Accountable Human Subject Research Data Processing using Lohpi

Aakash Sharma[1,*], *Thomas Bye* Nilsen[1], *Lars* Brenna[1], *Dag* Johansen[1], and *Håvard D.* Johansen[1]

[1]Department of Computer Science, UiT The Arctic University of Norway, Tromsø, Norway

Abstract. In human subject research, various data about the studied individuals are collected. Through re-identification and statistical inferences, this data can be exploited for interests other than the ones the subjects initially consented to. Such exploitation must be avoided to maintain trust with the researched population. We argue that keeping data-access policies up-to-date and building accountability on research data processing can reflect subjects' *consent* and mitigate data misuse. With accountability in mind, we are building *Lohpi*: a decentralized system for research data sharing with up-to-date access policies. We demonstrate our initial prototype with timely delivery of policy changes along with minimal access control overhead.

1 Introduction

ICT-centered research methodologies are being quickly and widely adopted in the fields of social sciences and humanities [1], fueled by advances in big-data systems, knowledge extraction, and machine-learning methods. Some researchers have raised concerns about this rapid adoption of new and unfamiliar technologies as they bring new challenges in research [2], in particular with regards to privacy and compliance with laws and regulations. If ICT-centered research methodologies are not implemented correctly, the researcher may not obtain the required ethics approval and fail to establish the trust needed to recruit volunteer participants that, for instance, epidemiology, sports science, psychology, social sciences, and humanities heavily rely on.

Recent regulations such as General Data Protection Regulation (GDPR) require explicit *informed consent* from participating individuals (hereinafter referred as *subjects*) for collecting and processing their data [3]. Researchers and Institutions must ensure that sensitive data of subjects is meticulously handled [4, 5]. World Health Organization [6] states that an ethics committee must protect subjects from any anticipated harm.

Perhaps the most common techniques to process data in compliance with these laws are anonymization and aggregation and are often recommended by ethics committees. However, weaknesses in known methods have led to multiple privacy violations [7, 8]. Advancements in statistical inferences and re-identification attack methodologies have made it relatively easy to identify discussed individuals in a study [9, 10]. Differential privacy [11] is often hailed as one of the advanced solutions to protect an individual's privacy in a dataset. However,

*e-mail: aakash.sharma@uit.no

it is difficult to use differential privacy in every possible scenario [7]. Yang et al. [12] and Garfinkel et al. [13] have highlighted issues with differential privacy. Existing differential privacy protocols assume a relatively simple data model with a centralized database. Misunderstandings about randomness and noise, limited access to micro-data, and accuracy are some of the raised concerns [13, 14].

Kroll et al. [7] argue the need for having global visibility in data usage to test the next generation of privacy-enhancing technologies. Researchers argue that building accountability around the applicable laws and the dynamic privacy risks landscape, is the way forward [7, 8]. Subjects' perception of privacy might change over time and depend upon the purpose data is collected for [15, 16]. Although data analysis techniques, such as *statistical inferences*, can blur the lines between sensitive and non-sensitive [7] data, the problems of informed consent, individual privacy, harm, and data re-identification are evident in big-data computing [9]. Inspired by Shneiderman [4], we argue for auditing, independent oversight, and trustworthy certification for research data sharing and processing.

In this paper, we present Lohpi: a system for safe and accountable research data sharing, enabled by a secure network substrate for distributing and applying up-to-date access policies. Lohpi takes a decentralized approach where research institutions can process data on their internal computing infrastructure and maintain control of valuable data assets. The key contribution of Lohpi is our compliant data analytic framework that encapsulates and manages distributed data assets. Data access policies reside as *meta-code* stored at file-system level [17], along with the data they govern and updated using *gossip*-based communication. We present our initial results and discuss future work.

2 Background

Data-driven research in social sciences and humanities relies heavily on the voluntary participation of subjects [19]. Metrics from the Dataverse project [20] show that more than 29 300 (21%) datasets are related to social sciences and 7040 (5%) of the datasets are from medicine, health, and life sciences. The subjects of these studies contributed different types of data. Personally identifiable information (PII), such as contact information, can potentially identify an individual and is typically anonymized to safeguard a subject's privacy. The collected data remain publicly available on repositories such as Dataverse [21]. However, multiple data sources can be linked without a subject's knowledge or *consent*, which may

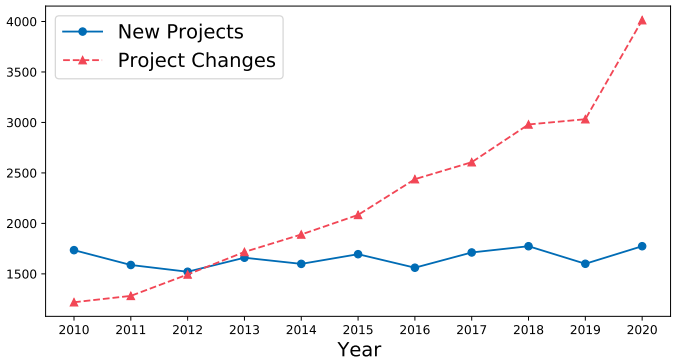

Figure 1. Different types of applications processed annually at the Norwegian ethics committee (REK) [18].

result in re-identification of the subject [22]. Protecting the data shared on a global scale is identified as one of the key challenges in the era of Big Data [22].

Typically, research projects concerning humans require regulatory approval from an ethics committee, a data protection officer (DPO), or an institutional review board (IRB). We collected data from annual reports of the Norwegian ethics committee (REK) [18] and identified two key metrics: *new projects* and *project changes*. Figure 1 shows the growing number of changes to existing projects. We contacted REK to understand what is considered a project change. A project change includes changes to people who have access to the collected data (*new researchers*), newly discovered risks for subjects (*new threats*), and even changes in conditions for dispensation from professional secrecy requirements (*new laws*). These changes require approval by a governing body.

Data collected for a specific research context are often used beyond its initially specified goals [23]. Also, data without any access control can be exploited by third parties. A dataset downloaded from a public repository might not reflect the current state of data sharing policies approved by an ethics committee. To the best of our knowledge, we are not aware of mechanisms that an ethics committee can use to verify compliant data processing by researchers. Anonymized data available in repositories such as Dataverse [21] can potentially also be re-identified and misused [5]. Consequently, we have built Lohpi as a platform for compliant data usage among researchers, which may identify a *rogue researcher* [24].

The FAIR Guiding Principles [25] are becoming an established standard for managing research data. The principles can be applied to data assets to make them Findable, Accessible, Interoperable, and Reusable. Holub et al. [26] proposed an extension to the FAIR data management principles by accounting for the privacy risks associated with research data. They map the flow of data from participants to research data repositories and highlight the trust and privacy aspects. A research project is considered compliant if the consent is obtained either from the participants or from an ethics committee [26]. Holub et al. also highlight the following competing interests in human data use: (a) protection of privacy of individuals, (b) reuse of data, and (c) complex ownership and economic interests; and conclude that anonymization cannot always protect individuals' privacy when data are shared. Instead, they advocate for checking compliance to research data before they are shared. By checking data usage against approved policies at any stage of a project, Lohpi extends this notion of compliance to the entire lifetime of the data. Note that Lohpi is designed to not limit collaboration among researchers. On the contrary, by building on accountability and oversight of research data processing, we conjecture that trust between researchers and the public can be improved [4, 7]. That may lead to improved participation in fields such as social sciences and humanities, which rely heavily on public participation.

2.1 Vision with Lohpi

In this section, we present our vision for Lohpi, and we refer to Figure 2 that describes a typical research data collection process with this system. A principal investigator (PI) or a team of co-PIs formulates a project protocol outlining research data collection and processing. The protocol provides details on the data that will be collected and how it will be processed and stored. The protocol also provides details on collaborators and measures to protect subjects' privacy while processing and sharing data. The project protocol is sent to an ethics committee (or some other regulatory unit). The ethics committee reviews the protocol and ensures that the data collection, processing, and sharing within the scope of the research project complies with applicable laws and regulations. The committee also reviews potential threats to the subjects' privacy and necessary measures put in place to safeguard privacy. The approval ensures that these measures are transformed into a verifiable data-access policy. The approval also

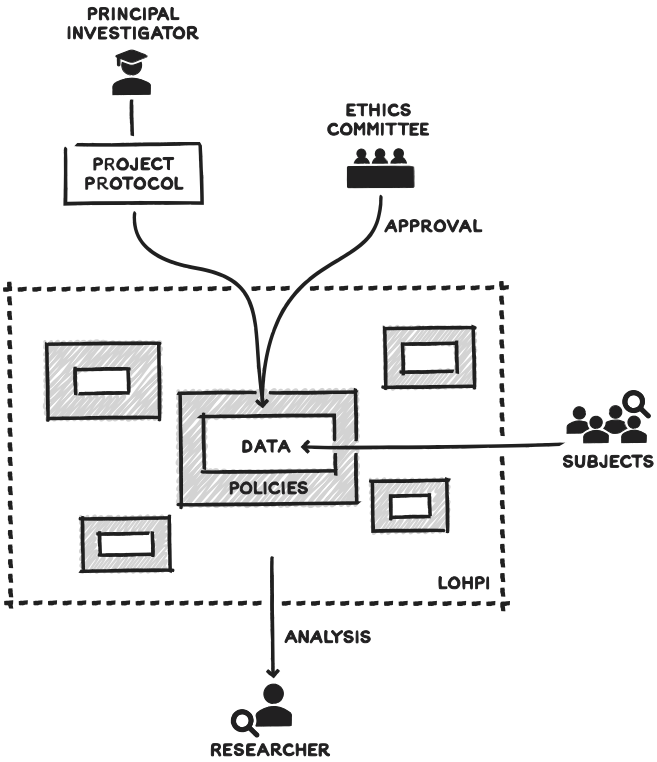

Figure 2. A research project involving data collection and dissemination with Lohpi.

means that at any time, a competent authority or a subject can request a compliance report on the project's data.

Ausloos et al. [2] argue that defining policies for responsible data-driven research should be an iterative process among the stakeholders. The data-access policy approved by the ethics committee is attached to the data collected in the project. The PIs retain their data assets, which are governed by the approved policies. These data assets individually owned by the PIs connect via gossips. The network as a whole provides a platform for researchers to do analyses. These analyses are compliant with the applicable laws and subjects' consents. Any changes to these policies, whether revocation of a *consent* or a newly added collaborator, are updated as policies disseminated through the network via a so-called gossip protocol. The stakeholders have oversight over the data they are responsible for [4, 7] and can iterate over compliant data-access policies as the threat models change.

Even though the data collected through a project might not change, the policies might. Lohpi facilitates the subjects' to have their requirements translated into verifiable policies. Researchers might not be familiar with applicable laws that apply and might be breaching them unknowingly. Lohpi enables compliant data processing and sharing which can prevent such breaches by keeping the policies up-to-date. The applicable laws endorsed by the ethics committee and the consents of the subjects of their data form an agreement for data processing. Lohpi keeps this agreement enforced by continuously monitoring and updating

data-access policies. A breach of trust can damage the relationship with the subjects. A compliant data-processing environment enabled by Lohpi will mitigate such risks and improve the relationship between researchers and the researched population.

3 Lohpi Overview

We now present an overview of Lohpi. Lohpi is intended to operate as a permissioned system, one that needs prior approval before being used by research organizations that cooperate on a potential large portfolio of research projects. Institutions host their project data at *storage nodes*, typically located either on a secure campus infrastructure or on a public cloud. These nodes form a data-storage substrate that runs our secure Fireflies overlay-network protocol [27]. This decentralized network of nodes stores the research data along with their recent data-access policies. Figure 3 shows the overview of Lohpi.

Researchers interested in accessing data are required to authenticate with one or more institutions.Lohpi allows institutions to join the network and integrate their identity management systems based on OpenID [28]. The stakeholders can issue policy changes that are propagated to the data storage network via *gossips* with the underlying Fireflies network. The compliance engine facilitates researchers to analyze the data and stakeholders to perform audits. As argued earlier, audits can provide a clearer picture to the stakeholders about the data use. A policy change is stored at the policy store. The policy store also propagates these changes into the data storage network as gossip messages.

We now briefly explain components of Lohpi (see Figure 3). *Subjects* refer to the researched population that contributes data about themselves in a project. An ethics committee, also known as Research Ethics Committee (REC), Institutional Review Board (IRB), or Data Protection Officer (DPO), is charged with the task of ensuring that a research project complies with all laws, regulations, and ethical standards. Therefore, throughout this paper, we show the functions of Lohpi in the context of an ethics committee concerning research data sharing and processing. A data storage node stores one or more study data. They are managed by institutions themselves with a Lohpi communication substrate running on them. The nodes can be hosted on an institution's infrastructure or a public cloud platform such as Microsoft

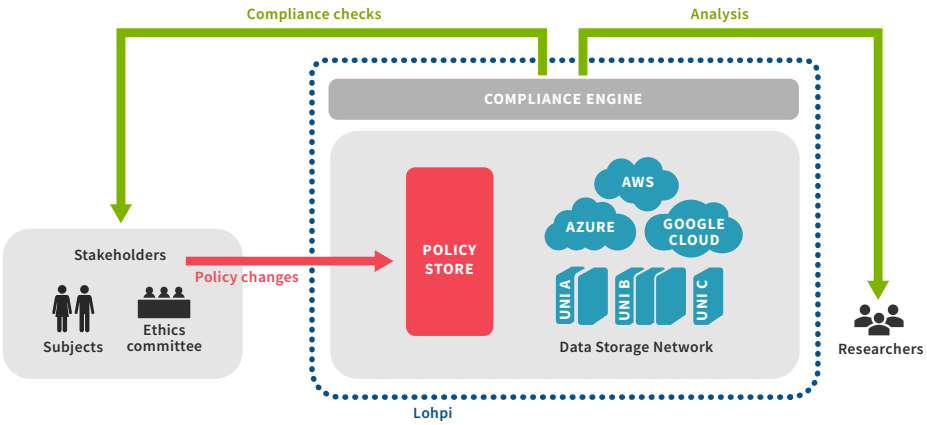

Figure 3. Lohpi system overview.

Azure, Amazon Web Service, or Google Cloud. The nodes form a data storage network based on Fireflies [27] and use TLS-based secure communication between them. The policy store stores and propagates policies for research data stored in the data storage network. It also stores policies' history in a *git*-like manner. The policy store can also probe the data storage network for any configuration issues or communication losses. The compliance engine performs audits requested by the stakeholders. A detailed description of Lohpi components is available in [29]. For brevity, the components are discussed briefly in this paper.

Instead of a centralized access control mechanism, each node has embedded data-access control. The policies for such control are updated via gossip messages. Once the policies reach the target node, they are encoded into the file system. Later in Section 5, we show the overhead of such access control. Lohpi is designed to support compliant processing by a benign user (researcher). However, a highly knowledgeable attacker or someone with physical access to the computer network can bypass these mechanisms. In addition to providing compliant access to researchers, Lohpi allows stakeholders to request compliance reports. These reports can be predefined to obtain a holistic view of data usage. For subjects, it may be of interest to see what their data is being used for [16] and update their policies.

4 Evaluation

A key property of Lohpi is the reliable dissemination of policy updates. Therefore, we evaluate the propagation of the updates as gossips, issued by an ethics committee and introduced to Lohpi by the policy store. We designed a set of micro-benchmarks to evaluate how much time it takes to propagate a data-access policy change. These experiments focus on the time required to propagate an update under different conditions.

Let ϕ be the percent of the data storage network that must receive a gossip message to consider it successful. σ represents the number of nodes to which the policy store multicasts the update directly. For example, if the policy store multicasts the message to one node, $\sigma = 1$. We begin by simulating the growth of the total number of nodes N in Lohpi. We assign a static value to ϕ and introduce policy updates by the policy store. To consider a policy update successful, the policy store must receive k acknowledgments from different nodes (see Eqn. 1). We measure the time elapsed after the policy store multicasts the message to σ set of nodes and then waits to receive k acknowledgments. We arbitrarily chose the message size to 512 KiB. We take measurements at least three times to calculate the uncertainty and plot them using error bars. After recording the first set of readings, we increase the value of σ, by doubling it and take a further set of readings.

$$k = \max\left(\lceil \phi \times |N| \rceil, \phi \times |N| + 1\right) \tag{1}$$

We also evaluate the overhead added by the access controls. First, we measure the *baseline* by reading multiple files from the file system without any access controls introduced by Lohpi. After enabling the access controls, we perform the same read operations and measure time. We measure the time required to read a large chunk of 1 GiB of data.

5 Results

In Figure 4 we show the time required for reaching at least two-thirds of the data storage network. We can observe that the time required to reach the acceptance level grows exponentially with the number of nodes (N) in the network. We also observe that by increasing the value of σ, we can propagate the message faster through the network. However, the gains are not significant at lower values of N. Only with $N \geq 32$, we start to observe significant

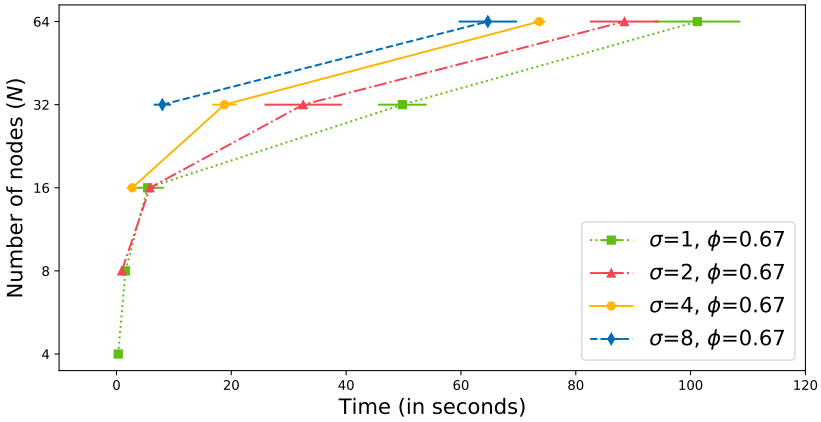

Figure 4. Time to reach k nodes with $\phi = 0.67$.

gains. Also, the variability in the time increases with the size of the network. Failures in the network or nodes can increase the propagation time, however, this can be mitigated.

We also evaluated the access control overhead for file read operations. The results (Figure 5) show that the overhead is significantly large ($\geq$15%) when the file sizes are smaller than 64 KiB. As the file size grows, the overhead becomes negligible.

6 Related Work

Dataverse [21] is a centralized repository where researchers can deposit their data. Researchers can add custom licenses. Once a dataset is downloaded from Dataverse, there are no mechanisms to restrict sharing through any other means such as over FTP or a USB drive. Wolley et al. [30] introduced the Automatable Discovery and Access Matrix (ADA-M) that

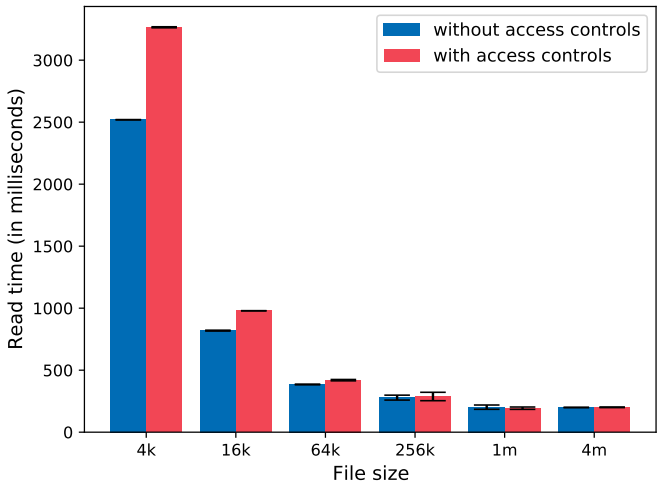

Figure 5. Read operation with data access controls.

allows stakeholders to confidently track, manage, and interpret applicable legal and ethical requirements. The ADA-M metadata profiles allow an ethics committee to evaluate and approve information models linked to a dataset. ADA-M facilitates responsible sharing outlined in the profile and allows the custodian to check the accesses against regulatory parameters. However, they do not mention any functionality about issuing updates to the profile. Alter et al. [31] presented Data Tags Suite (DATS), which can be used to describe data access, use conditions, and consent information. DATS provides a metadata exchange format without any compliance checking mechanisms. Havelange et al. [32] developed a blockchain-based smart contract to attach license requirements to a dataset. The datasets are encrypted and ADA-M profiles are attached with each dataset. A researcher accepts the contract and receives a token to decrypt the dataset. The researcher's data accesses are checked against the ADA-M profile for compliance. However, they require each researcher, dataset provider, and supervisory authority to have a node on the Ethereum-blockchain network. They do not provide any evaluation in their work.

7 Discussion

Our prototype implementation demonstrates that it is possible to propagate updated policies close to real-time. We conjecture that even with a larger distributed storage network, policy changes can be propagated within minutes. We also conjecture that transparency in research data processing can increase trust in research institutions. Adapting protection mechanisms to newly discovered threats to protect individuals involved in research can help sustain public trust [33]. With OpenID [28] Lohpi can integrate with existing authentication services used at various institutions.

While Lohpi's approach is not centralized, we conjecture that it can provide abstractions for an ethics committee. Such abstractions can periodically measure compliance on the data storage network and mitigate privacy risks. An expressive policy language like Guardat [34] can be realized using *meta-code* [17, 35]. We are also interested in building a tool to express research data usage protocol for streamlining ethics committee approvals and verifying compliance against an approved protocol. We are interested in making existing research data available on Lohpi.

8 Conclusion

We presented a distributed infrastructure to support compliant data analytics for human subject research. We demonstrated that a distributed gossiping network can ensure the timely delivery of policy changes. The architecture can scale even when a research project spans multiple regulatory bodies. In Lohpi, a data storage node can run on the public cloud or on-campus hardware. Institutions can easily join the network without the need to move their research data.

Acknowledgment

This work was funded in part by Research Council of Norway project numbers 263248 and 275516. We thank Katja Pauline Czerwinska for her assistance with the graphics.

References

[1] A. Weichselbraun, P. Kuntschik, V. Francolino, M. Saner, U. Dahinden, V. Wyss, *Adapting data-driven research to the fields of social sciences and the humanities*, Future Internet **13**, 1 (2021)

[2] J. Ausloos, R. Heyman, N. Bertels, J. Pierson, P. Valcke, *Designing-by-Debate: A Blueprint for Responsible Data-Driven Research & Innovation*, Springer pp. 47–63 (2018)

[3] G. Schneider, *Disentangling health data networks: A critical analysis of Articles 9(2) and 89 GDPR*, International Data Privacy Law **9**, 253 (2019)

[4] B. Shneiderman, *Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems*, ACM Transactions on Interactive Intelligent Systems **10**, 1 (2020)

[5] J.P. Daries, J. Reich, J. Waldo, E.M. Young, J. Whittinghill, A.D. Ho, D.T. Seaton, I. Chuang, *Privacy, anonymity, and big data in the social sciences*, Communications of the ACM **57**, 56 (2014)

[6] World Health Organization, *Research Ethics Committees : Basic Concepts for Capacity-Building* (World Health Organization, 2009), ISBN 9789241598002

[7] J.A. Kroll, N. Kohli, P. Laskowski, *Privacy and Policy in Polystores: A Data Management Research Agenda*, Lecture Notes in Computer Science **11721 LNCS**, 68 (2019)

[8] D. McGraw, C. Petersen, *From Commercialization to Accountability: Responsible Health Data Collection, Use, and Disclosure for the 21st Century*, Applied Clinical Informatics **11**, 366 (2020)

[9] K.W. Goodman, E.M. Meslin, *Ethics, Information Technology, and Public Health: Duties and Challenges in Computational Epidemiology*, in *Public Health Informatics and Information Systems* (Springer, 2014), pp. 191–209

[10] L.L. Roos, M. Brownell, L. Lix, N.P. Roos, R. Walld, L. MacWilliam, *From health research to social research: Privacy, methods, approaches*, Social Science and Medicine **66**, 117 (2008)

[11] C. Dwork, F. McSherry, K. Nissim, A. Smith, *Calibrating noise to sensitivity in private data analysis*, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Springer, 2006), Vol. 3876 LNCS, pp. 265–284, ISBN 3540327312, ISSN 03029743

[12] X.J. Zhang, X.F. Meng, *Differential privacy in data publication and analysis*, in *Jisuanji Xuebao/Chinese Journal of Computers* (2014), Vol. 37, pp. 927–949, ISSN 02544164

[13] S.L. Garfinkel, J.M. Abowd, S. Powazek, *Issues encountered deploying differential privacy*, in *arXiv* (2018), pp. 133–137, ISSN 23318422

[14] V.M. Suriyakumar, N. Papernot, A. Goldenberg, M. Ghassemi, *Chasing your long tails: Differentially private prediction in health care settings*, in *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021), pp. 723–734, ISBN 9781450383097, 2010.06667

[15] A.Z. Woldaregay, A. Henriksen, D.Z. Issom, G. Pfuhl, K. Sato, A. Richard, C. Lovis, E. Årsand, J. Rochat, G. Hartvigsen, *User expectations and willingness to share self-collected health data*, Studies in Health Technology and Informatics **270**, 894 (2020)

[16] A. Sharma, K.P. Czerwinska, L. Brenna, D. Johansen, p. Johansen, Håvard D., *Privacy perceptions and concerns in image-based dietary assessment systems: Questionnaire-based study*, JMIR Human Factors **7** (2020)

[17] H.D. Johansen, E. Birrell, R. Van Renesse, F.B. Schneider, M. Stenhaug, D. Johansen, *Enforcing Privacy Policies with Meta-Code*, in *6th Asia-Pacific Systems Workshop, APSys 2015* (ACM, 2015), p. 16, ISBN 9781450335546

[18] REK, *Regionale komiteer for medisinsk og helsefaglig forskningsetikk*, Accessed 31/04/2021 rekportalen.no

[19] R. Schroeder, *Big Data and the brave new world of social media research*, Big Data and Society **1** (2014)

[20] The Dataverse Project, *Metrics*, Retrieved 31/04/2021 dataverse.org/metrics (2021)

[21] G. King, *An introduction to the dataverse network as an infrastructure for data sharing* (2007)

[22] J. Salerno, B.M. Knoppers, L.M. Lee, W.W.M. Hlaing, K.W. Goodman, *Ethics, big data and computing in epidemiology and public health*, Annals of Epidemiology **27**, 297 (2017)

[23] A.S.Y. Cheung, *Moving beyond Consent for Citizen Science in Big Data Health Research*, SSRN Electronic Journal **16**, 15 (2017)

[24] M. Camden, *A 'Microdata for Research'sample from a New Zealand census*, Monographs of official statistics p. 117 (2005)

[25] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.W. Boiten, L.B. da Silva Santos, P.E. Bourne et al., *Comment: The FAIR Guiding Principles for scientific data management and stewardship*, Scientific Data **3**, 1 (2016)

[26] P. Holub, F. Kohlmayer, F. Prasser, M.T. Mayrhofer, I. Schlünder, G.M. Martin, S. Casati, L. Koumakis, A. Wutte, Z. Kozera et al., *Enhancing Reuse of Data and Biological Material in Medical Research: From FAIR to FAIR-Health*, Biopreservation and Biobanking **16**, 97 (2018)

[27] H.D. Johansen, R. Van Renesse, Y. Vigfusson, D. Johansen, *Fireflies: A secure and scalable membership and gossip service*, ACM Transactions on Computer Systems **33**, 1 (2015)

[28] D. Recordon, D. Reed, *OpenID 2.0: A platform for user-centric identity management*, in *Proceedings of the Second ACM Workshop on Digital Identity Management, DIM 2006. Co-located with the 13th ACM Conference on Computer and Communications Security, CCS'06* (2006), pp. 11–16, ISBN 1595935479

[29] A. Sharma, T.B. Nilsen, K.P. Czerwinska, D. Onitiu, L. Brenna, D. Johansen, H.D. Johansen, *Up-to-the-minute Privacy Policies via gossips in Participatory Epidemiological Studies*, Frontiers in Big Data **4**, 14 (2021)

[30] J.P. Woolley, E. Kirby, J. Leslie, F. Jeanson, M.N. Cabili, G. Rushton, J.G. Hazard, V. Ladas, C.D. Veal, S.J. Gibson et al., *Responsible sharing of biomedical data and biospecimens via the "Automatable Discovery and Access Matrix" (ADA-M)*, npj Genomic Medicine **3**, 1 (2018)

[31] G. Alter, A. Gonzalez-Beltran, L. Ohno-Machado, P. Rocca-Serra, *The Data Tags Suite (DATS) model for discovering data access and use requirements*, GigaScience **9** (2020)

[32] A. Havelange, M. Dumontier, B. Wouters, J. Linde, D. Townend, A. Riedl, V. Urovi, *LUCE: A blockchain solution for monitoring data License accoUntability and CompliancE*, arXiv (2019), 1908.02287

[33] Anna C. Mastroianni, *Sustaining Public Trust: Falling Short in the Protection of Human Research Participants*, Hastings Center Report **38**, 8 (2008)

[34] A. Vahldiek-Oberwagner, E. Elnikety, A. Mehta, D. Garg, P. Druschel, R. Rodrigues, J. Gehrke, A. Post, *Guardat: Enforcing data policies at the storage layer*, in *Proceedings of the 10th European Conference on Computer Systems, EuroSys 2015* (2015), pp. 1–16, ISBN 9781450332385

[35] D. Johansen, J. Hurley, *Overlay cloud networking through meta-code*, in *Proceedings - International Computer Software and Applications Conference* (IEEE, 2011), pp. 273–278, ISBN 9780769544595, ISSN 07303157