Designing a Service for Compliant Sharing of Sensitive Research Data

Aakash Sharma[0000−0003−3965−3271], Thomas Bye Nilsen[0000−0002−3602−451X], Sivert Johansen[0000−0003−3442−2466], Dag Johansen[0000−0001−7067−6477], and Håvard D. Johansen[0000−0002−1637−7262]

UiT The Arctic University of Norway, Tromsø, Norway
aakash.sharma@uit.no

Abstract. Data-driven research is increasingly becoming fueled by access to open datasets, often shared publicly on the Internet. However, many research projects study sensitive data. They cannot easily participate in this shift as access to their data is significantly controlled by ethical and regulatory constraints. This paper discusses the requirements for building a service that enables sensitive data for sharing between collaborators in a controlled manner. We argue that a decentralized service that maintains metadata, a global view on all data usage, and active policy combined with local monitoring and security enforcement can provide automated compliance checking. With such a service, researchers can share sensitive data with a broader community rather than limiting access to core project members.

Keywords: open data · data-sharing · compliance · sensitive data

1 Introduction

The Internet has changed the way researchers work, collaborate, and disseminate. Open Science is a cultural change [2]. The arguments concerning the benefits of Open Data are well established; for example allowing researchers to explore existing datasets in new ways [1, 4, 10]. Volunteers (hereafter written as *subjects*) contribute their data for research. The trustworthiness of the institution conducting the research plays a key role in a subject's willingness to contribute [17]. Privacy leaks or misuse can damage the reputation and affect future research studies [1, 4]. Fears of misuse of data may also restrict many researchers from sharing data openly [4, 20].

Researchers argue that these concerns can be mitigated by building accountability in research data sharing and processing [8]. Recent regulations, such as General Data Protection Regulation (GDPR) require researchers to abide by a subject's consent for data processing. GDPR also provides workarounds for public-funded research by entrusting a Regional Ethics Committee (REC) or an Institutional Review Board (IRB) to protect subjects' privacy. The public's trust in researchers is fragile [1]. The growing concerns regarding data breaches, data brokers, and indiscriminate profiling of users might change subjects' willingness to participate or continue participating in a research project.

The guidelines and complexity of compliance is a tedious job and requires a complex understanding of legal, ethical, and regulatory issues [5]. Often institutions employ large teams to assist researchers in making their data openly available [18]. Researchers' concern about misuse of their data is the leading reason given for not sharing data [4, 20]. As a result, research data may end up in silos accessible only to a limited few. A lot of work has been done for simplifying regulatory requirements, easy-to-create toolkits [5, 18, 21] and metadata formats for making research openly accessible [7, 12, 21]. However, additional regulations like *data sovereignty* [13] may further restrict Open Data. For example, medical research data is heavily regulated. Often movement of such sensitive data is restricted outside a nation's physical boundaries.

Open science and data-based collaboration require access to the same data regardless of international borders [2, 11]. As argued earlier, there might be regulatory restrictions limiting the sharing of sensitive data. The cloud provides an interesting platform for Open Data access to researchers with manageable services. Our contribution is a scalable cloud-based service that allows researchers to analyze sensitive data regardless of their location. We discuss related works in Section 2, and present the requirements for the service in Section 3. Later in Section 4, we present our system's design, key features, and limitations.

2 Related Work and Discussion

Dataverse [7] is a well-known data repository for sharing research data, which currently hosts tens of thousands of datasets. However, Dataverse does not support sensitive data. Datatags system [18] translates security and access requirements for sensitive data into a model set of six tags. Their approach simplifies the complex workings and guidelines for sharing datasets responsibly as they provide a decision tree for picking a correct tag for different requirements. The Datatags approach simplifies complex information flows for IRBs and RECs without specifying mechanisms for automated audits or enforcement.

Automatable Discovery and Access Matrix (A-DAM) [21] provides a *profile* as regulatory metadata for responsible sharing of biomedical assets. A-DAM provides a semi-automated approach for analyzing ethical and regulatory requirements for sharing and processing research data. Policy changes require a newer profile and reevaluation. Maguire et al. [9] proposed a metadata-based architecture for accountability. Similar to A-DAM, Maguire et al.'s approach attaches a static policy to a dataset, which is verified by a gatekeeper service. Their approach introduces validation against *context* by the gatekeeper. For sensitive data, they only briefly discuss adding encryption and keeping the keys under the control of the gatekeeper. In our earlier work Lohpi [15, 16], we argued that the changes occurring during a project's life cycle might affect its data security policies. Thus, we built support for accountability by keeping data security policies up to date securely and efficiently. We build upon existing works to semi-automate regulatory, ethical, and legal requirements. Our contribution is a scalable cloud-based service that allows researchers to analyze sensitive data regardless of their location. The service provides transparency to stakeholders, such as subjects regarding data sharing and data use.

Axelsson and Schroeder [1] argue that public trust is fragile, and once broken, might take years to re-build. Compliance and transparency are crucial for maintaining the fragile public trust in researchers. The complex set of guidelines, regulatory and legal requirements, and consent management increase complexity for data curators and researchers. And keeping sensitive data open is a challenging [1]. Even experts in various fields feel the lack of assurances [6] in existing practices. We provide compliance with audit-able data sharing of sensitive datasets. Our approach simplifies access to such datasets by automating compliance while maintaining compliance. The compliance requirements are derived from applicable regulatory and legal requirements. Many researchers have argued for building transparency for data-sharing, usage, and privacy protections [8, 14]. Along with these themes, our approach addresses the compliant sharing of sensitive data, especially data sovereignty. Thus, allowing sensitive datasets to reach a broader audience while fulfilling regulatory and legal compliance requirements. The built-in transparency allows stakeholders, such as subjects, to understand the usage of their data and answer questions like who, whom, and where, about their contributed data. Such transparency may improve the public's trust and participation in studies that rely heavily on volunteers.

3 Requirements for the Service

As argued earlier, Open Data should be able to reach a broader audience. Our goal is to build a service that supports sharing of sensitive data and addresses the compliance requirements. We now discuss the requirements for building services for researchers to share sensitive research data. The regulations, re-identification attack methods, and legal and ethical requirements may change over time. We conjecture that the following requirements are essential for building a service compliant with the legal, regulatory, and ethical requirements stipulated by concerned authorities/stakeholders. The service can adapt to changes that may affect data sensitivity and a subject's preference. Existing works like Datatags [18] and A-DAM [21] provide methods for computable ethical, legal, and regulatory requirements. Different security mechanisms can enforce these requirements [18]. Additionally, we include the *data sovereignty* requirement for sensitive data, which restricts the movement of data outside a nation's physical boundaries, even if the data are hosted, by a cloud service provider (CSP).

RQ 1. (Timely Dissemination of Data Policies) Data policies define the ethical, legal, and regulatory requirements attached with a dataset. The service should disseminate changes to data policies within a predefined time τ. Each dissemination should be secure, maintain integrity, and be logged, for auditing. Consent revocations and new approvals from an IRB or a REC can result in such changes.

A change in laws, regulations and institutional guidelines may result in a policy change as well.

RQ 2 (Data Sovereignty)**.** The service should ensure data sovereignty by verifying data residency. Any attempt of data movement which violates data sovereignty should be prevented and logged.

RQ 3 (Garbage Collection)**.** Once completed with the task, copies of data should be securely deleted and logged. No residual copies of the data or the dataset remain on an unsolicited location or machine.

RQ 4 (Auditing)**.** The service should log each operation and action in a distributed log. These operations and actions must be available for auditing by an IRB, REC or an independent auditing authority.

RQ 5 (Secure Computation)**.** The continuous access evaluation should be securely computed at the CSP. The attack surface for tampering with the data access policy should be limited.

4 System Overview

We now describe our approach and discuss how different components will fulfill the requirements discussed earlier (Section 3). We assume that each dataset has a unique identifier. Researchers who are interested in accessing a dataset can authenticate themselves. The existing data security policy allows authenticated researchers to access the dataset. The dataset in this example has data sovereignty constraints. A CSP has a data center in the same region as the dataset.

4.1 Workflow

Fig. 1 shows the system architecture with an example workflow. A researcher or simply a *user* is interested in analyzing a dataset hosted at an institution in another country. After authenticating herself using the web portal, the user configures a machine for her analysis (RQ4). The user can decide from multiple pre-configured container images which contain different data analysis software packages. These container images enable communication with a trusted substrate for exchanging data policy updates and logs (RQ1, RQ4). The daemon software is pre-installed and configured in these container images. For enhanced security, the container images are pre-configured to limit data egress and allow only a set of pre-approved packages. As the last step, the user chooses the dataset that she is interested in (RQ5).

Once configured, a new container instance with chosen software packages runs in the cloud which, exists in the same country as the dataset (RQ2). The user obtains access to the container running in the cloud. Upon initialization, the user needs to authenticate herself again for obtaining a copy of the dataset

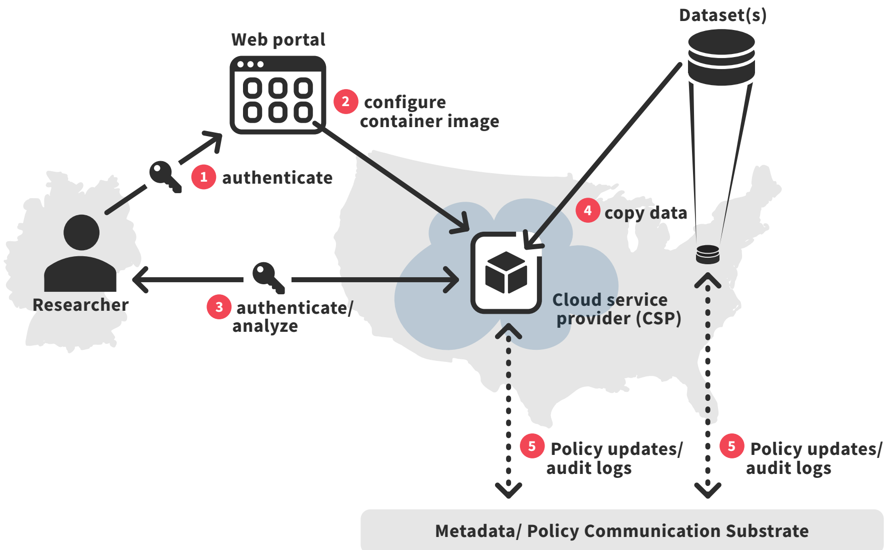

Fig. 1. An example workflow: A researcher from Germany wants to access a sensitive dataset from an institute in the USA. The movement of the data outside the USA is restricted. By leveraging a CSP, we facilitate sharing of the dataset for research without moving the data outside the USA. The accesses are logged for audit and transparency reports. The data security policies are updated using the communication substrate.

(RQ4). The instance also receives policy changes that might arrive while the user is working on the dataset after the initialization (RQ1).

The instance enforces *use-based* policies using Intel's software guard extensions (SGX) using approaches like [3] (RQ5). In-line monitoring using SGX can result in a performance penalty. For performance reasons, we use the delegated monitoring architecture proposed in Birrel et al. [3]'s work. The instance communicates with the metadata communication substrate and keeps the metadata/policy up-to-date (RQ1). The changes to a checked-out dataset's policy are disseminated through the substrate. Both the original dataset and the copy in the container receive the changes via the substrate. After receiving metadata updates, the compliance and access are reevaluated (RQ5). The daemon process routinely checks for compliance with the latest data security policies. These work behind the scenes and notify the user if additional inputs are required. Thus, making compliance easier for the user. In case of non-compliance, the user may lose access to the machine while saving the image to save her work-in-progress. After resolving the non-compliance issue, the user can regain access. The user can export the analysis' results in different pre-approved file formats to the web portal (RQ3). Through the portal, the user can obtain the results later. The results can be archived at the portal for cross-examination by auditors and reviewers for scientific peer-review processes.

At the end of the analysis, the user can terminate the instance, and the analysis scripts and the dataset copy are securely destroyed (RQ3). A user may also choose to save the current state of the container for reproducibility of results [19]. The sharing and accesses generate logs containing sanitized information for audits. The stakeholders (subjects, REC or IRB) can also view reports on data use and sharing and intervene if necessary. Oversight committees (RECs or IRBs) can review non-compliance incidents and take necessary actions. These actions can be in the form of policy updates propagated to every copy of the dataset. Securing a container image while preserving reproducibility is beyond the scope of this paper.

Limitations Our approach protects against a benign threat model and assumes an accountable adversary. The system may not protect against a sophisticated attacker. The availability of a cloud service provider (CSP)'s container services in the same administrative region as of a dataset's location is crucial for the data sovereignty requirements.

5 Conclusion

Regulatory compliance in research data sharing is a developing problem with newly introduced regulations and growing concerns about individual privacy. The relationship between subjects and research institutions relies heavily on trust for voluntary participation. Data sharing and use, compliant with the subjects' wishes is crucial for continued participation and sustaining trust. We discussed the requirements for building a service enabling compliant data use and sharing sensitive research data. We further presented our approach for building such a service and how it addresses those requirements. We plan to test the system with our partners from sports sciences and medical science, creating policy templates for legal and regulatory requirements for sensitive research data.

Acknowledgements We thank Katja Pauline Czerwinska for her assistance with the graphics. This work was funded in part by Research Council of Norway project numbers 263248 and 274451.

Bibliography

[1] Axelsson, A.S., Schroeder, R.: Making it open and keeping it safe: e-enabled data-sharing in Sweden. Acta Sociologica **52**(3), 213–226 (2009), ISSN 00016993, https://doi.org/10.1177/0001699309339799

[2] Bartling, S., Friesike, S.: Opening science: the evolving guide on how the web is changing research, collaboration and scholarly publishing, vol. 51. Springer Nature (2014), https://doi.org/10.5860/choice.51-6715

[3] Birrell, E., Gjerdrum, A., Van Renesse, R., Johansen, H., Johansen, D., Schneider, F.B.: SGX enforcement of use-based privacy. In: Proceedings of the ACM Conference on Computer and Communications Security, pp. 155–167 (2018), ISBN 9781450359894, ISSN 15437221, https://doi.org/10.1145/3267323.3268954

[4] Borgman, C.L.: Open Data, Grey Data, and Stewardship: Universities at the Privacy Frontier. Berkeley Tech. LJ **33**, 365 (2018), URL http://arxiv.org/abs/1802.02953

[5] Braunschweig, K., Eberius, J., Thiele, M., Lehner, W.: The State of Open Data. Limits of current open data platforms. Www2012 (2012)

[6] Hammack, C.M., Brelsford, K.M., Beskow, L.M.: Thought Leader Perspectives on Participant Protections in Precision Medicine Research. Journal of Law, Medicine and Ethics **47**(1), 134–148 (2019), ISSN 1748720X, https://doi.org/10.1177/1073110519840493

[7] King, G.: An introduction to the dataverse network as an infrastructure for data sharing (2007), https://doi.org/10.1177/0049124107306660

[8] Kroll, J.A., Kohli, N., Laskowski, P.: Privacy and Policy in Polystores: A Data Management Research Agenda. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **11721 LNCS**, 68–81 (2019), ISSN 16113349, https://doi.org/10.1007/978-3-030-33752-0_5

[9] Maguire, S., Friedberg, J., Nguyen, M.H., Haynes, P.: A metadata-based architecture for user-centered data accountability. Electronic Markets **25**(2), 155–160 (2015), ISSN 14228890, https://doi.org/10.1007/s12525-015-0184-z

[10] Molloy, J.C.: The open knowledge foundation: Open data means better science. PLoS Biology **9**(12) (2011), ISSN 15449173, https://doi.org/10.1371/journal.pbio.1001195

[11] Mulligan, A., Mabe, M.: The effect of the internet on researcher motivations, behaviour and attitudes. Journal of Documentation **67**(2), 290–311 (2011), ISSN 00220418, https://doi.org/10.1108/00220411111109485

[12] Pampel, H., Vierkant, P., Scholze, F., Bertelmann, R., Kindling, M., Klump, J., Goebelbecker, H.J., Gundlach, J., Schirmbacher, P., Dierolf, U.: Making research data repositories visible: The re3data.org registry. PLoS ONE **8**(11), e78080 (2013), ISSN 19326203, https://doi.org/10.1371/journal.pone.0078080

[13] Peterson, Z.N., Gondree, M., Beverly, R.: A position paper on data sovereignty: The importance of geolocating data in the cloud. 3rd USENIX Workshop on Hot Topics in Cloud Computing, HotCloud 2011 (2011)

[14] Schneider, G.: Disentangling health data networks: A critical analysis of Articles 9(2) and 89 GDPR. International Data Privacy Law **9**(4), 253–271 (2019), ISSN 20444001, https://doi.org/10.1093/idpl/ipz015

[15] Sharma, A., Nilsen, T.B., Brenna, L., Johansen, D., Johansen, H.D.: Accountable Human Subject Research Data Processing using Lohpi. In: Proceedings of the ICTeSSH 2021 conference (2021), https://doi.org/10.21428/7a45813f.80ebd922

[16] Sharma, A., Nilsen, T.B., Czerwinska, K.P., Onitiu, D., Brenna, L., Johansen, D., Johansen, H.D.: Up-to-the-Minute Privacy Policies via Gossips in Participatory Epidemiological Studies. Frontiers in Big Data **4**, 14 (2021), ISSN 2624909X, https://doi.org/10.3389/fdata.2021.624424

[17] Slegers, C., Zion, D., Glass, D., Kelsall, H., Fritschi, L., Brown, N., Loff, B.: Why Do People Participate in Epidemiological Research? Journal of Bioethical Inquiry **12**(2), 227–237 (2015), https://doi.org/10.1007/s11673-015-9611-2

[18] Sweeney, L., Crosas, M., Bar-Sinai, M.: Sharing Sensitive Data with Confidence: the Datatags System. Technology Science pp. 1–34 (2015), URL http://techscience.org/a/2015101601/

[19] Trisovic, A., Durbin, P., Schlatter, T., Durand, G., Barbosa, S., Brooke, D., Crosas, M.: Advancing Computational Reproducibility in the Dataverse Data Repository Platform. In: P-RECS 2020 - Proceedings of the 3rd International Workshop on Practical Reproducible Evaluation of Computer Systems, pp. 15–20 (2020), ISBN 9781450379779, https://doi.org/10.1145/3391800.3398173

[20] Winthrop, S.: Social considerations to make data FAIR-er: Understanding researchers' views on data "misuse" and credit. Septentrio Conference Series (1) (2019), https://doi.org/10.7557/5.4970, URL https://doi.org/10.6084/m9.figshare.10011788.v2;

[21] Woolley, J.P., Kirby, E., Leslie, J., Jeanson, F., Cabili, M.N., Rushton, G., Hazard, J.G., Ladas, V., Veal, C.D., Gibson, S.J., Tassé, A.M., Dyke, S.O., Gaff, C., Thorogood, A., Knoppers, B.M., Wilbanks, J., Brookes, A.J.: Responsible sharing of biomedical data and biospecimens via the "Automatable Discovery and Access Matrix" (ADA-M). npj Genomic Medicine **3**(1), 1–6 (2018), ISSN 20567944, https://doi.org/10.1038/s41525-018-0057-4