compare reuse of the same data by different reusers and how the same researchers reuse different data (Figure 2). The

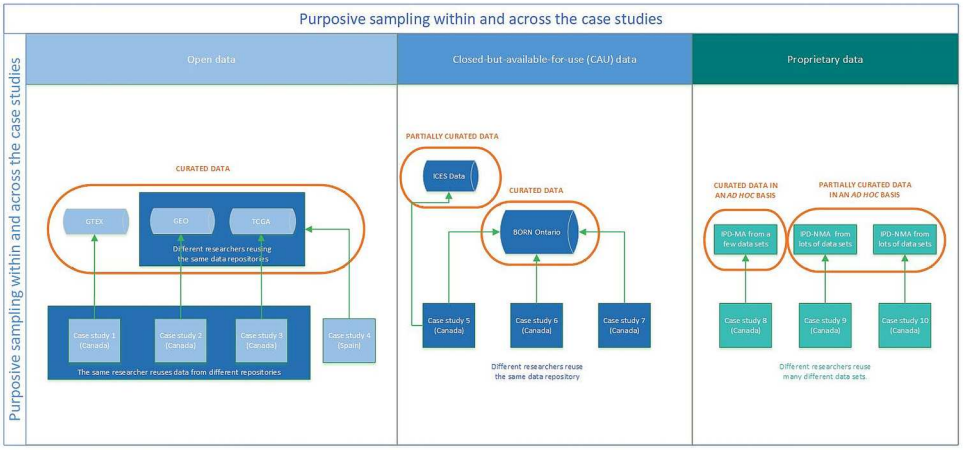

Figure 2. Purposive sampling.

open and CAU case studies allow us to compare how data from the same repositories (GEO Profiles, TCGA, BORN Ontario) are used by different researchers and how the same researcher reuses data from different repositories (case studies 1, 2, 3, 4 and 5). The case studies of proprietary data allow us to compare the location of and the request for data from multiple clinical trial studies by three different researchers. Unlike Hyman (1972), who only studied completed data reuse processes, we include ongoing cases of data reuse in order to track changes to the mechanism conditions and the sequence of events over time (Abbott 2001; Pettigrew 1990; Saldaña 2003). This was also aimed at minimizing participant bias in retrospective accounts of their actions (Tsoukas 2010). We also recruited cases of completed data reuse which allowed us to ensure that data reuse had occurred either as evidence for a scientific contribution or creation of background knowledge.

3.2. Case studies

All case studies comply with our definition of data reuse and are from the field of biohealth sciences. Four cases correspond to molecular and computational biology – three molecular biology in ovarian cancer and one computational biology in breast cancer. The researchers concerned reused data from GTEx (Genotype-Tissue Expression), GEO (Gene Expression Omnibus) Profiles, and TCGA (The Cancer Genome Atlas). All three repositories store curated genomic, epigenomic, transcriptomic, and proteomic data, and gene expression profiles. Curation of the data in these repositories is ongoing. Access is open but requires some identification and registration details to access clinical variables.

The other six cases are in the area of epidemiology and differ in terms of data access regime and institutional setting of primary data collection. Three cases reused CAU health administrative data from the BORN (Better Outcomes Registry & Network) Ontario data repository, and three reused proprietary individual participant or patient data (IPD) collected in randomized clinical trials (RCT). The IPD consist of meta-analysis addressing the same health issue but not the same research question(s) that motivated the original

data collection for each of the RCTs. According to our participants, curation levels differed across the six epidemiology case studies (see Figure 2 and Table 1). Table 1 summarizes the characteristics of each case study.

3.3. Data collection and analysis

With the exception of one case study in Spain included to increase variability, most data were collected in Canada. After obtaining ethics clearance, a recruitment message was sent to the directors of different Ottawa Hospital Research Institute health units and a BORN Ontario data repository knowledge translation specialist, asking them to forward the message to researchers in their units. We checked that all those who responded to our recruitment message met the sampling criteria. We also used the snowballing method to find cases.

Data collection involved 25 semi-structured open-ended interviews, lasting between 30 and 60 min, mainly but not exclusively with data reusers. With the exception of one reuser who did not respond to our request for a second interview, all reusers were interviewed at least twice. Data collection and analysis were conducted simultaneously at different stages (see Figure 3 corresponding to case study #4 and *Analysis* pp. 51 to 61 in the supplemental material). Ongoing cases were analyzed diachronically at two moments in the data reuse process – the beginning (time 1) and the end (time 2). In the cases of a completed data reuse process, we asked participants to provide us with a retrospective narrative detailing progression through their data reuse process (which might last months or years) from beginning (time 1) to the end (time 2).

The interpretive qualitative analysis was grounded in a 'systemic' approach to understanding the interaction of the variables in a complex environment (Miles and Huberman 1994), focused on reusers' decision-making with respect to their horizon – or scientific goal – and all the elements of the data-reuse mechanism. Decision-making is not observable directly (Tsoukas 2010); thus, our units of observation and analysis were constituted by the sequence of events prompted by the researcher's data reuse. Using process-tracing methods (Beach 2017; Beach and Pedersen 2013; Bril-Mascarenhas, Maillet, and Mayaux 2017) we translated the occurrences or actual happenings (Abbott 2001) in the data reuse process into researcher decisions. In their first interview, the researchers were asked to describe the data reuse process and detail the related challenges and why they chose to overcome them. Based on this information we constructed data reuse process workflow diagrams that participants could validate in the second interview (e.g. Figure 4 and *Analysis* pp 40–50 in the supplemental material). These diagrams supported a proper understanding of both the process and the researcher's decision-making logic.

Publications can be used to complement the analysis of data reuse by providing details perhaps overlooked by interviewees (Hyman 1972); therefore, we also analyzed interviewees' publications that used secondary data. In addition, we examined the characteristics of the data repositories and some of the data sets from the randomized clinical trials and conducted within and across case analyses. Each individual condition and element was compared against the other conditions and elements in order to identify patterns in the data.

Table 1. Details of the ten case studies.

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8	Case 9	Case 10
Type of data based on the access regime	Open Data				Closed-but-available-for-use (CAU) health administrative data				Proprietary data	
Reuser(s)/ researcher(s)	David Cook (real name with participant's permission)	David Cook	David Cook	Jaume Forés and Joan Climent (real names with participants' permission)	Deshayne Fell (real name with participant's permission)	Mary Smith (pseudonym)	Sarah Wilson (pseudonym)	Nicole Langlois (real name with participant's permission)	Claire Johnson (pseudonym)	Areti Angeliki Veroniki (real name with participant's permission)
Data repository	GTEx	GEO Profiles	TCGA	GEO Profiles and TCGA	BORN Ontario & ICES	BORN Ontario	BORN Ontario	Data sets from 8 randomized clinical trials (RCT)	Data sets from 67 RCTs	Data sets from 108 RCTs
Were the data curated?	Yes	Yes	Yes	Yes	ICES (Partially) BORN (Yes)	Yes	Yes	On an ad hoc basis	Partially and on an *ad hoc* basis	Partially and on an *ad hoc* basis
Ongoing or completed	Completed	Ongoing	Completed	Ongoing	Completed	Completed	Ongoing	Completed	Ongoing	Ongoing
Name of project or research question	Chromatin regulators: jacks of all states	Transcriptional and epigenetic determinants of the epithelial-to-mesenchymal transition in ovarian cancer	Undetermined name (related to cancer)	The gene and biological relationship between the autism spectrum and cancer: the role of TRIM29	Influenza illness and influenza vaccination during pregnancy and risk of preterm birth and fetal death	The effect of maternal obesity on stillbirth and neonatal death	A research study in epidemiology related to maternal and neonatal health	Low-molecular-weight heparin and recurrent placenta-mediated pregnancy complications: a meta-analysis of individual patient data from randomized controlled trials.	Tentative title: Mass deworming to improve developmental health and wellbeing of children in low-income and middle-income countries	Tentative title: Comparative safety and effectiveness of cognitive enhancers for Alzheimer's dementia: a systematic review and individual patient data network meta-analysis
Type of reward or social commitment		Obtain a PhD		Obtain a PhD	Obtain a PhD	Publication(s)	Publication(s)	Publication(s)	Publication(s)	Publication(s)
Sub-discipline and topics	Molecular biology (applied to research on ovarian cancer)	Molecular biology (applied to research on ovarian cancer)	Molecular biology (applied to research on ovarian cancer)	Computational biology (applied to research on breast cancer)	(Perinatal) Epidemiology	Epidemiology	Epidemiology	Clinical in nature, but it can fall into epidemiology	Epidemiology. Global health.	Clinical in nature but it can fall into epidemiology
Country	Canada	Canada	Canada	Spain	Canada	Canada	Canada	Canada	Canada	Canada
Number of interviews with the reuser(s)	3	2	2	3 with Jaume Forés and 2 with Joan Climent	3	2	3	2	2	2
Number of interviews with other participants	Director of the Cancer Therapeutics Program			–	1 interview with an ICES data steward & curator and 2 interviews with a Knowledge Translation Specialist of the BORN Ontario Repository			–	–	–

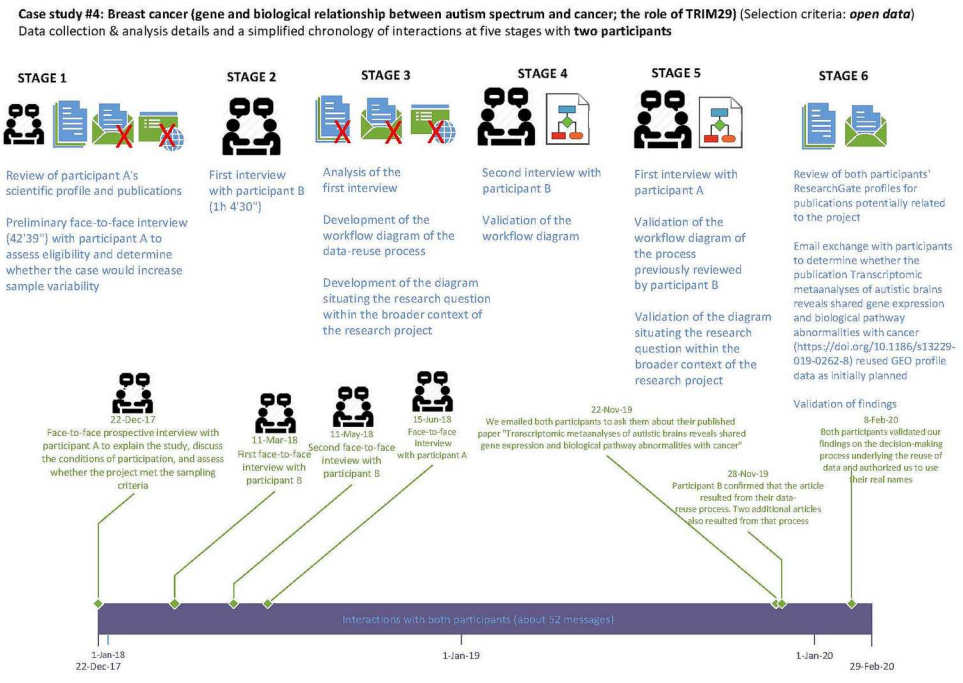

Figure 3. Details of data collection and analysis (case study #4).

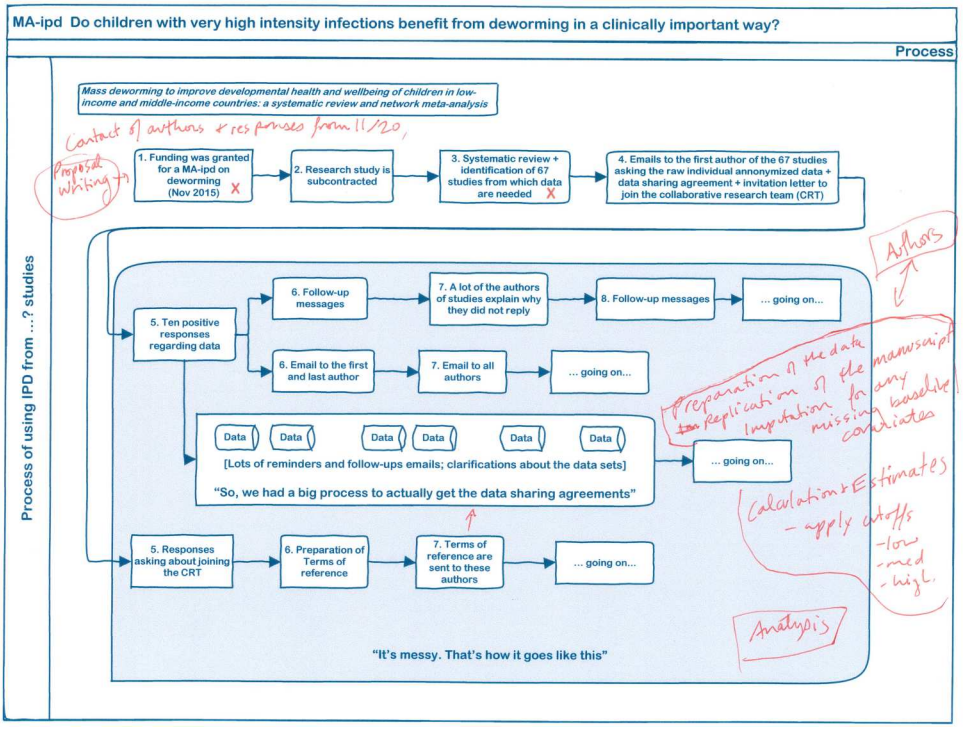

Figure 4. Workflow diagram of the data reuse process reviewed by the participant (red handwritten text). Case study #9.

4. Findings and discussion

Our discussion and findings are presented in bulleted paragraphs in subsections 4.1 to 4.4. Whenever appropriate in each section, we refer to the specific access regime and level of data curation.

4.1. (Un)anticipated challenges

Although our participants had engaged with data sets that they believed they would be able to access and use, they made decisions and initiated courses of action without knowing many of the actual costs and challenges that would be involved when accessing and analyzing the data.

4.1.1. Locating, finding, and obtaining data from different access regimes

All participants knew about the data that they thought they could use for their contributions and their source. However, in the case of proprietary data, although clinical trial studies were relatively easy to identify from publications, as Argie (Areti Angeliki Veroniki) explained, it was not always easy to identify the holder of the data set or to obtain the data owner's contact details or the specific study identification number (ID):

> I was surprised but we couldn't locate the specific study ID that the sponsor required, and, on the other hand, I was really surprised that some of the sponsors could not locate the studies and they were provided with the title, with the journal, even the publication, the authors, all these details we were able to identify, […] but they weren't able to identify the study in their database. (case #10)

It took Argie nearly three years to obtain enough, though not all, of the proprietary data from the clinical trials sponsors – mostly pharmaceutical companies – to make a scientific contribution. This prolonged process caused conflict among the team's members, with some favoring waiting to obtain more data before publishing the contribution, and some who felt that they had waited long enough since there was no certainty about how many more data sets they would be able to obtain.

Among the researchers involved in the six epidemiology cases, compared to researchers accessing CAU and open data, those involving especially individual patient or participant data (IPD) experienced even greater difficulties when trying to obtain data. Claire Johnson explained that there was uncertainty about the amount of IPD that her team could obtain:

> But we're optimistic that we will have enough. 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 … (case #9)

Deshane Fell was unable to obtain all the variables she needed for her study, a limitation that, similar to Claire's team, she acknowledged and later addressed in the manuscript. In one case where there was agreement about sharing proprietary data, the first author of one clinical trial data set died before the data were shared – an unexpected event which resulted in the team working on thrombophilia having to conduct the IPD meta-analysis without the deceased author's data. Also, open and CAU data came with some access costs. Jaume Forés and David Cook could not have anticipated encountering restrictive-access walls in an open access data repository to access less unprocessed data

and clinical variables from TCGA. Although Sarah Wilson was very familiar with BORN Ontario data, it took her some time to realize that the repository did not contain the variables she had expected to find.

4.1.2. Decoding and interpreting data with different levels of curation

All the interviewees had excellent cognitive capabilities as a result of their education and training and the skills they had acquired. Nevertheless, all of them referred to problems related to interpreting and understanding some of the variables, and especially those representing the facts or objects that the data represented. Deshayne told us that understanding data hosted at the Institute for Clinical Evaluative Sciences (ICES) was harder than doing the statistical analysis:

> I would say 80%, easily 80%, of the work is just in understanding those various databases, manipulating them, linking them all together, and getting a data set that you can actually analyze. Only 20% is the analysis. That's kind of the easy part. The hard part is understanding and synthesizing these big data. (case #5)

Although GTEX data had a higher level of data curation than ICES data, David found the interpretation of GTEX data equally difficult:

> Format is tricky […] when you are looking at these types of data you never know what you are looking at. […] uhh … so … understanding the data is more than just a formatting thing. What am I looking at? You know, what did the numbers in every cell, what exactly do these mean […]? (case #1)

4.1.3. Ethical, legal, and bureaucratic challenges

The data also presented ethical, bureaucratic, and legal hurdles. In trying to obtain individual patient data (IPD) from clinical trials, Claire Johnson (case #9) realized that she would probably be unable to get access to some of the data sets due to differences in the study ethical review process before and after 1990. Studies published before 1990 did not require an ethics certificate number or a consent form. Deshayne Fell needed to link data from two different institutions, and told us that she got 'caught in the legal and administrative negotiations between the two organizations that had to happen in order to allow that transfer of data … '. She could never have anticipated the length of time it would take for BORN Ontario and ICES to sign a data sharing agreement:

> 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: 'This data sharing agreement will take 18 months' […] And so you never – like I never knew when I would ultimately [obtain the data-sharing agreement]. (case #5)

Argie and her team also experienced a delay of a few months to allow their 'lawyer to make several amendments to the data-sharing agreements' (case #10).

4.1.4. Technical obstacles

While there are no legal impediments to using open data, the process involved some technical issues, which have made decoding the data a huge task requiring huge computational capacity to process TCGA data:

> I got access to [data]. But then only to find that it was completely unprocessed data that we would have to do, and we're talking of 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. (case #3)

In Claire's case, the problems were software-related; the team needed a range of different software to open the data files, and some members did not 'have the right software' for a particular file. They also '[did] not know which file [had] the data [they] want[ed] so that [was] a little bit more challenging' (case #9). For Mary Smith (case #6), the problem was related to obtaining secondary data in the format she needed.

4.2. (A)synchronic interdependencies among conditions and resulting uncertainty

In this subsection, we show how changes to one of the elements or conditions affect other elements and conditions. In most cases, the data reuse process was lengthy. The most extreme case was case study #10: it took Argie nearly three years to obtain not all but just enough proprietary data from the clinical trials' sponsors to make a scientific contribution. Supplemental material *Analysis* pp. 35 and 36 provides a summary of the temporal analysis, which shows that changes to/impacts on all elements and conditions of the theorized data-reuse mechanism are likely to occur during such a prolonged process. As discussed in subsection 4.1, most of the problems encountered could not have been anticipated; they mainly emerged asynchronously and introduced a high level of uncertainty into the data reuse process.

4.2.1. Changes to researchers and data

To our knowledge, researchers' cognitive frameworks, training, and skills to obtain, interpret, and thus repurpose data remained constant in every case study. However, over time, the researchers inevitably developed a better understanding of the data as they obtained more variables, metadata, or data codebooks.

In case study #7, the researcher's relational structure changed. Initially, Sarah Wilson as a public health officer whose performance was not driven by the reward system in science began reusing BORN Ontario data. Thus, at the beginning of the data reuse process, conditions C3 (data are an initial satisficing option) and C5 (an expected scientific contribution exists) were not met because Sarah's goal was not to make a scientific contribution but to produce data-backed knowledge as the basis for an epidemiological surveillance report for the public health office. This affected how Sarah Wilson could access data and the types of recommendations and claims she could make based on her findings. When the public health office decided to stop the study, Sarah decided to continue working on the same area as a researcher affiliated to a research institution. This changed her relational structure and meant that conditions C3 and C5 were met. When we asked her for more detail on how her relational structure affected her causal powers and liabilities, she responded:

> Yeah, okay. Well, I'll start with the transfer. So, a transfer can happen between BORN and the [public health] agency, but the amount of paperwork and process involved is at a very high level. So, it would be the hospital – I don't actually understand the BORN process, but in our place it would be senior management negotiating like memorandums of agreement and

> understanding. […] For the purposes of why I joined BORN [as a public health officer], that data transfer didn't quite exist because that's institution to institution, right?. […] When I work with the government and we're trying to find something in surveillance, we don't make policy recommendations. I'm not a policy analyst. I'm a surveillance analyst or an epidemiologist. So, what I say is the rates of diabetes are increasing, continued surveillance is important, and rates of diabetes are increasing in this province and this province. […]. I never make the policy recommendation. I don't make the intervention recommendation. […] And so I don't do that in my role because that's just not my role and that's just not how we operate. If I'm doing this outside in academia I can say whatever I want because now I can say, 'Okay, the rates are going up. We need to weight management interventions. We need earlier screening'. And I can be very prescriptive with what I think is important based on the research I've done because I'm no longer coming from the agency because there's a lot of implications there, right? (case #7)

Sarah finally published a scientific contribution related to maternal and neonatal health with BORN Ontario data as a researcher and not as a public health officer.

In case study #3, the data properties changed, and this affected condition C3 (data are an initial satisficing option). Initially, the TCGA project data were satisficing (C3 was met) and were expected to be sufficient to provide background knowledge to help David's lab with its experiments. However, it was soon realized that the TCGA data were unprocessed (raw data). David and the rest of the team decided that the information that they expected to obtain to support their experimentation with primary data did not justify the time and extra effort that would be required to process the data. They lost interest in TCGA data due to their lack of technological and computing capacity required to process the data (C3 was unmet). However, sometime later, David revisited the data in the TCGA repository and found that the data collectors had now released processed data, which made the TCGA data satisficing (C3 was now met again).

4.2.2. How access regimes determine the sequence of conditions C1, C2, and C3

With open data, conditions C1 (researcher knows the data exist), C2 (data are obtained), and C3 (data are an initial satisficing option) may occur almost simultaneously. However, in the case of CAU data, condition C2 is not necessarily met before the researcher finds the data satisficing (C3). For instance, in case #6, after Mary obtained the data (C2), the value of condition C3 changed and the data were no longer a satisficing option. Mary requested record-level data from BORN staff for a student project due to begin in the summer for a four-month period. However, Mary and her student were able to access only data rates and not even aggregated data. Obtaining record-level data would have taken between 12 and 18 months.

In the case of proprietary data, there is constant uncertainty about obtaining all the data required. Our findings show that this uncertainty can persist over several years. While initially, before the data are obtained, they may seem satisficing (C3), if the researcher manages to obtain the data, this can change the value of this condition (C3). Once data are obtained, they can stop being satisficing. In cases #8, #9, and #10, the researchers needed multiple data sets from several different clinical trial studies. Obtaining the data was impeded by legal and ethical requirements related to the data-sharing process. Conditions C1, C4 (the idea of collecting primary data is not a satisficing

option), and C5 were met. However, condition C3 could be met or unmet only after data were obtained (C2).

Our analysis shows that despite not having all the data sets and all the values, the interviewees had *enough* to allow them to write and publish their scientific contributions, meaning that condition C2 (data are obtained) was met in part. The (long) wait for access to the data was sustained by the researcher's horizon (see discussion in section 4.3).

4.3. Scientific horizon and researcher tenacity

Uncertainty about eventually obtaining the data and whether the variables and their format would be usable meant that the interviewees were forced to exercise patience or abandon the data reuse process. Argie described how the decision to be patient was often accompanied by serious doubts:

> Of course I've thought about [giving up]. 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 […] 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. (case #10)

In most cases, the long wait was sustained by the researcher's willingness or need to pursue a particular scientific horizon i.e. to make a publishable scientific contribution, obtain a PhD degree, etc. Only in case #1 did David Cook abandon pursuing his research goal after considering his personal network of social, material, and institutional relations in the context of his bounded individual horizon and deciding that other scientific goals were more important:

> As for the GTEx project, I kind of gave up on writing [the perspectives article with GTEx data]. I thought the work I was doing on the project was informative, but I ended up prioritizing some other projects [in the lab] that I figured may be more beneficial to me [for his research career]. (case #1)

Case study #3 exemplifies how background knowledge can trigger data reuse but may not be sufficient to sustain the process if the problems that arise are considered too great to try to overcome. Initially, TCGA data were to be used as background knowledge to help the lab with its experiments, but when the team realized the scale of the barriers to reuse, they stopped the process. David explained that:

> 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. (case #3)

Not all decisions and actions were related to uncertainty about waiting times. Researchers were required continuously to make decisions and take action to adjust an initially envisioned contribution to the data they were able to locate, access, decode, and recontextualize. These changes depended on the researcher's flexibility in relation to sacrificing an initial scientific horizon as discussed below.

4.4. Sacrificing the initial scientific horizon with cognitive flexibility and practical actions

On reaching a self-imposed or required scientific horizon deadline, the researchers reassessed their expectations about the research goal and research method based on constant dialogue related to fitting secondary data and research questions in a dialogical way. These proactive and flexible cognitive-related actions included tweaking the research questions based on the limitations encountered or use of a different research method. For instance, when Deshayne realized that she had missing information on influenza for one of the five women in her study, she 'had to drop that outcome completely' from the peer-reviewed paper (case #5). Mary originally requested individual BORN Ontario data, but due to BORN staff time constraints and work overload, was given access to only some aggregated data. She and her master's student then 'modified [their initial research question] based on the data [they] were able to obtain in the timeframe that [they] needed [to carry out the study]' (case #6).

On finding that data on patients taking insulin and those not taking it were poorly captured due to the many unknowns in the variable, Sarah's (case #7) solution was to tweak her research question. She also adapted the research method by regrouping the variable values. Similarly, in the case of GTEX data, David said that he:

> revaluate[d] how [he] was looking at, and it kind of inspired a new research question. […] So, [he] had to rework the way [he] was kind of looking at the data, rather than saying let's look at the expression gene A or gene B.

When we asked him to define his research question, he responded that:

> I have a pretty good rough idea … what my question is right now, but after all the data comes together, I mean, the question might shape and change a little bit. (case #1)

Argie, who was frustrated by the tedious and slow process of obtaining or accessing data from pharmaceutical companies, considered adapting the research questions or adapting 'the protocol that [they] published in the beginning to what [they] ha[d] at hand so far'. When we told her that our understanding was that protocols could not be modified, she agreed and responded:

> But in an appendix, let's say, we will say what are the differences from the protocol and the current paper, current complications. So we have to be clear and transparent, and we will mention all the differences […]. And expect them to be all. For example, even from the methodology part, I was hoping they would apply in one-stage analysis, but now I'm forced to apply two-stage analysis because I don't have all the data in my computer. I have access to specific platforms for specific data, and so this actually forces me to do a case analysis. So, these differences should be clearly reported in the [appendix]. (case #10)

Other respondents resorted to alternative practical and managerial resources and actions. Claire's employment of purposive actions to overcome a wide range of practical problems associated with early phases of secondary data collection is illustrative:

> […] 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?' And then we have studies going back to 1975 and so some of them have moved on. So, one of them had retired, one of them has gone with no forwarding address, and there are

> still 10 that we haven't identified a contact. […] 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. […] It's messy. That's how it goes like this. […] Another one, we offered support, she didn't take us up on it. She's the dean, so she's very busy and the data [are] at another institution. She finally wrote to the other institution because we asked her, we put some pressure on her from friends and colleagues. […] And then, the third one, we offered support. Oh, because she was having a terrible crisis in her country and had to deal with it, and we said that we could provide some of the support if that would help to get the data, 'cause it's a recent study and we know she has the data, and she said, 'It's just too much of a crisis, that won't help'. Finally, we put some pressure on her from higher professionals than me and she sent the data. (case #9)

Even in the case of authors who were willing to share their (proprietary) data, the re-use process was not straightforward. Some authors lacked the resources and time to curate and reformat the data. Claire 'offer[ed] [them] a stipend to support like research assistant time to prepare the data or whatever'. Another impediment to sharing data was the pre-1990 RCT ethical review process. Claire 'asked for an addendum to [their] ethics and they granted it, that authors could simply write [them] a letter to explain what ethical procedures were followed and why they [couldn't] provide it'.

Cognitive and practical actions were complementary, intertwined, and often concurrent. Only in retrospect did Argie realize that her contribution had probably not been worth the effort, time, funding, and energy invested. She explained the relevance of her scientific horizon for supporting her efforts to reuse the data:

> So, what I would say is that I believe it depends on the clinical field, on the clinical question, so sometimes we're desperate and we need answers. Questions that have not been answered and cannot be answered using aggregated data. So in this case, I would say that, 'yeah, it's worth all this effort'. But now I still can't answer the clinical question using the aggregated data. No, I wouldn't try that again. So, it depends on what you're trying to do, what you're trying to answer. (case #10)

Without the willingness to sacrifice initial scientific horizons and without the cognitive capacity and tenacity to make multiple adjustments, such as tweaking research questions or employing different methods to the available data and variables, our interviewees' data reuse process would have stopped at an early state, and the outcome would have been #1 of the theorized data-reuse mechanism; that is, no use of secondary data after trying or considering the options.

5. Conclusion

This study seeks to contribute to a longstanding debate about the codification, circulation, and reuse of knowledge in scientific research. Specifically, our focus was the debate highlighted by Prof. Steinmueller and colleagues more than 20 years ago (Cohendet and Steinmueller 2000; David and Steinmueller 1991; Foray and Steinmueller 2003) over the challenges associated with effective codification and decodification of knowledge, which are necessary to realize the full potential of knowledge sharing and data reuse in science. We suggest that this debate is even more relevant today when so much scientific knowledge is produced and codified in the form of research data that, in principle, is available for reuse by others.

We examined ten cases across molecular/computational biology and epidemiology and showed that data reuse is a dynamic, socially embedded process shaped by the interplay of both synchronic and diachronic relations among different elements and conditions that significantly increase the uncertainty involved in the process. It is precisely this uncertainty in combination with researchers' tenacity, scientific horizon, and flexibility in sacrificing a horizon that enable reuse of data. Our analysis highlighted the centrality of the researcher's bounded horizon (Aleixos-Borrás and López 2025) in sustaining reuse. Limited time, limited resources, and institutional conditions, for example, as well as the researcher's relational structure can be constraints, but determination to contribute to science can result in purposive actions – either practical actions such as negotiating access or cognitive actions such as redefining the research questions. This dual role of agency underscores that data reuse is neither merely technical nor solely cognitive but is a dialogic process of fitting the data context to the researcher's scientific goals (Aleixos-Borrás and López 2025).

Our findings demonstrate also that while access regimes, curation levels, and other socio-technical aspects are important, they do not fully determine whether or how data are reused. In particular, our evidence challenges the assumption that restrictive access regimes and poorly curated data deter reuse. In our study, in the case of proprietary or poorly curated data (case studies #8, #9 and #10) and CAU data (case studies #5, #6 and #7), participants did not abandon the data reuse process. In fact, when the goal is a significant scientific contribution, researchers invest considerable efforts to overcome even very costly obstacles. Open data (case studies #1, #2, #3 and #4) allow a data reuse process that is more efficient in terms of the time and effort spent on accessing data.

It would seem that the researcher's determination to achieve a meaningful scientific contribution is a decisive factor in pursuit of data reuse. That is, the scientific goal is more important than the convenience or availability of (non-)curated data, and what ultimately triggers and sustains data reuse. Researchers persist despite unanticipated costs, uncertainties, and technical or legal barriers, and mobilize both practical/managerial and cognitive strategies to adjust their project and expectations to what the data afford. This is explained by the theorized data-reuse mechanism that also shows that uncertainty, intertwined relations with a scientific horizon, and satisficing decision-making rather than optimal calculation often underpin sustained efforts over time.

These insights contribute to the literature both conceptually and methodologically. Conceptually, we have highlighted the need to reinsert researcher agency in contemporary discussion of open science, data infrastructures, and the so-called FAIR – findable, accessible, interoperable, and reusable principles, which show that data reuse depends as much on human motivation and capacity as on technical systems. Methodologically, the data-reuse mechanism offers a useful framework for future inquiries in other disciplines and contexts. The mechanism is conceived as the result of the researcher's ongoing decisions to achieve a prospective goal (a scientific contribution) for which the researcher expects to be rewarded. The mechanism draws attention to the extent to which it is the envisioned scientific contribution that motivates and fosters the process of data reuse which is often given sparse attention in existing accounts. In other words, the theorized explanatory model is underpinned by an understanding of causation deriving from teleological decision-making theory, in which X (the cause)

does not lead to Y (the effect). Rather, it is Y (the effect) that triggers, impels, and sustains X (the cause). It is the researchers' scientific horizon that 'puts the process [of reusing data] in motion' (Poole and Van de Ven 2010, 551). Thus, the data-reuse mechanism captures the process of calculating a bounded decision which ultimately allows the researcher to reuse data in order to create knowledge and make a scientific contribution.

Our study focused on two health-related subfields and a small number of cases. Future research should aim to broaden the empirical base by exploring how the data-reuse mechanism operates in other domains, under different institutional conditions, and under different reward structures. This might reveal alternative or conjunctural mechanisms that sustain data reuse in science.

In sum, data reuse can be understood as a satisficing process in which the researcher's agency is mobilized in pursuit of a scientific contribution. Data, infrastructure, and institutional context matter and the findings from this study stress the need for policy support for accessibility to and robust curation of data. We also show that ultimately it is the researcher's vision of making a knowledge contribution that sets the reuse process in motion and sustains it over time.

Acknowledgements

We are grateful to all who participated in our study. We are grateful also to two anonymous reviewers for their feedback and to feedback from the journal's guest editors Aldo Geuna and Maria Savona. Collection of data from the nine Canadian case studies was made possible by project NOVA DOMUS-CHEMEDPHO (Ref.: 545700-EM-1-2013-1-ES-ERAMUNDUS-EMA22)

Author contributions

CRediT: **Inma Aleixos-Borrás:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Writing – original draft, Writing – review & editing; **Pablo D'Este:** Conceptualization, Formal analysis, Methodology, Writing – review & editing; **José J. López:** Conceptualization, Formal analysis, Methodology, Writing – review & editing.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by NOVA DOMUS CHEMEDPHO (ERAMUS MUNDUS ACTION 2) [grant number 545700-EM-1–2013-1-ES-ERAMUNDUS-EMA22; 2015].

ORCID

Pablo D'este http://orcid.org/0000-0002-4040-8209

References

Abbott, Andrew Delano. 2001. *Time Matters: On Theory and Method*. Chicago: University of Chicago Press.

Aleixos-Borrás, Inma, and José Julián López. 2025. "A Hermeneutic Dialogical Understanding of Data Reuse across Different Access Regimes." *Science as Culture* 34 (1): 114–135. https://doi.org/10.1080/09505431.2024.2375218.

Ancori, B., A. Bureth, and P. Cohendet. 2000. "The Economics of Knowledge: The Debate about Codification and Tacit Knowledge." *Industrial and Corporate Change* 9 (2): 255–287. https://doi.org/10.1093/icc/9.2.255.

Ansbacher, H. L. 1950. "The Problem of Interpreting Attitude Survey Data." *Public Opinion Quarterly* 14 (1): 126–138. https://doi.org/10.1086/266155.

Baxter, Pamela, and Susan Jack. 2008. "Qualitative Case Study Methodology: Study Design and Implementation for Novice Researchers." *The Qualitative Report* 13 (2): 544–559. https://doi.org/10.46743/2160-3715/2008.1573.

Beach, Derek. 2017. "Process-Tracing Methods in Social Science." In *Oxford Research Encyclopedia of Politics*. online edn, edited by Erin Hannah. New York, NY: Oxford Academic. https://doi.org/10.1093/acrefore/9780190228637.013.176.

Beach, Derek, and Rasmus B. Pedersen. 2013. *Process-tracing methods: Foundations and guidelines*. Ann Arbor: University of Michigan Press.

Borgman, Christine L. 2007. *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Cambridge, MA: The MIT Press.

Borgman, Christine 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, Christine L. 2015. *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge: The MIT Press.

Bril-Mascarenhas, Tomas, Antoine Maillet, and Pierre-Louis Mayaux. 2017. "Process Tracing. Inducción, Deducción e Inferencia Causal." *Revista de Ciencia Política* 37 (3): 659–684. https://doi.org/10.4067/S0718-090X2017000300659.

Caldas, Alexandre. 2003. "Are Newsgroups Extending 'Invisible Colleges' into the Digital Infrastructure of Science?" *Economics of Innovation and New Technology* 21 (1): 43–60. https://doi.org/10.1080/10438590303123.

Cohendet, Patrick, and W. Edward Steinmueller. 2000. "The Codification of Knowledge: A Conceptual and Empirical Exploration." *Industrial and Corporate Change* 9 (2): 195–209. https://doi.org/10.1093/icc/9.2.195.

Cowan, Robin. 2001. "Expert Systems: Aspects of and Limitations to the Codifiability of Knowledge." *Research Policy* 30 (9): 1355–1372. https://doi.org/10.1016/S0048-7333(01)00156-1.

Cowan, Robin, Paul A David, and Dominique Foray. 2000. "The Explicit Economics of Knowledge Codification and Tacitness." *Industrial and Corporate Change* 9 (2): 211–253. https://doi.org/10.1093/icc/9.2.211.

Curty, Renata Gonçalves. 2015. "Beyond 'Data Thrifting': An Investigation of Factors Influencing Research Data Reuse in the Social Sciences (2015)." Dissertations-ALL.266. http://surface.syr.edu/etd/266.

Curty, Renata Gonçalves, Kevin Crowston, Alison Specht, Bruce W. Grant, and Elizabeth D. Dalton. 2017. "Attitudes and Norms Affecting Scientists' Data Reuse." *PLoS One* 12 (12): 1–22. https://doi.org/10.1371/journal.pone.0189288.

David, Paul A, and W. Edward Steinmueller. 1991. "The Impact of Information Technology upon Economic Science." *Prometheus* 9 (1): 35–61. https://doi.org/10.1080/08109029108631904.

David, Paul A., and W. Edward Steinmueller. 2003. "Introduction." *Economics of Innovation and New Technology* 12 (1): 1–3. https://doi.org/10.1080/10438590303122.

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

Faniel, Ixchel M., Rebecca D. Frank, and Elizabeth Yakel. 2019. "Context from the Data Reuser's Point of View." *Journal of Documentation* 75 (6): 1274–1297. https://doi.org/10.1108/JD-08-2018-0133.

Faniel, Ixchel M., Adam Kriesberg, and Elizabeth Yakel. 2012. "Data Reuse and Sensemaking among Novice Social Scientists." *Proceedings of the ASIST Annual Meeting* 49 (1): 1–10. https://doi.org/10.1002/meet.14504901068.

Finholt, Thomas A. 2003. "Collaboratories as a New Form of Scientific Organization." *Economics of Innovation and New Technology* 12 (1): 5–25. https://doi.org/10.1080/10438590303119.

Foray, Dominique, and W. Edward Steinmueller. 2003. "The Economics of Knowledge Reproduction by Inscription." *Industrial and Corporate Change* 12 (2): 299–319. https://doi.org/10.1093/icc/12.2.299.

Garmon Bibb, Sandra 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, John. 2004. "What is a Case Study and What is it Good For?" *American Political Science Review* 98 (2): 341–354. https://doi.org/10.1017/S0003055404001182.

Gläser, Jochen. 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. https://www.ts.tu-berlin.de/fileadmin/fg226/TUTS/TUTS-WP-1-2012.pdf.

Gläser, Jochen, and Grit Laudel. 2013. "Life With and Without Coding: Two Methods for Early-Stage Data Analysis in Qualitative Research Aiming at Causal Explanations." *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* 14 (2): 1–23. https://doi.org/10.17169/fqs-14.2.1886.

Gregory, Kathleen, Paul Groth, Andrea Scharnhorst, and Sally Wyatt. 2020. "Lost or Found? Discovering Data Needed for Research." *Harvard Data Science Review* 2 (2): 1–51. https://doi.org/10.1162/99608f92.e38165eb.

Hemphill, Libby, Amy Pienta, Sara Lafia, Dharma Akmon, and David A. Bleckley. 2022. "How Do Properties of Data, Their Curation, and Their Funding Relate to Reuse?" *Journal of the Association for Information Science and Technology* 73 (10): 1432–1444. https://doi.org/10.1002/asi.24646.

Huang, Can, and Luc Soete. 2025. "Reconciling Open Science with Technological Sovereignty." *Economics of Innovation and New Technology* 8599:1–15. https://doi.org/10.1080/10438599.2025.2459764.

Hyman, Herbert Hiram. 1972. *Secondary Analysis of Sample Surveys*. New York: Wiley.

Irwin, Sarah, and Mandy Winterton. 2011. "Debates in Qualitative Secondary Analysis: Critical Reflections Paper." 3349 (4): 1–23. ISSN: 1758 3349.

Kim, Youngseek, and Ayoung Yoon. 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.

Latour, Bruno. 1987. *Science in Action: How to Follow Scientists and Engineers through Society*. Cambridge: Harvard University Press.

Latour, Bruno. 2005. *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford: Oxford University Press.

Latour, Bruno, and Steve Woolgar. 1986. *Laboratory Life: The Construction of Scientific Facts*. Princeton, NJ: Princeton University Press.

Leonelli, Sabina. 2015. "What Counts as Scientific Data? A Relational Framework." *Philosophy of Science* 82 (5): 810–821. https://doi.org/10.1086/684083.

Leonelli, Sabina. 2016. *Data-Centric Biology: A Philosophical Study*. Edited by The University of Chicago Press. Chicago: The University of Chicago Press.

Mansell, Robin, and W. Edward Steinmueller. 2000. *Mobilizing the Information Society. Strategies for Growth and Opportunity*. Oxford, NY: Oxford University Press.

Marshall, Catherine, and Gretchen B. Rossman. 2011. *Designing Qualitative Research*. Los Angeles: SAGE Publications, Inc.

Mayntz, Renate. 2004. "Mechanisms in the Analysis of Social Macro-Phenomena." *Philosophy of the Social Sciences* 34 (2): 237–259. https://doi.org/10.1177/0048393103262552.

Miles, Matthew B., and A. Michael Huberman. 1994. *Qualitative Data Analysis: An Expanded Sourcebook*. California: SAGE Publications, Inc.

Murillo, Angela Patricia. 2016. "Data Sharing and Data Reuse: An Investigation of Descriptive Information Facilitators and Inhibitors (Dissertation)."

Niu, Jinfang. 2009a. "Overcoming Inadequate Documentation (Dissertation)."

Niu, Jinfang. 2009b. "Perceived Documentation Quality of Social Science Data."
Olson, Gary, and Judith Olson. 2003. "Mitigating the Effects of Distance on Collaborative Intellectual Work." *Economics of Innovation and New Technology* 21 (1): 27–42. https://doi.org/10.1080/10438590303117.
Pasquetto, Irene V. 2018. "From Open Data to Knowledge Production: Biomedical Data Sharing and Unpredictable Data Reuses." https://escholarship.org/uc/item/1s1814cj.
Pasquetto, Irene V, Bernadette M Randles, and Christine L Borgman. 2017. "On the Reuse of Scientific Data." *Data Science Journal* 16 (8): 1–9. https://doi.org/10.5334/dsj-2017-008.
Pettigrew, Andrew 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. https://doi.org/10.1287/orsc.1.3.267
Pienta, Amy M, Dharma Akmon, Justin Noble, Lynette Hoelter, and Susan Jekielek. 2018. "A Data-Driven Approach to Appraisal and Selection at a Domain Data Repository." *International Journal of Digital Curation* 12 (2): 362–375. https://doi.org/10.2218/ijdc.v12i2.500.
Poole, Marshall S, and Andrew Van de Ven. 2010. "Empirical Methods for Research on Organizational Decision-Making Processes." In *Handbook of Decision Making*, edited by Paul C Nutt and David C Wilson, 543–580. Chichester: John Wiley & Sons, Ltd.
Reichenbach, Hans. 1938. *Experience and Prediction: An Analysis of the Foundations and the Structure of Knowledge*. Chicago: University of Chicago Press.
Rojas, Fabio. 2017. *Theory for the Working Sociologist*. New York : Columbia University Press.
Saldaña, Johnny. 2003. *Longitudinal Qualitative Research: Analyzing Change through Time*. Walnut Creek, CA: AltaMira Press.
Sayer, Andrew. 2000. *Realism and Social Science*. London: Sage Publications.
Sayer, Andrew. 2010. *Method in Social Science: A Realist Approach*. Oxon: Routledge.
Simon, Herbert A. 1955. "A Behavioral Model of Rational Choice." *The Quarterly Journal of Economics* 69 (1): 99–118. https://doi.org/10.2307/1884852.
Simon, Herbert A. 1957. *Models of Man: Social and Rational. Mathematical Essays on Rational Human Behavior in a Social Setting*. New York: Wiley.
Simon, Herbert A. 2000. "Bounded Rationality in Social Science: Today and Tomorrow." *Mind & Society* 1 (1): 25–39. https://doi.org/10.1007/BF02512227
Stake, Robert E. 2005. "Qualitative Case Studies." In *The SAGE Handbook of Qualitative Research*, edited by Norman K Denzin and Yvonna S Lincoln, 443–466. Thousand Oaks: Sage Publications.
Stake, Robert E. 2006. *Multiple Case Study Analysis*. New York: The Guilford Press.
Steinmueller, W Edward. 2000. "Will New Information and Communication Technologies Improve the 'Codification' of Knowledge?" *Industrial and Corporate Change* 9 (2): 361–376. https://doi.org/10.1093/icc/9.2.361
Tsoukas, Haridimos. 2010. "Strategic Decision Making and Knowledge: A Heideggerian Approach." In *Handbook of Decision Making*, edited by Paul C Nutt and David C Wilson, 379–402. Chichester: John Wiley & Sons, Ltd.
Wallis, Jillian C, Elizabeth Rolando, and Christine L Borgman. 2013. "If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology." *PLoS One* 8 (7): e67332. https://doi.org/10.1371/journal.pone.0067332.
Wolfe, David A., and Matthew Lucas. 2001. "Investing Knowledge in Universities. Rethinking the Firm's Role in Knowledge Transfer." In *Knowledge Management in the Innovation Process*, edited by John de la Mothe and Dominique Foray, 173–191. Boston, MA: Springer US.
Yin, Robert K. 2003. *Case Study Research: Design and Methods*. Thousand Oaks: Sage Publications.
Yoon, Ayoung. 2016a. "Data Reusers' Trust Development." *Journal of the Association for Information Science and Technology* 68: 946–956. https://doi.org/10.1002/asi.23730.
Yoon, Ayoung. 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, Ayoung. 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.

Zimmerman, Ann 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, Ann 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.

Zuiderwijk, Anneke, and Helen Spiers. 2019. "Sharing and Re-using Open Data: A Case Study of Motivations in Astrophysics." *International Journal of Information Management* 49:228–241. https://doi.org/10.1016/j.ijinfomgt.2019.05.024.