Data & Policy (2025), 7: e41
doi:10.1017/dap.2025.16

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Facing the ambiguities of participation in data-driven projects: a systematic literature review

Judith Fassbender[1,2], Irina Kuehnlein[2] and Tristan Henderson[1]

[1]School of Computer Science, University of St Andrews, St Andrews, UK
[2]Alexander von Humboldt Institute for Internet and Society, Berlin, Germany
Corresponding author: Judith Fassbender; Email: judith.fassbender@hiig.de

Received: 08 October 2024; **Revised:** 17 January 2025; **Accepted:** 24 March 2025

Keywords: data governance; data handling; participation; systematic literature review

Abbreviations: AI, artificial intelligence; PC, participation-at-core; PC-LG, Participatory-at-core including labour and governance; PI, participatory-informed; PI-G, participatory-informed focusing on governance; PI-L, participatory-informed focusing on labour

Abstract

Participation is a prevalent topic in many areas, and data-driven projects are no exception. While the term generally has positive connotations, ambiguities in participatory approaches between facilitators and participants are often noted. However, how facilitators can handle these ambiguities has been less studied. In this paper, we conduct a systematic literature review of participatory data-driven projects. We analyse 27 cases regarding their openness for participation and where participation most often occurs in the data life cycle. From our analysis, we describe three typical project structures of participatory data-driven projects, combining a focus on *labour and resource participation* and/or *rule- and decision-making participation* with the general set-up of the project as *participatory-informed* or *participatory-at-core.* From these combinations, different ambiguities arise. We discuss mitigations for these ambiguities through project policies and procedures for each type of project. Mitigating and clarifying ambiguities can support a more transparent and problem-oriented application of participatory processes in data-driven projects.

Policy Significance Statement

Our study provides policymakers with strategies to mitigate ambiguities around participation in data-driven projects. We propose a set of principles to enable more transparent, clearly communicated, and problem-oriented forms of participation, which are relevant for policymakers on different levels. Policymakers in funding institutions may embed those principles in their funding schemes, for example, by educating facilitators or asking for a declaration of how the principles are followed or not followed in funded projects. Policymakers within organisations facilitating participatory data-driven projects can follow our principles to guide their own projects and embed them in project-internal policies and procedures. In addition to the principles, our study includes projects where internal policy documents, such as data access protocols and data governance agreements, were developed as part of participatory approaches. Such approaches can help realise policies that are more aligned with the collective interest in their governance.

This research article was awarded Open Data badge for transparent practices. See the Data Availability Statement for details.

1. Introduction

Participation has become a mainstream topic in a variety of contexts (Marres, 2015, 16). In respect to data and data-driven technologies, participatory approaches are having a heyday; participation and power imbalances in data-intensive contexts, such as artificial intelligence (AI) systems, are discussed both in the academic debate (e.g. Falco, 2019; Bondi et al., 2021; Delgado et al., 2022) and in policy- as well as practice-focused environments (e.g., Mozilla, n.d.; Ada Lovelace Institute, 2021).

At the same time, dissatisfaction with various forms of participatory approaches prevails. In discussions of participatory matters, ambiguities between the aim of levelling power imbalances and the risks of ineffectiveness, exploitation, or window dressing are repeatedly highlighted (Arnstein, 1969; Birhane et al., 2022; Sloane et al., 2022). Groves et al. (2023) show that some practitioners in commercial AI Labs have concerns about engaging in potentially exploitative and tokenistic forms of participatory practices (Groves et al., 2023, 1169). Their study further shows a lack of a shared understanding of the value and utility of participation (Groves et al., 2023, 1168).

Recent empirical work on participatory data-driven projects in the academic literature focuses mainly on qualitative studies of a small number of cases (Falco, 2019; Birhane et al., 2022; Delgado et al., 2022). Kelty et al. (2015)provide a framework based on a much larger number of cases in which they outline dimensions of participation but do not explicitly focus on recommendations. The Ada Lovelace Institute (2021) provides a report that focuses on practical processes on a case-by-case basis and looks at generally facilitating participatory data stewardship. To complement these approaches and to contribute to a better understanding and addressing of the ambiguities of participation in data-driven projects, we conducted a large-scale systematic review of implemented projects described in the literature. We focus on the data level of the cases to derive implications that apply to a broader set of technologies.

For this purpose, we reviewed and screened 1,642 items and analysed the descriptions of 27 projects that rely on data for their functioning and incorporate participatory elements. The project descriptions were analysed from two angles: their openness to participation and the processes used within the data life cycle.

In the discussion, we look at two dominant forms of participatory processes—labour and resource participation and rule- and decision-making participation—and highlight aspects connected to participatory ambiguities, such as possible countermeasures. We define three types of participatory data-driven projects—projects that focus on participation in governance, projects that focus on participation in resource and labour aspects, and projects that combine both—to help address ambiguities and create more clarity in projects for participants and the public.

2. Background

The discussion of participatory approaches is often characterised by an ambiguity between an "attractive" image (Himmelreich, 2023)—in the sense of caring for people's interests, co-determination, and levelling power imbalances—and, on the other hand, the potential for exploiting participants and misusing the participatory appeal to evoke a desirable public image of a participatory project. This ambiguity was captured early on in the late 1960s in the much-cited Ladder of Participation (Arnstein, 1969). As the degree of power in the hands of the participants decreases, so too does the licence to call an approach actual participation, according to Arnstein. The highest rungs of the ladder—*citizen control*, *delegated power*, and *partnership*—implicitly carry the mentioned promises of participation, which others frame as the expectation to render a project "inclusive, equitable, robust, responsible, and trustworthy" (Birhane et al., 2022). The hopes in participatory practices pave the way for evoking a false-positive image, which has been referred to as "participation-washing" (Sloane et al., 2022). Others point to forms of participation as "window dressing" (Gilman, 2022, 507) when instrumentalised by facilitators to enforce contested decisions. Within Arnstein's framework, such practices would all be located on the lower ranks—*placation*, *consultation*, *informing*, *therapy*, and *manipulation* (Arnstein, 1969). Critiques connected to further ambiguities point at instances when participatory methods are used to exploit participants for

https://doi.org/10.1017/dap.2025.16 Published online by Cambridge University Press

laborious tasks to enhance private interests (Resnik et al., 2015; Sloane et al., 2022). An additional important point of critique concerns the use of participation to shield power imbalances (Birhane et al., 2022; Riley and Mason-Wilkes, 2024). Next to deliberate efforts by facilitators to reduce their own responsibilities through implementing participatory processes, a diffusion of roles in participation and co-creation organically brings a diffusion of responsibilities (Steen et al., 2018; Zehner and Ullrich, 2024). The outlined aspects show that the ambiguities surrounding participation in data-driven projects are multi-layered. They relate to the question of what actually qualifies as participatory, what forms a participatory project can take, and the lack of clarity about the purpose, roles, and processes of participatory approaches.

To support a better analysis of participatory approaches in various dimensions, Kelty et al. (2015) developed a framework based on comprehensive case studies in fields related to information studies. The seven dimensions that they draw out describe modalities as well as effects of participation in those cases: *educative dividends* for participants (1), participation in *tasks* and setting *goals* (2), participation in *controlling resources* (3), the possibility for participants to *exit* (4), and *voice* (5) their opinions, showing outcomes of participation in *visible metrics* (6) and the *affective/communicative capacity* (7) experienced through participation. They further point out the diversity in what is understood as participation.

The ambiguities of participatory approaches presumably result in part because of the lack of an agreed-upon definition of participation. Understandings of participation range from theoretical differentiations, for example, between a sociological and a political view, to procedural differentiations in participation. Carpentier (2016) contrasts a sociological perspective of participation as being part of a process—for example, buying something is understood as participating (Carpentier, 2016, 71f)—with a political perspective, in which participation is understood as a means to counteract power asymmetries with a focus on decision-making (Carpentier, 2016, 72f). Schrögel and Kolleck (2018) highlight siloed discussions within the discourse of participatory science between *doing science* in and a *dialogue about science* (Schrögel and Kolleck, 2018, 78). Sloane et al. differentiate in the context of AI systems between participation as *work*, as *consultation*, and/or as *justice*; the first two modes describe processes where it is designed *for* participants, whereas the latter describes a mode of designing *with* participants. Sloane (2024) further describes AI systems as "deeply participatory" (Sloane, 2024, 2), as the data used could only be collected with the participation of technology users, bringing us closer to the described sociological understanding of participation.

While Sloane et al. (2022) and Birhane et al. (2022) provide categories for assessing participation in regards to AI systems, guidance for how to handle ambiguities in participatory data-driven projects is developing sparsely and in siloed discourses, for example, in the citizen science discourse (Resnik et al., 2015). The academic debate on the participatory process in data-driven projects mostly focuses on particular technologies. The predominant focus is on participation in machine learning in specific application contexts, such as smart cities (Falco, 2019), law (Delgado et al., 2022), or on affected groups (Queerinai et al., 2023). In regard to normative approaches, authors study the possibility to realise ethical or socially good AI through participation in capacity-building approaches (Bondi et al., 2021) or value-sensitive design methods (Gerdes, 2021). Given that the fine-grained implementation of participatory approaches is context-dependent, as can be seen, for example, in projects regarding indigenous data sovereignty (Pyper et al., 2018; Love et al., 2022), a comparison of a larger number of cases is needed to provide insights into common project types. We therefore conduct a systematic literature review to analyse participatory data-driven projects. In addition to reviewing a larger number of cases for comparison, we focus on the data level and specifically look at participation in the data life cycle (a model that describes the higher-level work steps relating to the data aspect of a project in chronological order; e.g., Faundeen et al., 2014). The choice of data-driven projects remains relevant to technologies such as machine learning, but is not limited to one technology. Kelty et al. (2015), as well as researchers from the Ada Lovelace Institute (2021), provide frameworks and guidance applicable to broader data-driven technologies based on large-scale case studies; we aim to complement

https://doi.org/10.1017/dap.2025.16 Published online by Cambridge University Press

their work by focusing on project structure in combination with participatory processes, and how to handle the ambiguities of participation across different project types.

Our research questions are:

RQ1 How is participation in data-driven projects described in the empirical literature with regard to the openness of the project for participation and the nature of the processes along the data life cycle?

RQ2 What typical project types can be found, and how can we clarify participatory ambiguities in the process for participants and the public?

3. Method

To identify cases of data-driven projects with a participatory element, we conducted a systematic literature review (Kitchenham and Charters, 2007) by surveying the literature and then analysed the identified 27 cases from two angles:

1. We analysed the structural *openness* of the cases using the Participatory Science Cube (Schrögel and Kolleck, 2018) and mapped who participates and the degree of participation in data handling and decision-making.
2. We looked at *where* and *how* in the data life cycle participatory processes were implemented and described these processes.

The *Participatory Science Cube* (Section 3.2.1) is crucial for our understanding of participation in this paper. Within our analysis, we identify something as participation as soon as external parties who do not belong to the facilitating organisation either actively participate in a data handling task and/or are included in a participatory process regarding the governance. There is no intention, however, to make this a normative measure for the evaluation of participation. We consider projects to be data-driven when the handling or processing of data within the project is a necessary aspect for the functioning of the project.

3.1. Systematic literature review

To extract cases from the literature, we followed Kitchenham and Charters (2007) by selecting research databases, constructing two search strings to filter potentially relevant papers, and defining inclusion and exclusion criteria, which were employed by two reviewers to systematically identify relevant cases.

To cover the most relevant conference proceedings and journals in computer science, we chose the *ACM Digital Library* (Association for Computing Machinery, 2024) and the *IEEE Xplore* (Institute of Electrical and Electronics Engineers, 2024) as data sources. Given the interdisciplinary nature of the topic, *Scopus* (Elsevier, 2024) was added to cover other disciplines.

We derived our search terms from a report on participatory data stewardship (Ada Lovelace Institute, 2021) with additional terms based on our knowledge of the subject matter and discussions with fellow researchers.

We divided the search into two parts to separate search rationales (Table 1). Search String I combined modes of handling data (Data Stewardship, Data Governance, and Data Curation) with attributes that refer directly to participation or indicate a participatory character (e.g., participatory, cooperative, or democratic). Search String II considered modes that implicitly indicate a participatory approach or an orientation towards collective interests (e.g., Data Stewardship, Data Commons, Data Trusts, or Data Curation). In addition, we restricted the search to a 10-year period (2012–2022) and papers written in English. The search string was adapted depending on the options of the different databases.

The search was conducted on 20 September 2022, and 1,642 items were found and imported to the Zotero reference manager (Corporation for Digital Scholarship, 2024). A detailed listing of the search results is part of Table 2; Figure 1 shows the reduction of the results, and Table 3 contains the inclusion and exclusion criteria used to do so.

https://doi.org/10.1017/dap.2025.16 Published online by Cambridge University Press

Table 1. *Search strings used in the literature review*

Name	Rationale	Search string
Search String I	Mode of data handling + attribute implying participation, mentioned within the abstract, in a publication between 2012 and 2022, marked as a research article	[[Abstract: "data stewardship"] OR [Abstract: "data governance"] OR [Abstract: "data curation"]] AND [[Abstract: participatory] OR [Abstract: participation] OR [Abstract: deliberative] OR [Abstract: deliberation] OR [Abstract: collaboration] OR [Abstract: collaborative] OR [Abstract: fair] OR [Abstract: citizen] OR [Abstract: community] OR [Abstract: collective] OR [Abstract: cooperative] OR [Abstract: social] OR [Abstract: democratic] OR [Abstract: donated] OR [Abstract: donation] OR [Abstract: accountable]] AND [Publication Date: (2012–01–01 TO 2022–12–31)] Applied Filters: Research Article
Search String II	Mode of data handling inherently implying participation, mentioned within the abstract, in a publication between 2012 and 2022, marked as a research article	[Abstract: "data stewardship"] OR [Abstract: "data commons"] OR [Abstract: "data trust"] OR [Abstract: "data care"] OR [Abstract: "data cooperative"] AND [Publication Date: (2012–01–01 TO 2022–12–31)] Applied Filters: Research Article

Table 2. *Initial numbers of papers found in the databases*

Database	Initial extraction (date: 20 September 2022)
ACM Digital Library	Search String I: 47 Search String II: 27
IEEE Xplore	Search String I: 61 Search String II: 110
Scopus	Search String I: 602 Search String II: 795
Total	1,642

3.2. Analysis of the cases

To apply the *Participatory Science Cube*, we analysed the identified cases qualitatively and coded the cases using MAXQDA (VERBI Software, 2024). We describe the derived models and coding categories below.

3.2.1. Participatory science cube

The *Participatory Science Cube* (Schrögel and Kolleck, 2018) is a model to systematically describe and compare participatory science projects. It was developed on the basis of models from science governance, citizen science, and other participatory research frameworks, which were combined with the Democracy Cube by Fung (2006). The dimensions represent the following questions: Who participated? (reach); How is knowledge produced? (epistemic/doing); What ought to be done? (normative/deciding) (Schrögel and Kolleck, 2018, 87–91).

https://doi.org/10.1017/dap.2025.16 Published online by Cambridge University Press

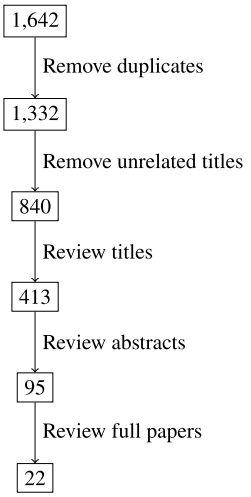

Figure 1. *Reducing the total sample of papers.*

Table 3. *Exclusion and inclusion criteria*

Exclusion criteria	Inclusion criteria
1. The paper is not written in English. 2. The paper is incomplete. 3. The paper is not peer-reviewed. 4. The title does not indicate a possibility of participation in data handling or decision-making. 5. The described case is described in more detail in another paper in the batch. 6. The paper contains too little information to analyse it.	1. The paper describes an implemented case of a data-driven project with a participatory element. 2. The described case includes participation of entities from outside of the organisation that facilitates the project in data handling and/or decision-making in the project. 3. The described case is implemented.

While the framework is designed to map participatory science projects, most of the questions still apply to data-driven projects. To use the cube, we needed to make some minor adjustments, described below and in Figure 2. In the original paper, the scales are defined as a spectrum, and the cube was intended to be used for singular cases and as a support for qualitative case descriptions. To compare and categorise cases, we needed to apply distinct values on each axis for each project. To incorporate the notion of the spectrum, we introduced half values, representing projects between stages. To assign values, we introduced distinct definitions of the stages described below. These were based on the original terminology from the cube, but we further introduced half values for where a case is positioned between two stages. Finally, we position "Facilitators" where "Scientists" were originally positioned in the model. By facilitators, we refer to those who are responsible for implementing participation in a project, which can be units or people in an organisation or autonomous actors. An organisation can be, for example, a university/research group, a collective, or an NGO.

Reach dimension: The *reach* dimension moves between *experts* and the *broad public* (Schrögel and Kolleck, 2018, 88) and evaluates *who* is addressed by the participatory processes. Experts may come from different fields, such as academia, NGOs, politics, or industry, and their expertise can relate to technological or domain knowledge. Domain knowledge can refer to specific subjects as well as lived experiences. The greater the range of the reach dimension, the lower the specificity of the expertise that

https://doi.org/10.1017/dap.2025.16 Published online by Cambridge University Press

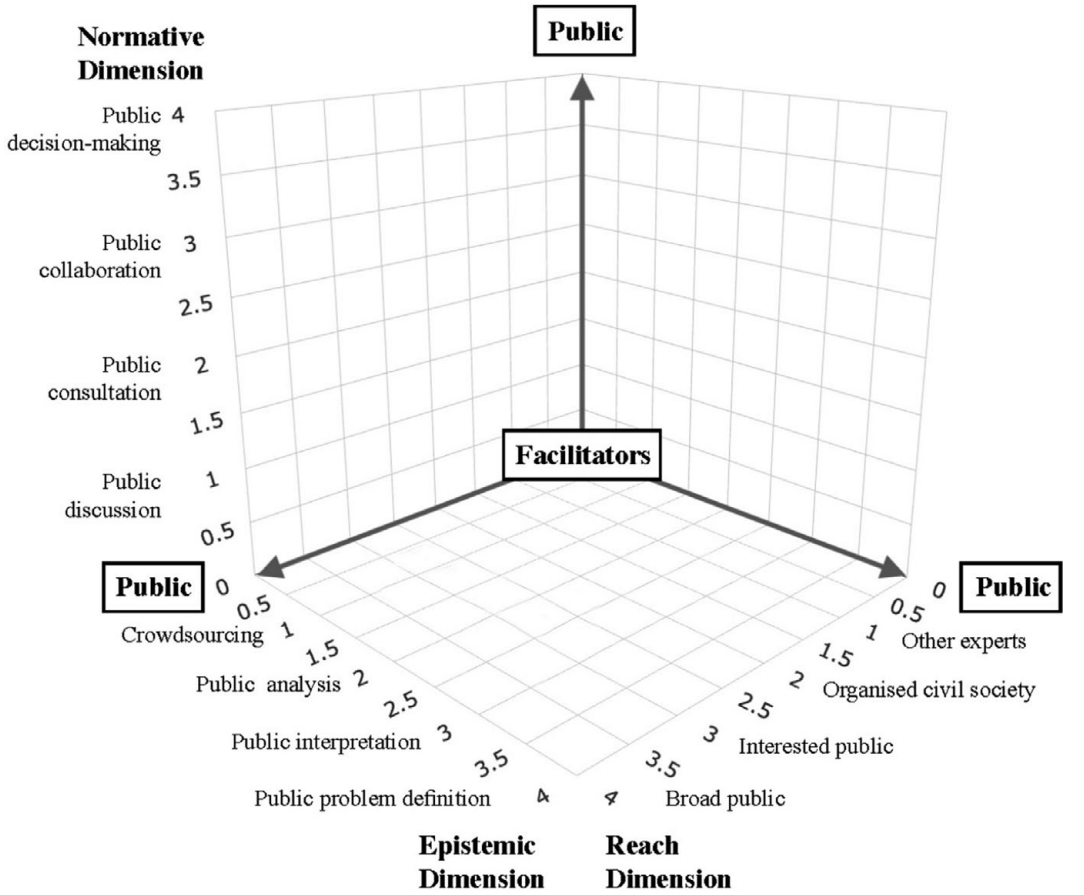

Figure 2. *The Participatory Science Cube.*
Source: Adapted from Schrögel and Kolleck *(2018); we replace "Scientists" with "Facilitators" to make it applicable to a wider range of cases.*

can be assumed. While members of a civil society organisation can participate as experts, they can also participate as representatives of collective interests. Depending on how they are included in the project, the evaluation changes. For cases in which this distinction is not sharp, we assign half-values (Table 4).

Normative dimension: The *normative* dimension focuses on "questions of values and norms as well as questions of preferences and interests" (Schrögel and Kolleck, 2018, 89). It deals with decision-making processes on governance questions that may result in project policies, budget decisions, or restrictions within the project. We define the stages based on increasing decisive power in the hands of participants. Low-ranking examples include general discussions or information sessions, whereas high-ranking examples include co-designing policies or policy-making (Table 5).

Epistemic dimension or data dimension: The *epistemic* dimension deals with the degree of participation in knowledge production; we use it in application to data handling while using the same idea and terms of the scale, but with a slightly different focus. As data are inherent in knowledge production processes, we use this dimension to describe the participation in data handling with a decreasing degree of restrictions put on participants in handling the data (Table 6).

The cases are analysed based on the information given in the source papers (for more details, see the Supplementary Material; Fassbender et al., 2025). We focused on the openness of the projects for participants and, therefore, did not analyse the kinds of facilitators but mentioned them in the case descriptions in the Supplementary Material.

https://doi.org/10.1017/dap.2025.16 Published online by Cambridge University Press

Table 4. *Definition of the stages of the reach dimension with their values on the scale*

Stage of the reach dimension	Value	Description
Other experts	1	Regards the inclusion of experts outside of the facilitating organisation in their professional capacity
Organised civil society	2	Regards the inclusion of people who are part of an organisation that represents a special interest
Interested public	3	Considers people who act in a private capacity and have some kind of precursor, connecting them to the project, a special interest in the topic, or a specific attribute such as a health condition or a place of residence
Broad public	4	Describes an unspecified participant group; when no precondition is needed, a process is theoretically open to participate with no restriction

Table 5. *Definition of the stages of the normative dimension with their values on the scale*

Stage of the normative dimension	Value	Description
Public discussion	1	Describes general informing of the public or exchange between participants and facilitators without a concrete decision at stake or any kind of bindingness
Public consultation	2	Describes processes in which input from the participants on a concrete decision or matter is gathered with the intent to implement it
Public collaboration	3	Describes a shared decision-making process between participants and facilitators (e.g., including a negotiation)
Public decision-making	4	Describes that decisive power is in the hands of the participants and the facilitators enacting the decision

Table 6. *Definition of the stages of the epistemic dimension with their values on the scale*

Stage of the epistemic dimension	Value	Description
Crowdsourcing	1	Participants knowingly provide data, but have no further influence on the data
Public input for analysis	2	Participants collect data but are closely guided by the facilitators
Public collaboration for interpretation	3	Participants handle the data with less guidance from the facilitators, and have more impact own on the data that are collected and/or how they are curated
Public problem definition and interpretation	4	Participants define what problem is solved on the basis of the data

3.2.2. Data life cycle

The cube gives a structural perspective on the openness of the project towards participants. For a better understanding of the underlying processes and project realities, we added a contextualised perspective, employing the data life cycle to see *where* and *how* participatory processes were employed.

Data life cycle models provide a version of the data handling process in procedural steps. A variety of models exist, differing in the level of detail and to what degree they aim to capture the messy realities

https://doi.org/10.1017/dap.2025.16 Published online by Cambridge University Press

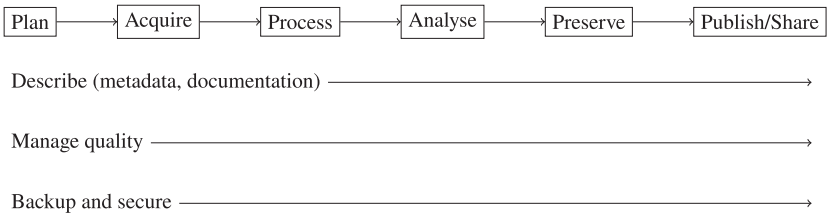

Figure 3. *Visualisation of the data life cycle.*
Source: Visually adapted from Faundeen et al. (2014, 2).

underlying the idealised depictions. Faundeen et al. (2014) provide a data life cycle model for scientific data (Figure 3). The model is simplified to a comparably high degree, which is useful for considering a larger number of heterogeneous cases. The life cycle provides the following steps: *plan*, *acquire*, *process*, *analyse*, *preserve*, and *publish/share.* Underlying tasks, concerning the whole cycle, are framed as cross-cutting model elements: *describe (metadata* and *documentation*), *manage quality*, and *backup and secure* (Faundeen et al., 2014). The description of the general steps can be found in Table 7.

4. Results

Our analysis identified different categories, which we use as a lens to view participation in data-driven projects. In the first step, we group the cases according to their conceptual set-up (Section 4.1). The application of the participatory cube showed two different project layouts regarding openness for participation (Section 4.2). The analysis of the focus points of participatory processes in the data life cycle showed clear hotspots of participatory attention and suggests two kinds of participatory processes (Section 4.3).

4.1. Emergent groups of cases

We cluster the cases in four separate groups: *Research Projects*, *Data Collections*, *Data-Driven Products/Services*, and *Data Activism Initiatives* (Table 8). The groups are not mutually exclusive and are based on the dominant framing of the cases in the source paper; for example, if a project is explicitly described as a data activism initiative but also has a research basis, we placed it in the data activism group. The majority of these projects follow or support an epistemic aim and are connected to research endeavours; 11% of the cases are projects that used data for an application in a service or product. *Health & Medicine* and *Environmental* topics dominate the field with encapsulating 77% of the cases in the sample.

4.1.1. Research Projects

The first group comprises 11 *Research Projects* (R1-11). Six projects (R1, R2, R4, R5, R7, R8, and R9) concentrate on *environmental issues*; two cases deal with *health* conditions (R3 and R10) and two with

Table 7. *Steps in the data life cycle as defined in* Faundeen et al. *(2014)*

Plan	Focuses on the design of the project with all its elements; Faundeen et al. suggest that the data management plan is the main output of this stage
Acquire	Concerns the collection of data through different techniques, either collecting new data or existing data for reuse
Process	Preparation of the data to be processed further, concerning its (inter)operability, defining elements, calibrations, to get the data in a state to be analysed
Analyse	Describes the exploration of the data via statistics, testing of hypotheses in the model
Preserve	Concerns the long-term storage of the data
Publish/share	Focuses on the accessibility of the data for external parties and the connected rules for that step

https://doi.org/10.1017/dap.2025.16 Published online by Cambridge University Press

Table 8. *The case studies fall into four groups*

Data Collections	Research Projects	Data-Driven Products/Services	Data Activism Initiatives
C1 BigMouth (Walji et al., 2022) **C2** Childhood Obesity Data Initiative (CODI) (Kraus et al., 2022) **C3** Dementias Platform UK (DPUK) (Bauermeister et al., 2020) **C4** MIDATA.coop (Vayena and Blasimme, 2017) **C5** National COVID–19 Chest Imaging Database (NCCID) (Cushnan et al., 2021) **C6** OneFlorida Data Trust (Hogan et al., 2022) **C7** The Open Data Commons for Spinal Cord Injury (odc-sci.org) (Torres-Espín et al., 2021) **C8** Paediatric Cancer Data Commons (Plana et al., 2021) **C9** PIONEER Hub in Acute Care (Gallier et al., 2021) **C10** RPGEH: Research Program on Genes, Environment and Health (Tai et al., 2019)	**R1** Analysing Indigenous Cultural and Natural Resource Management (Robinson et al., 2021) **R2** Astrophysics Data Systems All-Sky Survey (Cohen et al., 2015) **R3** Citizen Science Symptom Study (Murray et al., 2021) **R4** Co-VITAS (Aubin et al., 2020) **R5** EcoPrairie (Baker and Karasti, 2018) **R6** Game with Words (Cohen et al., 2015) **R7** Great Backyard Bird Count (Cohen et al., 2015) **R8** iNaturalist (Cohen et al., 2015) **R9** NABat (Reichert et al., 2021) **R10** Pathways TB Project (Love et al., 2022) **R11** SETI at Home (C. Cohen et al., 2015)	**P1** Crowdsourcing Open Pedestrian Network Data (Bolten and Caspi, 2022) **P2** Chemical Hazard Data Commons (Kokai et al., 2020) **P3** Wikidata/ Wikiprojects (Kanke, 2021)	**A1** Fatal Encounters (Currie et al., 2019) **A2** Making Sense (Kosovo) (Currie et al., 2019) **A3** Data Rescue (Walker et al., 2018)

other topics (R6 and R11). While the projects take place in academic or academia-related environments, 6 of the 11 cases are further framed as citizen science projects (R2, R3, R6, R7, R8, and R11) 5 of those (R2, R6, R7, R8, and R11) are described in the same paper (C. Cohen et al., 2015).

4.1.2. Data Collections

Ten cases focus on creating data collections to be shared with researchers (C1–C10). All 10 cases deal with *health* data, which mostly stem from care facilities and are compiled from patient treatment records. Each collection is organised around a shared topic, such as specific diseases, conditions, treatment situations, or a geographical area. We assume that in most cases, datasets are created, but we do not know the exact technical organisation of the data collections.

https://doi.org/10.1017/dap.2025.16 Published online by Cambridge University Press

4.1.3. Data-Driven Products/Services

The third group is *Products/Services*, where data enables an application or service (P1–P3), in distinction to an epistemic aim. This group includes one mapping case (P1) and two encyclopaedia-related cases (P2 and P3).

4.1.4. Data Activism Initiatives

Three *Data Activism Initiatives* (A1–A3) are all mentioned in one paper, while one case (A3) occurs in one more paper of the sample and is analysed on the basis of Walker et al. (2018) as the description is more elaborate. The projects are framed as data activism (Currie et al., 2019, 1): "groups devoted to representing a contentious political issue through data, either by producing their own data, collecting "missing data," or keeping vulnerable data in the public domain" (Currie et al., 2019, 2). The projects deal with different topics: preservation of public data feared to become censored (A3), documentation of incidents of police killings (A1), and a political activist campaign based on air pollution data (A2).

4.2. Openness of the cases to participation: participatory-at-core or participatory-informed

Locating the cases in the Participatory Science Cube (Schrögel and Kolleck, 2018) identified two dominant participatory project characteristics: projects that are *participatory-at-core* and projects that are *participatory-informed.* Those are structural characteristics that should not be used in isolation to describe a project, but rather provide the first element for building the project types we later discuss.

4.2.1. Participatory-informed

A characteristic that we term ***participatory-informed (PI)*** regards projects that have a strong focus on their openness for participation, either on governance participation or on participation in data handling. Therefore, we find projects with this characteristic in two different areas of the participatory cube, which also describe two starkly different types of projects. In one area, we see projects that rely on participation in data handling but not in the project governance (Figure 4). Those cases are open to a broader audience, ranging from *organised civil society/interested public* (2.5) to an *interested public* (3). These projects are typically citizen science projects (R2, R3, R6, R7, and R11) where participation takes place in the data acquisition, analysis, and processing. Another example is a data activism case (A1), in which participants submit data collected from different sources to the project to support building a statistic on killings by the police in the United States; the project is hierarchical and run by one journalist; therefore, no participation in governance takes place (Figure 5). The cases in this area share their strong focus on labour participation.

In the other area, we see projects that include participation in project governance but implement hands-on participation to a minimal degree (C2, C4, C5, C7–C10, and R5) (Figure 5). These projects are mainly health data collections. How participation is realised differs in who is participating and how the participation is facilitated. Patients (data subjects) and scientists (data users) are the main participating audiences. The inclusion of patients differs by the directness of their impact and their representation. In one case, facilitators interview a variety of stakeholders regarding their attitude towards data sharing (C10). These interviews informed the facilitators in drafting policies, which has a lower direct impact on participants but includes their positions in the policy-making process. The other extreme can be found in a data cooperative, where participants decide on the data access requests (C4). Between these ends, we find a case facilitated by the NHS where patients took part as in a workshop to co-draft the data sharing procedures (C9). While in the mentioned examples, patients represented patient interests as lay-experts, in other cases, professional experts are the main participants (C1, C2, C5, and C7). Here, experts take part in decision-making aspects of a project, for example, via data access committees (C1, C5, and C7) in which participants represent different perspectives or co-decide on the governance structure of the project (C2). The cases in this area show a varying degree of governance participation but share a focus on participation in the normative aspects of the projects.

4.2.2. Participatory-at-core

The other project characteristic is termed ***participatory-at-core (PC)*** and regards projects that employ participatory processes at different moments in the data life cycle and combine participation in

https://doi.org/10.1017/dap.2025.16 Published online by Cambridge University Press

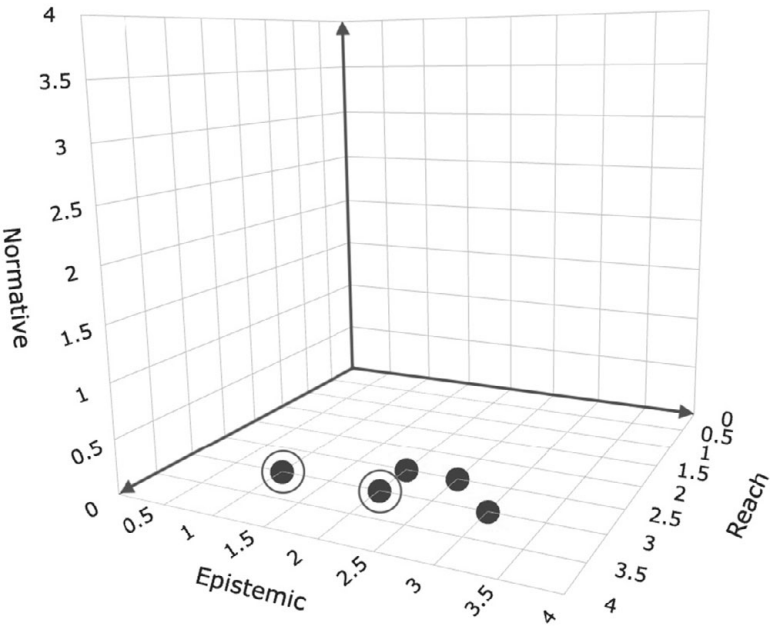

Figure 4. *PI-L projects and their position in the participatory science cube; each dot represents one case, and each circle represents an additional case. PI-L cases tend to have a high reach in the participant group; those participants tend to be an interested public. This is matched with a focus on participation in data handling (epistemic dimension) and a tendency for* no participation in the data governance of the project *(normative dimension).*

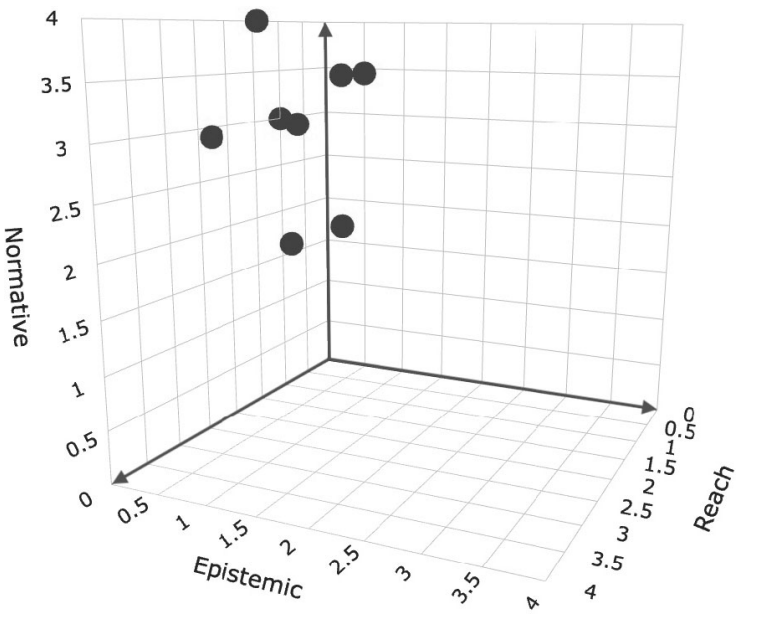

Figure 5. *PI-G projects and their position in the participatory science cube; each dot represents one case. PI-G projects tend to have a lower reach in the participant group; those participants can be called lay-experts and/or experts. This is matched with a focus on participation in the governance of the project (normative dimension) and a tendency for* very low participation in the data handling of the project *(epistemic dimension).*

https://doi.org/10.1017/dap.2025.16 Published online by Cambridge University Press

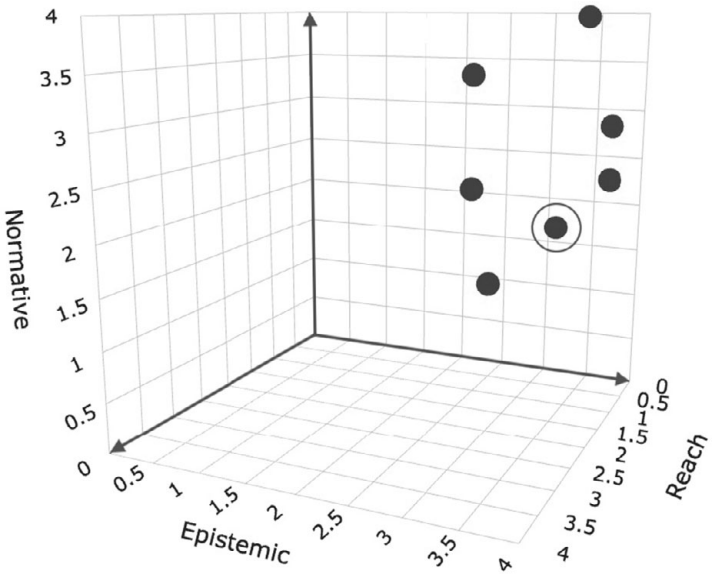

Figure 6. *PC-LG projects and their position in the participatory science cube; each dot represents one case, and each circle represents an additional case. PC-LG cases tend to have a* medium reach in the participant group*; those participants can be called lay-experts and/or an interested public. This set-up tends to be matched with* participation in a variety of data handling tasks*, including the use of the data (epistemic dimension) and a tendency for higher participation in the governance of the project (normative dimension).*

governance and data handling to a similar and high extent (Figure 6). Their participatory set-up is a core objective of the projects to realise, maintain, and shape them; most of such projects cannot exist without participation.

Typical PC examples include a data activism initiative that campaigns based on air pollution data that was collected by participants (A2); the project is governed through democratic decision-making procedures by the participants. Another example is self-organised Wikiprojects to improve Wikidata (P3). An additional case leaning towards PC is a knowledge database framed as a commons on chemical hazards (P2), which is curated and used by experts and lay-experts. The project ranks highly in terms of data handling participation but ranks less highly in participation in governance aspects, because the facilitators want to protect the project against lobbying. Another project collects data for pedestrian navigation (P1), where the transfer of data stewardship and connected governance responsibilities is designed to happen gradually from facilitators to participants.

4.3. Participatory processes in the data life cycle—labour and resource participation and participation in rule- and decision-making

As the cube does not provide insights into the processes themselves, we analysed them separately. We found participatory processes at each step of the data life cycle. Most can be found at the acquisition and sharing/publication steps. After describing the processes along the steps of the data life cycle, we provide a distinction in labour- and resource-related participation as well as rule- and decision-making focused participation.

4.3.1. Plan

Planning is an important step for rule-making in a project, given the negotiation and set-up of, for example, data governance agreements, memoranda of understanding, data management plans or

https://doi.org/10.1017/dap.2025.16 Published online by Cambridge University Press

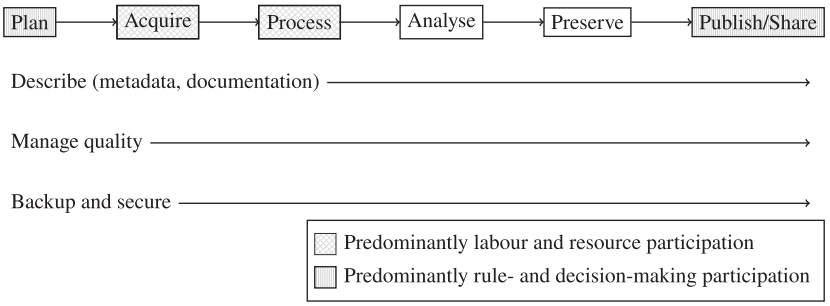

Figure 7. *The most relevant steps for labour/resource participation in the life cycle are acquire and process, and for decision- and rule-making participation, plan and publish/share.*

protocols for data access requests, or the data collection (Figure 7). The planning step is in a few cases explicitly subject to participation (C2, C8, C9, C10, and R10). In two cases, policies or binding procedures are developed. One project focused on the adjustment of the planning step during the project (R10). Together with indigenous representatives, the facilitators created a binding data governance agreement. In the other case, the facilitators developed a request protocol for health data access within a project together with patients (C9). In two health data collections, a broader range of participants is involved; in one, 52 people, including both scientists and patients, were interviewed about data sharing within a biobank. The interviews were used to inform the data-sharing policy (C10). In contrast, a data commons on cancer (C8) and a data initiative on obesity (C2) developed policy procedures involving (domain-) experts as participants.

4.3.2. Acquire

Data acquisition, a resource- and often labour-intensive step, is connected to participation in 21 of 27 projects (C1–C10, R3, R4, R6, R7, R8, R11, P1, A1–A3). Many projects collect new data, suited to their own needs. We see data gathering in five citizen science projects (R3, R6, R7, R8, and R11), two distributed research projects (R4 and R9), one mapping case (P1), and three data activism cases (A1–A3). Only the health data collections rely on a secondary use of existing data, for example, compiled patient records (C1, C2, C4, C5, C6, C9, and C10) or data from previous research studies (C3, C7, and C8). For the health data collection, the acquisition of secondary data is inherently connected to *sharing* the data.

As data collection is regularly conducted by a number of different, often non-expert collectors, management of activities and data quality are recurring topics. This regards the training/managing of data collectors to sample data in a consistent manner (P1 and R4), the support of the process with software or other means (R3, R6, R8, R9, R11, and P1), or data quality (R4, R9, P1, and A1). We also see recurring discussions of the management of especially distributed sampling sites (R4, R9, and P1) and the engagement of participants (R4, R7, R8, P1).

4.3.3. Process

The processing step includes labour-intensive aspects that typically require more technical knowledge than the data acquisition. So, some aspects of the data processing are carried out by project facilitators, for example, managing data interoperability (P1) or validation (A1). The definition of elements in the data is, however, frequently subject to participatory processes, for example, where participants, often (lay)-experts, provide their knowledge on a subject matter (P2 and P3), classify data following a protocol (P1 and R2), or when researchers engage in collaborative data curation efforts (C2 and C7).

When a project aims at fostering collaboration among participants, it is usually reflected in discussion spaces (P2 and P3), instructive pages (P3), and related support for finding collaborators (P2) and measures to manage participants' behaviour (P2).

https://doi.org/10.1017/dap.2025.16 Published online by Cambridge University Press

As the life cycle model is an idealisation, the transitions from one step to the other are not always discrete. For example, in a mapping effort, the data collection is organised in a way that processing elements of the data are included in the data collection (P1 and R3).

4.3.4. Analysis

With regard to participation, this step plays no major role. The health data collections promote independent use of the data by researchers. Most of the other projects appear to conduct the analysis in the facilitating team and/or do not specify participation happening in this step of the data life cycle. A notable exception is a research project that co-investigated research questions with an indigenous community and supported this participating group in using the data in their own right, which led to storing the data in a different format. Nevertheless, the process is connected to the planning step of the data life cycle and to the analysis (R1).

4.3.5. Preserve

The preservation of data is only involved in participation in one case, where the data stem from a research project that ran for several decades and lost its funding (R5). The data needed to be migrated to an external library.

4.3.6. Publish/share

After planning, publishing and sharing the data is the next most prominent moment for rule- and decision-making in the data life cycle. Publishing and sharing are frequent subjects for decision-making participation, mostly in health data collection and in the form of data access request committees. The projects focus in large parts on the researchers using the data (C1, C2, C6, and C8), or different stakeholder groups (C2, C5, C7, C9, and C10), which frequently include patients or patient perspectives (C5, C7, and C10). One exception, which heavily focuses on prioritising patients/data subjects, is a data-cooperative (C4) that is explicitly built around patient control over data access. Recurring topics, related to participation, are consent or the lack of consent by patients, which is addressed by participatory processes as a counterbalance (C5 and C9), and the management of data access, which is addressed in different ways: data cooperative models (C4), data request protocols/agreements (C2 and C9), or access committees (C1, C5, and C10). Further topics related to data access are the participatory drafting of policies and procedures (C2, C8, and C9), committees overseeing the projects (C1 and C2), or general agreement on publishing data (R4).

4.3.7. Labour and resource participation

Following our analysis, we can identify one hotspot for participatory processes in the acquisition of data and other data handling tasks; we call such processes *labour and resource participation.* By this, we mean the performance of processes in the data life cycle that take effort, as well as the donation of data for secondary use. While receiving existing data is less labour-intensive and includes more coordination and communication tasks than manual tasks, it is a contribution of data as a resource to the project. We find that *acquiring* and *processing* data are the most relevant steps for labour and resource participation.

4.3.8. Rule- and decision-making participation

The other process category regards participation in the data governance of the project; we call this *rule- and decision-making participation.* We consider participation as *rule-making* when it concerns the drafting of overarching prescriptions, for example, a protocol for data access, the data collection, or the set-up of a board. Participation in executive governance decisions, for example, expert participants in data access boards granting or denying data access on the basis of an application, is considered participation in *decision-making.* Rule- and decision-making participation is dominant in the *planning* and *publishing/sharing* steps. The planning step is more closely connected to rule-making participation, for example, drafting policies. The publishing/sharing aspect is connected to decision-making participation, for example, following protocols for data access/data sharing.

https://doi.org/10.1017/dap.2025.16 Published online by Cambridge University Press

5. Discussion

By combining the structural components *participatory-at-core* (PC) and *participatory-informed* (PI) with the process components *labour and resource participation* (L), and *rule- and decision-making participation* (G), we can observe three typical project types: projects with a PC structure that include labour and governance processes to a comparable degree (*PC-LG*), projects with a PI structure that focus on governance participation (*PI-G*), and projects with a PI structure that focus on labour participation (*PI-L*). The relation between labour participation and governance participation in the project types is decisive in determining which ambiguities need to be clarified.

In the following, we first discuss the ambiguities of resource/labour participation and decision-/rule-making participation, to then attribute these ambiguities to the three project types. We note how to clarify ambiguities for the different configurations and what possible countermeasures exist for each project type. We do not discuss the differentiation between participatory-informed and participatory-at-core in isolation, as the category does not indicate the nature of participation in the projects in general but merely an element of the project structure. We can see this from the vastly differing projects that share the feature of being participatory-informed. The distinction between the project structures in combination with the labour/governance component does, however, help us to provide a more precise description of the participatory configurations in the project. Therefore, this component is discussed as part of the three project types.

5.1. Ambiguity of participation in labour and resource participation

While labour/resource participation can provide insights into the inner workings of a project for participants and provide an educational dividend (cf. Kelty et al., 2015), it is also a form of participation with a higher risk of being exploitative (Resnik et al., 2015; Sloane et al., 2022). Further, projects implementing labour participation can profit from the generally positive image of participation without necessarily aligning a project with people's interests, but at the same time, they have the potential to save costs through participatory processes. The possibility to save labour costs through participation stands against the fact that facilitating participation is expensive in itself (Groves et al., 2023), among others, due to raised (transaction) costs (Steen et al., 2018). Responsible forms of participation can be expected to be even more resource-intensive, due to extra care measures.

Otherwise, labour participation can also support the realisation of projects oriented towards public and collective interests; we find, for example, a navigation system for pedestrians, encyclopaedias, or commons projects with a strong labour participation. In several cases that are oriented towards a public value, it is mentioned that the projects fill a gap left by public institutions (P1) or are intended to hold public institutions accountable (A1–A3). One case explicitly mentions that the project objective was not picked up by other actors due to missing incentives (P1). Elsewhere, projects are not conducted by public actors due to conflicts of interest, as the project objective is to observe and/or hold authorities accountable (A1–A3). Splitting laborious tasks among participants can be an enabling factor in non-commercial and resource-scarce settings. The cases in the sample had a tendency to be located in a non-profit realm. Labour and resource participation can support the realisation of data-driven projects that are less affected by a for-profit logic and can therefore realise publicly desirable projects even though they are not profitable. Therefore, the evaluation of such approaches needs to be contextualised. Even though the exploitative potential of labour participation is heightened in projects that aim at enhancing private interests, labour participation in a non-profit context is not free from an exploitative potential.

5.2. Ambiguity of participation in rule- and decision-making

Participation in rule- and decision-making is closely connected to the idea of political participation (cf. Carpentier, 2016). However, we see that it is a less dominant form of participation in our cases. This category broadly corresponds to Sloane et al.'s (2022) understanding of participation as *justice* and the normative dimension in the participatory cube (Schrögel and Kolleck, 2018), but understands

https://doi.org/10.1017/dap.2025.16 Published online by Cambridge University Press

this kind of participation in a procedural manner—not to a fixed intended end. This means that participation in rule- or decision-making does not necessarily need to be connected to *justice* but may also be employed in an, for example, managerial manner. The differentiation between labour and resource participation has overlaps with Kelty et al. (2015) dimension *goals and tasks*, in which the former overlaps more with rule- and decision-making participation and the latter with labour and resource participation. Our categories provide additional context to the nature of the tasks and mode of goal setting.

It is important to clarify that participation in those aspects *can* increase accountability, but should not be misunderstood by facilitators as a reduction of obligations on their side. This kind of process inherits a heightened risk and potential to obscure and veil responsibilities (Steen et al., 2018; Zehner and Ullrich, 2024). This suggests that an explicit and clear attribution of roles and connected responsibilities is necessary. The selection of participants additionally impacts governance participation, especially if processes aim at representation. Such processes carry the risk of enhancing existing power imbalances. Further, possible selection biases of facilitators, for example, to confirm their own beliefs or issues with narrow out-reach channels, should be counterbalanced with context-sensitive selection strategies.

5.3. How to face participatory ambiguities in participatory data-driven projects

In the following, we discuss how to address ambiguities in the project types that combine the project configurations (PI/PC) with the categories of participatory processes (L/G).

5.3.1. PI-L—projects focusing on labour participation

PI-L projects likely focus on participatory processes in the data life cycle steps, *acquisition*, *data processing*, and/or *data analysis.* They tend to be *open to a broad group of participants*, depending on the affordances that follow from what is necessary to participate. Such a participatory approach can help to realise projects that enhance collective interests but are difficult to realise in a profit-oriented environment. In this kind of project, costs for laborious and resource-intensive tasks may be reduced through participatory processes. PI-L projects have, at the same time, a heightened risk for exploitation, especially if the project enhances private interests.

Labour/resource participation should be considered as a resource contribution to the project in the same way as financial or managerial contributions by stakeholders. Facilitators should clarify for themselves and for participants/the public who has contributed what to the implementation and maintenance of the project. If labour and resource participation take place in isolation, this decision should be explained and justified.

In addition, facilitators should explicitly communicate how participants are considered in the benefits of the project, as participants directly contribute to the output of the project. Benefits for participants can be differentiated into *direct* and *indirect* benefits. Non-monetary direct benefits may be the experience and insights in otherwise hidden processes, such as the outcome of a project, for example, a service that participants can use (P1, P3, and P2). Indirect benefits can be collective/public benefits such as accessible research results as public knowledge (e.g., R3, R7, and R9) or the impact of data activism cases, holding public institutions accountable (A1–A3). The clear communication of these aspects can enable participants to make informed decisions for or against contributing to a project.

Therefore, our baseline recommendation for PI-L projects is transparent expectation management: facilitators should communicate towards participants what their realm of impact is and how they contribute to the project. Additionally, facilitators should justify an isolated implementation of labour participation and clarify the project benefits. To raise the legitimacy of such an approach, benefits should exist for participants and the project and/or should be collective benefits. Those benefits should be exemplified and named explicitly. In addition, participation in rule- and decision-making aspects of the project should be considered; participants should at least have insight into the decision-making of the project.

https://doi.org/10.1017/dap.2025.16 Published online by Cambridge University Press

5.3.2. PI-G—projects focusing on participation in governance

PI-G projects likely focus on participation in the data life cycle steps of *planning* and/or data *sharing/publishing*; the participatory processes are likely open to *(lay-) experts.* The expertise may regard the application domain, technological aspects, or the needs of an interest group. Such a participatory approach can align aspects of a project with the interests represented by the participants. In this way, collective interests can be considered in PI-G projects that are otherwise not participatory or community-oriented. At the same time, such projects run the risk to veil the accountabilities of facilitators and potentially reinforce existing power imbalances in the representation of participants.

Facilitators need to draft governance participation with contextual care. Among others, this can mean cautiously limiting participation and marking what is not up for discussion/cannot be altered through participation. In one case, on chemical hazards of building materials (P2), we saw an explicit restriction on participation with the intention to prevent lobbying. In another case, governance responsibility was gradually handed over to participants to ensure a working project (P1). This suggests that simply increasing rule- and decision-making participation is not necessarily better for collective interests and that a contextualised design of governance participation can help to improve the project's functioning. Additionally, recognition by facilitators of their accountability for the project and its impact is crucial, along with the explicit allocation of other roles and responsibilities within a project.

To counterbalance possible selection biases of facilitators and include additional positions, it can help to install mechanisms for contestation rather than relying on deliberative formats (Crawford and Lumby, 2013; Cohen and Suzor, 2024). Further, informal channels can be helpful for participants and the public to voice their opinions (Kelty et al., 2015), for example, through simple feedback options, contact information, and an open project culture.

In PI-G projects, responsibilities need to be assigned clearly and should not be "outsourced" to participants. It may be helpful to explicitly state the objectives of the participatory processes and to clarify what can be decided by whom for what reason. It should be clarified how participants are determined by stating the selection procedure and the reasoning behind that. A participant selection by facilitators can contribute to avoiding important, but possibly confrontational, aspects that facilitators may not know about or want to spare. This can be counterbalanced by including mechanisms for contestation.

5.3.3. PC-LG—projects combining governance and labour participation

PC-LG projects focus on participation in potentially *all steps of the data life cycle* and combine labour and resource participation with participation in rule- and decision-making; the participatory processes are likely open to *(lay-) experts* and an *interested public.* This project type often builds on community organising and will thus need heightened attention and resources for community facilitation. PC-LG configurations are found in projects that are difficult to realise in a profit-driven environment and are likely to be non-profit projects. They can realise, for example, data activism efforts or public interest technologies. Those projects are likely set up to be aligned with the needs and wishes of participants and are realised through their labour and resource input. For PC-LG projects, it is a question of whether facilitators manage to build and foster a reliable community or can rely on an existing community.

In this project configuration, the dimension "Collective, Affective, and Communicative Experience of Participation" (Kelty et al., 2015, 483) is naturally more pronounced, given the high intensity and variation of participation. Such an experiential focus may be explicitly desirable, for example, because it developed organically through a community or is considered important for the specific project domain. Yet, for PC-LG projects, there is a risk of overriding other important aspects through an experiential focus; this may affect other parts of the problem orientation, impact, and success of the project. This again can lower the positive experience of being a part of something for participants. Further, participation may not be a suitable solution for every aspect of the data life cycle.

The weight of the experiential dimension and its function should be recognised by facilitators. The experiential dimension should be considered in the problem orientation of the project. This could mean

https://doi.org/10.1017/dap.2025.16 Published online by Cambridge University Press

that the connection of the participants among one another and the experience of being part of something belong to the intended outcome of the project and the problem it tackles. Further, impact measures are important in this project type to avoid making participation a means in itself.

The three project types are obviously simplifications describing much more messy realities, but we hope they can contribute to discussing and designing more problem-oriented and transparent approaches to participation in data-driven projects.

5.4. Future work

To close, we highlight some calls to action for the advancement of more problem-oriented and transparent participatory projects, future research, and a short reflection on the use of the model.

Generally, we see a necessity to develop a standard for what elements to consider when aiming at providing problem-oriented and transparent participatory processes. Providing a catalogue of questions and/or checklists can act as a guiding tool for facilitators. Further, such resources can function as a reference to support arguments for better practices within an organisation and as a reference for funders to require respective practices. These questions should look at the *problem orientation* of the project, *communication* and *transparency measures*, the *scope of participation*, the *role attribution*, and direct/indirect *benefits* for participants and the public.

While our research focused on facilitators in the sense of the people who are responsible for the projects, it needs to be recognised that their decisions are not made in a vacuum. Their decisions may be constrained by the availability of resources and (conflicting) interests in an organisation, such as given competencies. Those elements need to be investigated further to provide context-aware recommendations on how to improve participatory practices and to decide where other approaches are more promising.

If the model is to be used further, we note one caveat. We used the terminology provided by the authors of the Participatory Science Cube to define the stages on the scales. To use the model on more general data-driven projects, it may be advisable to use different terms for some stages. This regards the first stage on the epistemic dimension "crowdsourcing" as the lowest level contains less agency than the term in its usual use implies. Further, including a stage "affected public" on the reach dimension is advisable as this group plays an important role in more general data-driven projects; it should be located between "interested public" and "broad public." We also recommend using the model with data collected for the use of the model, given its specificity.

The analysis based on the cube does not consider the important aspect of who the facilitators are and what kind of organisation they belong to, such as how participants are selected or if they enter the process through self-selection. Looking at those aspects can be a valuable addition.

6. Conclusion

We analyse 27 cases of participatory data-driven projects regarding their openness to participation and participatory processes along the data life cycle. Regarding their openness, we found two project configurations: projects that are *participatory-at-core* (PC), which cannot be thought of without participation and implement participation in a variety of steps in the data life cycle, and projects that are *participatory-informed* (PI), which employ participation sporadically in the data life cycle. We find that many projects employ participation regarding the acquisition, publishing, and sharing of the data, while the planning step is less often subject to participation. Those processes can be divided into participatory processes focused on *labour and resources* (L) as well as processes regarding the governance of the project concerning *rule- and decision-making participation* (G). We compare these to formulate three typical project types—PI-L, PI-G, and PC-LG—and discuss their ambiguities and potential mitigations to provide greater clarification. PI-L projects can enable the realisation of projects that are underfunded but have a heightened risk of one-sided exploitation. Isolated labour and resource participation should prioritise collective benefits in their outcomes, be justified explicitly towards participants, and involve expectation management. Granting insights into the governance of

https://doi.org/10.1017/dap.2025.16 Published online by Cambridge University Press

the projects are possible countermeasures. PI-G projects can support the alignment of a project with people's needs and wishes, but they also have a heightened risk of veiling accountabilities and increasing existing power asymmetries among participants. Countermeasures can be a public justification of the participant selection, contestation mechanisms, and an explicit attribution of accountabilities. PC-LG can be a collective realisation of projects according to people's needs and wishes, which would be hard to follow in a profit-oriented environment. At the same time, those projects may have the potential to employ participation predominantly for experiential reasons. The experiential dimension should be taken into account in the problem orientation of projects to avoid using participatory processes for their own sake.

The countermeasures are part of the governance of the project, particularly the policies and procedures. Further research is needed on what those countermeasures can look like in more detail, as well as how to formalise them in the project policies. Additionally, the interplay between further participatory dimensions and contextual expressions of the project types should be investigated.

7. Limitations

As touched upon prior, our use of the Participatory Science Cube for a larger number of projects in combination with using secondary sources—instead of in-depth and first-hand case studies—causes a *trade-off between precision/detail and the ability to compare a larger number of cases.* The use of secondary sources meant that the *descriptions of the cases were not tailored to our analysis.* Therefore, it is possible that we needed to evaluate, for example, normative participation with a value of 0, as no such participation was mentioned in the description, which does not have to mean that there actually is no normative participation. The use of secondary data may influence the reproducibility of our evaluations. To counterbalance this, we additionally published the justification of our analysis in the Supplementary Material. This is next to the inability to evaluate the success of the described approaches, which is a limitation. Additionally, it is likely that not all participatory processes present in the cases were described, as this was not the main focus of the analysed papers.

Lastly, our findings may not be *representative*, given the limitations of our methodology. The evaluation of the cases can be tracked in detail in the Supplementary Material.

Supplementary material. The Supplementary Material on the cases and placement in the Participatory Science Cube can be accessed at https://doi.org/10.6084/m9.figshare.28226423.

Acknowledgements. We thank the anonymous reviewers for their extensive feedback. We are also grateful for valuable feedback from Hadi Asghari, Freya Hewett, Freia Kuper, and Theresa Züger at different stages of the work.

Author contribution. Conceptualisation: J.F. and T.H.; Data curation: J.F. and I.K.; Data visualisation: J.F. and T.H.; Investigation: J.F.; Methodology: J.F. and T.H.; Supervision: T.H.; Writing—original draft: J.F.; Writing—review and editing: J.F. and T.H. All authors approved the final submitted draft.

Funding statement. This research is part of J.F.'s doctoral research, which is funded by a grant from the German Ministry of Education and Research (01IS20058) at the Alexander von Humboldt Institute for Internet and Society and a School of Computer Science Handsel Scholarship from the University of St Andrews.

Competing interests. The authors declare none.

Ethical standard. The research only uses secondary data and meets all ethical guidelines of the University of St Andrews Teaching and Research Ethics Committee.

References

Ada Lovelace Institute (2021) Participatory Data Stewardship. Available at https://www.adalovelaceinstitute.org/wp-content/uploads/2021/11/ADA_Participatory-Data-Stewardship.pdf (accessed 23 May 2025).

Arnstein SR (1969) A ladder of citizen participation. *Journal of the American Institute of Planners 35*(4), 216–224. https://doi.org/10.1080/01944366908977225.

Association for Computing Machinery (2024) ACM Digital Library. Available at https://dl.acm.org/ (accessed 23 May 2025).

Aubin I, **Cardou F**, **Boisvert-Marsh L**, **Garnier E**, **Strukelj M and Munson A** (2020) Managing data locally to answer questions globally: The role of collaborative science in ecology. *Journal of Vegetation Science 31*(3), 509–517. https://doi.org/10.1111/jvs.12864.

Baker KS and Karasti H (2018) Data care and its politics: Designing for local collective data management as a neglected thing. In *Proceedings of the 15th Participatory Design Conference: Full Papers—Volume 1*. New York: Association for Computing Machinery, pp. 1–12. https://doi.org/10.1145/3210586.3210587.

Bauermeister S, **Orton C**, **Thompson S**, **Barker RA**, **Bauermeister JR**, **Ben-Shlomo Y**, **Brayne C**, **Burn D**, **Campbell A**, **Calvin C**, **Chandran S**, **Chaturvedi N**, **Chêne G**, **Chessell IP**, **Corbett A**, **Davis DHJ**, **Denis M**, **Dufouil C**, **Elliott P**, **Fox N**, **Hill D**, **Hofer SM**, **Hu MT**, **Jindra C**, **Kee F**, **Kim CH**, **Kim C**, **Kivimaki M**, **Koychev I**, **Lawson RA**, **Linden GJ**, **Lyons RA**, **Mackay C**, **Matthews PM**, **McGuiness B**, **Middleton L**, **Moody C**, **Moore K**, **Na DL**, **O'Brien JT**, **Ourselin S**, **Paranjothy S**, **Park KS**, **Porteous DJ**, **Richards M**, **Ritchie CW**, **Rohrer JD**, **Rossor MN**, **Rowe JB**, **Scahill R**, **Schnier C**, **Schott JM**, **Seo SW**, **South M**, **Steptoe M**, **Tabrizi SJ**, **Tales A**, **Tillin T**, **Timpson NJ**, **Toga AW**, **Visser PJ**, **Wade-Martins R**, **Wilkinson T**, **Williams J**, **Wong A and Gallacher JEJ** (2020) The dementias platform UK (DPUK) dataportal. *European Journal of Epidemiology 35*(6), 601–611. https://doi.org/10.1007/s10654-020-00633-4.

Birhane A, **Isaac W**, **Prabhakaran V**, **Diaz M**, **Elish MC**, **Gabriel I and Mohamed S** (2022) Power to the people? Opportunities and challenges for participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '22)*. New York: Association for Computing Machinery, pp. 1–8. https://doi.org/10.1145/3551624.3555290.

Bolten N and Caspi A (2022) Towards operationalizing the communal production and management of public (open) data: A pedestrian network case study: A pedestrian network case study in operationalizing communal open data. In *ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS)*. New York: Association for Computing Machinery, pp. 232–247. https://doi.org/10.1145/3530190.3534821.

Bondi E, **Xu L**, **Acosta-Navas D and Killian JA** (2021) Envisioning communities: A participatory approach towards AI for social good. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. New York: Association for Computing Machinery, pp. 425–436. https://doi.org/10.1145/3461702.3462612.

Carpentier N (2016) Beyond the ladder of participation: An analytical toolkit for the critical analysis of participatory media processes. *Javnost—The Public 23*(1), 70–88. https://doi.org/10.1080/13183222.2016.1149760.

Cohen C, **Cheney L**, **Duong K**, **Lea B and Unno Z** (2015) Identifying opportunities in citizen science for academic libraries. *Issues in Science and Technology Librarianship 79*, 1–13. https://doi.org/10.5062/F4BR8Q66.

Cohen T and Suzor NP (2024) Contesting the public interest in AI governance. *Internet Policy Review 13*(3). https://doi.org/10.14763/2024.3.1794.

Corporation for Digital Scholarship (2024) Zotero. Available at https://www.zotero.org/ (accessed 23 May 2025).

Crawford K and Lumby C (2013) Networks of governance: Users, platforms, and the challenges of networked media regulation. *International Journal of Technology Policy and Law 1*(3), 270. https://doi.org/10.1504/IJTPL.2013.057008.

Currie M, **Paris B and Donovan J** (2019) What difference do data make? Data management and social change. *Online Information Review 43*(6), 971–985. https://doi.org/10.1108/OIR-02-2018-0052.

Cushnan D, **Berka R**, **Bertolli O**, **Williams P**, **Schofield D**, **Joshi I**, **Favaro A**, **Halling-Brown M**, **Imreh G**, **Jefferson E**, **Sebire NJ**, **Reilly G**, **Rodrigues JCL**, **Robinson G**, **Copley S**, **Malik R**, **Bloomfield C**, **Gleeson F**, **Crotty M**, **Denton E**, **Dickson J**, **Leeming G**, **Hardwick HE**, **Baillie K**, **Openshaw PJ**, **Semple MG**, **Rubin C**, **Howlett A**, **Rockall AG**, **Bhayat A**, **Fascia D**, **Sudlow C**, **NCCID Collaborative and Jacob J** (2021) Towards nationally curated data archives for clinical radiology image analysis at scale: Learnings from national data collection inresponse to a pandemic. *Digital Health* 7. https://doi.org/10.1177/20552076211048654.

Delgado F, **Barocas S and Levy K** (2022) An uncommon task: Participatory design in legal AI. *Proceedings of the ACM on Human–Computer Interaction 6*(CSCW1), 1–23. https://doi.org/10.1145/3512898.

Elsevier (2024) Scopus. Available at https://www.scopus.com/ (accessed 23 May 2025).

Falco G (2019) Participatory AI: reducing AI bias and developing socially responsible AI in smart cities. In *2019 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*. New York: IEEE, pp. 154–158. https://doi.org/10.1109/CSE/EUC.2019.00038.

Fassbender J, **Kuehnlein I and Henderson T** (2025) Supplementary material—facing the ambiguities of participation in data-driven projects: A systematic literature review [Dataset]. https://doi.org/10.6084/m9.figshare.28226423.v1.

Faundeen J, **Burley TE**, **Carlino JA**, **Govoni DL**, **Holl SL**, **Hutchison VB**, **Martín E**, **Montgomery ET**, **Ladino C**, **Tessler S and Zolly LS** (2014) *The United States Geological Survey Science Data Lifecycle Model* (Tech. Rep. No. 2013-1265) (ISSN: 2331-1258 Publication Title: Open-File Report). U.S. Geological Survey. https://doi.org/10.3133/ofr20131265.

Fung A (2006) Varieties of participation in complex governance. *Public Administration Review 66*(s1), 66–75. https://doi.org/10.1111/j.1540-6210.2006.00667.x.

Gallier S, **Price G**, **Pandya H**, **McCarmack G**, **James C**, **Ruane B**, **Forty L**, **Crosby BL**, **Atkin C**, **Evans R**, **Dunn KW**, **Marston E**, **Crawford C**, **Levermore M**, **Modhwadia S**, **Attwood J**, **Perks S**, **Doal R**, **Gkoutos G**, **Dormer R**, **Rosser A**, **Fanning H and Sapey E** (2021) Infrastructure and operating processes of PIONEER, the HDR-UK data hub in acute care and the workings of the data trust committee: A protocol paper. *BMJ Health and Care Informatics 28*(1), e100294. https://doi.org/10.1136/bmjhci-2020-100294.

https://doi.org/10.1017/dap.2025.16 Published online by Cambridge University Press

Gerdes A (2021) A participatory data-centric approach to AI ethics by design. *Applied Artificial Intelligence 36*, 1–19. https://doi.org/10.1080/08839514.2021.2009222.

Gilman ME (2022) Beyond window dressing: Public participation for marginalized communities in the datafied society. *Fordham Law Review 91*, 503.

Groves L, **Peppin A**, **Strait A and Brennan J** (2023) Going public: The role of public participation approaches in commercial AI labs. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. New York: Association for Computing Machinery, pp. 1162–1173. https://doi.org/10.1145/3593013.3594071.

Himmelreich J (2023) Against "democratizing AI.". *AI & Society 38*(4), 1333–1346. https://doi.org/10.1007/s00146-021-01357-z.

Hogan W, **Shenkman E**, **Robinson T**, **Carasquillo O**, **Robinson P**, **Essner R**, **Bian J**, **Lipori G**, **Harle C**, **Magoc T**, **Manini L**, **Mendoza T**, **White S**, **Loiacono A**, **Hall J and Nelson D** (2022) The OneFlorida data trust: A centralized, translational research data infrastructure of statewide scope. *Journal of the American Medical Informatics Association 29*(4), 686–693. https://doi.org/10.1093/jamia/ocab221.

Institute of Electrical and Electronics Engineers (2024) IEEE Xplore. Available at https://ieeexplore.ieee.org/Xplore/home.jsp (accessed 23 May 2025).

Kanke T (2021) Knowledge curation work in Wikidata WikiProject discussions. *Library Hi Tech 39*(1), 64–79. https://doi.org/10.1108/LHT-04-2019-0087.

Kelty C, **Panofsky A**, **Currie M**, **Crooks R**, **Erickson S**, **Garcia P**, **Wartenbe M and Wood S** (2015) Seven dimensions of contemporary participation disentangled. *Journal of the Association for Information Science and Technology 66*(3), 474–488. https://doi.org/10.1002/asi.23202.

Kitchenham B and Charters S (2007) Guidelines for performing systematic literature reviews in software engineering. Available at https://legacyfileshare.elsevier.com/promis_misc/525444systematicreviewsguide.pdf (accessed 23 May 2025).

Kokai A, **Blake A**, **Dedeo M and Lent T** (2020) Building shared information infrastructure for chemical alternatives assessment. *Elementa 8*(23), 26. https://doi.org/10.1525/elementa.422.

Kraus E, **Scott K**, **Zucker R**, **Heisey-Grove D**, **King R**, **Carton T**, **Daley M**, **Deakyne Davies S**, **Block J**, **Haemer M**, **Goodman A**, **Garrett N and Davidson A** (2022) A governance framework to integrate longitudinal clinical and community data in a distributed data network: The childhood obesity data initiative. *Journal of Public Health Management and Practice 28*(2), E421–E429. https://doi.org/10.1097/PHH.0000000000001408.

Love R, **Hardy B-J**, **Heffernan C**, **Heyd A**, **Cardinal-Grant M**, **Sparling L**, **Healy B**, **Smylie J and Long R** (2022) Developing data governance agreements with indigenous communities in Canada: Toward equitable tuberculosis programming, research, and reconciliation. *Health and Human Rights 24*(1), 21–33. Available at https://www.scopus.com/inward/record.uri?eid=2-s2.0-85132859260&partnerID=40&md5=d68c779c2666a00182cf9e1e962781dc.

Marres N (2015) *Material Participation*. London: Palgrave Macmillan. https://doi.org/10.1007/978-1-137-48074-3.

Mozilla (n.d.) Mozilla Common Voice. Available at https://commonvoice.mozilla.org/ (accessed 29 April 2023).

Murray B, **Kerfoot E**, **Chen L**, **Deng J**, **Graham M**, **Sudre C**, **Molteni E**, **Canas L**, **Antonelli M**, **Klaser K**, **Visconti A**, **Hammers A**, **Chan A**, **Franks P**, **Davies R**, **Wolf J**, **Spector T**, **Steves C**, **Modat M and Ourselin S** (2021) Accessible data curation and analytics for international-scale citizen science datasets. *Scientific Data 8*(1), 297. https://doi.org/10.1038/s41597-021-01071-x.

Plana A, **Furner B**, **Palese M**, **Dussault N**, **Birz S**, **Graglia L**, **Kush M**, **Nicholson J**, **Hecker-Nolting S**, **Gaspar N**, **Rasche M**, **Bisogno G**, **Reinhardt D**, **Zwaan C**, **Koscielniak E**, **Frazier A**, **Janeway K**, **S Hawkins D**, **Kolb E and Volchenboum S** (2021) Pediatric cancer data commons: Federating and democratizing data for childhood cancer research. *JCO Clinical Cancer Informatics 5*, 1034–1043. https://doi.org/10.1200/CCI.21.00075.

Pyper E, **Henry D**, **Yates E**, **Mecredy G**, **Ratnasingham S**, **Slegers B and Walker J** (2018) Walking the path together: Indigenous health data at ICES. *Healthcare Quarterly 20*(4), 6–9. https://doi.org/10.12927/hcq.2018.25431.

Queerinai OO, **Ovalle A**, **Subramonian A**, **Singh A**, **Voelcker C**, **Sutherland DJ**, **Locatelli D**, **Breznik E**, **Klubicka F**, **Yuan H**, **Hetvi J**, **Zhang H**, **Shriram J**, **Lehman K**, **Soldaini L**, **Sap M**, **Deisenroth MP**, **Pacheco ML**, **Ryskina M**, **Mundt M**, **Agarwal M**, **Mclean N**, **Xu P**, **Pranav A**, **Korpan R**, **Ray R**, **Mathew S**, **Arora S**, **John S**, **Anand T**, **Agrawal V**, **Agnew W**, **Long Y**, **Wang ZJ**, **Talat Z**, **Ghosh A**, **Dennler N**, **Noseworthy M**, **Jha S**, **Baylor E**, **Joshi A**, **Bilenko NY**, **Mcnamara A**, **Gontijo-Lopes R**, **Markham A**, **Dong E**, **Kay J**, **Saraswat M**, **Vytla N and Stark L** (2023) Queer in AI: A case study in community-led participatory AI. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. New York: Association for Computing Machinery, pp. 1882–1895. https://doi.org/10.1145/3593013.3594134.

Reichert B, **Bayless M**, **Cheng T**, **Coleman J**, **Francis C**, **Frick W**, **Gotthold B**, **Irvine K**, **Lausen C**, **Li H**, **Loeb S**, **Reichard J**, **Rodhouse T**, **Segers J**, **Siemers J**, **Thogmartin W and Weller T** (2021) NABat: A top-down, bottom-up solution to collaborative continental-scale monitoring. *Ambio 50*(4), 901–913. https://doi.org/10.1007/s13280-020-01411-y.

Resnik DB, **Elliott KC and Miller AK** (2015) A framework for addressing ethical issues in citizen science. *Environmental Science & Policy 54*, 475–481. https://doi.org/10.1016/j.envsci.2015.05.008.

Riley J and Mason-Wilkes W (2024) Dark citizen science. *Public Understanding of Science 33*(2), 142–157. https://doi.org/10.1177/09636625231203470.

Robinson C, **Kong T**, **Coates R**, **Watson I**, **Stokes C**, **Pert P**, **McConnell A and Chen C** (2021) Caring for Indigenous data to evaluate the benefits of Indigenous environmental programs. *Environmental Management 68*(2), 160–169. https://doi.org/10.1007/s00267-021-01485-8.

https://doi.org/10.1017/dap.2025.16 Published online by Cambridge University Press

Schrögel P and Kolleck A (2018) The many faces of participation in science. *Science & Technology Studies 32*, 77–99. https://doi.org/10.23987/sts.59519.

Sloane M (2024) Controversies, contradiction, and "participation" in AI. *Big Data & Society 11*(1), 20539517241235862. https://doi.org/10.1177/20539517241235862.

Sloane M, **Moss E**, **Awomolo O and Forlano L** (2022) Participation is not a design fix for machine learning. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '22)*. New York: Association for Computing Machinery, pp. 1–6. https://doi.org/10.1145/3551624.3555285.

Steen T, **Brandsen T and Verschuere B** (2018) The dark side of co-creation and co-production: Seven evils. In *Co-Production and Co-Creation: Engaging Citizens in Public Services*, 1st Edn. New York: Routledge, pp. 284–293. https://doi.org/10.4324/9781315204956.

Tai C, **Harris-Wai J**, **Schaefer C**, **Liljestrand P and Somkin C** (2019) Multiple stakeholder views on data sharing in a biobank in an integrated healthcare delivery system: Implications for biobank governance. *Public Health Genomics 21*(5–6), 207–216. https://doi.org/10.1159/000500442.

Torres-Espín A, **Almeida C**, **Chou A**, **Huie J**, **Chiu M**, **Vavrek R**, **Sacramento J**, **Orr M**, **Gensel J**, **Grethe J**, **Martone M**, **Fouad K**, **Ferguson A**, **Alilain W**, **Bacon M**, **Batty N**, **Beattie M**, **Bresnahan J**, **Burnside E** and **the STREETFAIR Workshop Participants** (2021) Promoting FAIR data through community-driven agile design: The open data commons for spinal cord injury (odc-sci.org). *Neuroinformatics 20*, 203–219. https://doi.org/10.1007/s12021-021-09533-8.

Vayena E and Blasimme A (2017) Biomedical big data: New models of control over access, use and governance. *Journal of Bioethical Inquiry 14*(4), 501–513. https://doi.org/10.1007/s11673-017-9809-6.

VERBI Software (2024) MAXQDA. Available at https://www.maxqda.com/ (accessed 23 May 2025).

Walji M, **Spallek H**, **Kookal K**, **Barrow J**, **Magnuson B**, **Tiwari T**, **Oyoyo U**, **Brandt M**, **Howe B**, **Anderson G**, **White J and Kalenderian E** (2022) BigMouth: Development and maintenance of a successful dental data repository. *Journal of the American Medical Informatics Association 29*(4), 701–706. https://doi.org/10.1093/jamia/ocac001.

Walker D, **Nost E**, **Lemelin A**, **Lave R and Dillon L** (2018) Practicing environmental data justice: From data rescue to data together. *Geo: Geography and Environment 5*(2), e00061. https://doi.org/10.1002/GEO2.61.

Zehner N and Ullrich A (2024) Dreaming of AI: Environmental sustainability and the promise of participation. *AI & Society 40*, 2605–2617. https://doi.org/10.1007/s00146-024-02011-0.

Cite this article: Fassbender J, Kuehnlein I and Henderson T (2025). Facing the ambiguities of participation in data-driven projects: a systematic literature review. *Data & Policy*, 7: e41. doi:10.1017/dap.2025.16