Computer-supported ethical rules for collaboratively sharing data

Rui Zhao, Malcolm Atkison
Petros Papapanagiotou, Jacques Fleuriot
School of Informatics
University of Edinburgh
{rui.zhao,m.atkinson,pe.p,jdf}@ed.ac.uk

Christian Pagé
CECI
Université de Toulouse, CNRS, Cerfacs
christian.page@cerfacs.fr

ABSTRACT

Rapid response to pressing societal and business challenges requires consortia joining forces to focus their resources including specialist skills, knowledge, methods and data. The formation and maintenance of the required collaborations depends on rules that cover ethical, legal, privacy and business issues. As complexity and scale grow, maintaining compliance and understanding becomes very difficult. This pushes collaborations to polarized extremes, *everything public* or *everything private*. We propose that computer support for rules will ease the burden on the people and enable more sophisticated forms of collaborations. This requires a (semi-)automated framework incorporating diverse stakeholders to shape the rules, govern their interpretation, and deliver unbiased diagnosis and mitigation of conflicts. Its robust depends on formalisation of rules crossing organisational and technical boundaries, enabling automatic composition and modification of rules propagated through multi-input-multi-output cross-boundary data flows. This requires a research campaign with sufficiently diverse contributors. We report two case studies to clarify requirements, and then reshape the requirements for building the framework. Our early prototype suggests a direction for that research. This will underpin future systems supporting agile, extensive and sustainable collaborations.

CCS CONCEPTS

• **Security and privacy** → **Usability in security and privacy**; • **Social and professional topics** → **Computing / technology policy**.

KEYWORDS

data policy, cross-boundary collaboration, data governance

INTRODUCTION

Large-scale collaboration between independent institutions is increasing in importance in a wide range of endeavours. In those collaborations, data sharing and data governance are crucial for success and sustainability. Many tasks depend on flexibly sharing data, which is only permitted if all stakeholders believe their constraints on its use will be respected. This requires them to trust that their rules will not be broken in other parts of the federation or by future users. As citizen science and Internet of Things (IoT) become more pervasive, sustainability and agility will be even more pressing. We can not expect everyone to repeatedly invest time in finding, reading and understanding the data use policies.

To handle ethical or regulatory issues, existing research covers general discussion [9, 12, 13], legal or regulatory requirements and actions [2, 19], and computer-supported compliance handling [6, 11, 14]. Even though research has been carried out in building a digital data market [4], there is still limited support for writing and handling cross-institutional data policies. This diminishes the agility of data sharing and collaboration between different bodies. The consequence is *polarization*, where data policies are invariably pushed to one of two extremes:

(1) no or minimal policies for data use, sometimes called “Open Data”;
(2) strongly restrictive policies granting limited access after training with enforced constraints.

In practice, although the essential requirements lie between these two, the lack of support forces data governors to choose the restricted option to avoid any risk. This issue is worsening as collaboration needs to be more pervasive and agile but the time spent on agreeing or respecting data use policies under current models ends up taking a significant proportion of the time and effort, inhibiting needed developments. This also exposes the problem that communication between the data providers and the data users is not well-established. For data-users aiming to achieve specific goals through collaborative R&D, policies are hard-to-find and hard-to-understand. Their need to focus on the task at hand is a common reason why even those most willing to comply fail to comply with data rules.

Therefore, the need for computer support to handle the ethical and governance issues for data has emerged. Existing research demonstrated computer support for intra-institution compliance is feasible, and we argue that it is also feasible for cross-boundary data-policy handling. Computer automation for data rules is unlikely to replace human effort completely, but it can significantly reduce the effort needed, increase the levels of compliance, and improve trust. Many ethical concerns will automatically vanish with sufficient encoding of policies and automatic process matching for declassification – computers will not be biased given unbiased directives.

CASE STUDIES

In this section, we present two cases representing different types of restrictions in data policies and demonstrate why the current practice does not work well. A summary is presented afterwards.

Climate data

The global infrastructure and rules for sharing climate data has a long history with careful governance [5]. Successive campaigns coordinated among research centers, often including downscaling from these and regional analyses of observations lead to data that is open to the public[1]. One aim is applications for using the data for public welfare, e.g. mitigating hazards.

These uses include influencing infrastructure planning, helping farmers plan crops and assessing insurable risk. They are often delivered via paid-for services permitted provided the sources of the data are properly acknowledged. Those running the services mix commercial-in-confidence data with the public data and need to maintain those confidences. Problems arise when they use the source's established authority without properly communicating the uncertainties. Even worse, they may use that implicit authority but distort the implications of the data for financial gain – in one case, a flood risk was deliberately understated in a map ostensibly derived from a regional severe Alpine precipitation hydrology model – undermining the model's reputation. To reduce misuse, mechanisms are needed to track data use and compliance with a data-authority's policies.

[1]Mainly distributed through the Earth System Grid Federation.

Healthcare data

Healthcare data include patients' private information, making them inherently sensitive. Strict rules are imposed on researchers using such data e.g., to develop precision medicine [1, 8, 18].

As an example, the electronic Data Research and Innovation Service (eDRIS)[2] of Public Health Scotland enables the use of data from the National Health Service for research purposes through a rigorous approval process [15]. This is initiated by a researcher submitting an application detailing the structure and size of data they require and how its use will deliver public benefit. Applications are approved by the Public Benefit and Privacy Panel for Health and Social Care, which may impose additional constraints, restrictions and requirements on the use of data. The researcher is required to attend training on private and ethical use of data. They then gain access to a protected virtual environment, a so-called *data safe haven*, where the data and approved software are placed. The panel must also approve any export of research results from the data safe haven, and may impose restrictions on its use and publication. Conformance is manually validated by a member of eDRIS attached to each project.

The entire process is very time-consuming. A lot of time is used for non-project related tasks, such as training, writing reports and waiting for decisions and data release. As another example, even though there are known automated methods for anonymization, e.g. averaging over a certain number of data points, the NHS always requires manual review of anonymized data and approval by the panel. There does exist some work on improving this, e.g. [17], but the general protocol for doing such research has not changed.

[2]www.isdscotland.org/Products-and-Services/eDRIS/

Emerging requirements

Several requirements are emerging. First, open data still require policies easily found and understood. Compliance is required in the same way as for private data. Second, derived data is normally governed by the same rules as the original data, but sometimes revised rules derived from the original ones and the processing may apply. This poses two requirements: identify the derived rules, and propagate them with result data. Third, automated compliance is possible for certain tasks, e.g. declassification.

SOLUTION VISION – A SEMI-AUTOMATED FRAMEWORK

The situation can be significantly improved with an appropriate computer-supported framework. Its design should meet the following requirements:

(1) Providers trust that their rules will be effective and start using more sophisticated rules;
(2) People who intend to comply with the rules will be well-informed e.g., by appropriate prompts;
(3) Rules that need to be adapted to meet new needs, e.g., legal changes, are propagated effectively;
(4) The rule language needs also to be understandable by humans and support auditing.

Given these requirements and building on current research, we propose a framework:

- A computer understandable and actionable formal language to describe the data use rules / policies (actionable [6, 11], expressive [7], formal [16]);
- A corresponding mechanism to propagate and dynamically apply rules where necessary, and modify them when data processing combines or changes the data ([10, 14]);
- A standard protocol to associate and retrieve data and its policies (e.g. [3]);
- A way to verify that the framework or reasoning is executed correctly (e.g. [20]).

Since different data will be combined, processed and propagated to different downstream uses, the framework also needs to be able to combine, change and apply the data policies to all potential users, in multi-input-multi-output data flows. This is the weak point for most existing research, whose designs often only consider single-input situations or their output is no longer checked. Our prototype language and system, reported in [21], demonstrates the potential of such a language. The work takes provenance as source, and our current work provides a formal foundation for the language.

In addition, the potential change of rules can be previewed by using the reasoner to check if the effect is as expected. The data governor, e.g., to respond to changed legislation, can then decide whether to deploy the changes (propagating it to subsequent data use) or make further revisions.

CONCLUSION

In this article, we presented the challenge of maintaining ethical and legal compliance for data in multi-bodied collaboration environments. Based on two brief case studies, we argued that a computer-supported approach is where the future lies, and outlined the key aspects of such a framework.

If the necessary R&D is completed, within five years, data providers and users will describe and discuss the data policies supported by such a framework, and carry out their research with confidence that the data use policies will be complied with. This will convince a larger audience of its advantage, and pave the way for incremental adoption and improved data-use rules that better serve diverse stakeholders' interests. Automation will only help, the rules and compliance with rules will remain a human responsibility. We hope others will join us in addressing this challenge.

ACKNOWLEDGMENTS

This work is partially supported by the EU H2020 project DARE, No. 777413.

REFERENCES

[1] Akram Alyass, Michelle Turcotte, and David Meyre. 2015. From big data analysis to personalized medicine for all: challenges and opportunities. *BMC medical genomics* 8, 1 (2015), 1–12.

[2] Cesare Bartolini, Andra Giurgiu, Gabriele Lenzini, and Livio Robaldo. 2017. Towards Legal Compliance by Correlating Standards and Laws with a Semi-automated Methodology. In *BNAIC 2016: Artificial Intelligence (Communications in Computer and Information Science)*, Tibor Bosse and Bert Bredeweg (Eds.). Springer International Publishing, 47–62.

[3] Sean Bechhofer, Iain Buchan, David De Roure, Paolo Missier, John Ainsworth, Jiten Bhagat, Philip Couch, Don Cruickshank, Mark Delderfield, Ian Dunlop, Matthew Gamble, Danius Michaelides, Stuart Owen, David Newman, Shoaib Sufi, and Carole Goble. 2013. Why linked data is not enough for scientists. *Future Generation Computer Systems* 29, 2 (Feb. 2013), 599–611. https://doi.org/10.1016/j.future.2011.08.004

[4] Reginald Cushing, Onno Valkering, Adam Belloum, and Cees de Laat. 2019. Towards a New Paradigm for Programming Scientific Workflows. In *2019 15th International Conference on eScience (eScience)*. 604–608. https://doi.org/10.1109/eScience.2019.00083

[5] Paul N. Edwards. 2010. *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. The MIT Press.

[6] Eslam Elnikety, Aastha Mehta, Anjo Vahldiek-Oberwagner, Deepak Garg, and Peter Druschel. 2016. Thoth: Comprehensive Policy Compliance in Data Retrieval Systems. In *Proceedings of the 25th USENIX Conference on Security Symposium (SEC'16)*. USENIX Association, Berkeley, CA, USA, 637–654. https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/elnikety

[7] Yehia Elrakaiby, Frédéric Cuppens, and Nora Cuppens-Boulahia. 2012. Formal enforcement and management of obligation policies. *Data & Knowledge Engineering* 71, 1 (Jan. 2012), 127–147. https://doi.org/10.1016/j.datak.2011.09.001

[8] Peter S. Hall and Andrew Morris. 2017. Chapter 15 - Predictive Analytics and Population Health. In *Key Advances in Clinical Informatics*, Aziz Sheikh, Kathrin M. Cresswell, Adam Wright, and David W. Bates (Eds.). Academic Press, 217 – 225. https://doi.org/10.1016/B978-0-12-809523-2.00015-7

[9] L. Hutton and T. Henderson. 2018. Toward Reproducibility in Online Social Network Research. *IEEE Transactions on Emerging Topics in Computing* 6, 1 (Jan. 2018), 156–167. https://doi.org/10.1109/TETC.2015.2458574

[10] Håvard D. Johansen, Eleanor Birrell, Robbert van Renesse, Fred B. Schneider, Magnus Stenhaug, and Dag Johansen. 2015. Enforcing Privacy Policies with Meta-Code. In *Proceedings of the 6th Asia-Pacific Workshop on Systems (APSys '15)*. ACM Press, Tokyo, Japan, 1–7. https://doi.org/10.1145/2797022.2797040

[11] Günter Karjoth, Matthias Schunter, and Michael Waidner. 2002. Platform for Enterprise Privacy Practices: Privacy-Enabled Management of Customer Data. In *Privacy Enhancing Technologies (Lecture Notes in Computer Science)*. Springer, Berlin,

Heidelberg, 69–84. https://doi.org/10.1007/3-540-36467-6_6

[12] Janne Lahtiranta, Sami Hyrynsalmi, and Jani Koskinen. 2017. The False Prometheus: Customer Choice, Smart Devices, and Trust. *SIGCAS Comput. Soc.* 47, 3 (Sept. 2017), 86–97. https://doi.org/10.1145/3144592.3144601

[13] Jonathan A. Obar and Anne Oeldorf-Hirsch. 2020. The biggest lie on the Internet: ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society* 23, 1 (Jan. 2020), 128–147. https://doi.org/10.1080/1369118X.2018.1486870

[14] Thomas F. J.-M. Pasquier, Jatinder Singh, David Eyers, and Jean Bacon. 2017. CamFlow: Managed Data-sharing for Cloud Services. *IEEE Transactions on Cloud Computing* 5, 3 (July 2017), 472–484. https://doi.org/10.1109/TCC.2015.2489211 arXiv: 1506.04391.

[15] Stephen Pavis and Andrew D Morris. 2015. Unleashing the power of administrative health data: the Scottish model. *Public Health Res Pract* 25, 4 (2015), e2541541.

[16] Livio Robaldo and Xin Sun. 2017. Reified Input/Output logic: Combining Input/Output logic and Reification to represent norms coming from existing legislation. *Journal of Logic and Computation* 27, 8 (Dec. 2017), 2471–2503. https://doi.org/10.1093/logcom/exx009

[17] David Robertson, Fausto Giunchiglia, Stephen Pavis, Ettore Turra, Gabor Bella, Elizabeth Elliot, Andrew Morris, Malcolm Atkinson, Gordon McAllister, Areti Manataki, Petros Papapanagiotou, and Mark Parsons. 2016. Healthcare data safe havens: towards a logical architecture and experiment automation. *The Journal of Engineering* 2016, 11 (Oct. 2016), 431–440. https://doi.org/10.1049/joe.2016.0170

[18] Eric J Topol. 2015. *The patient will see you now: the future of medicine is in your hands*. Basic Books.

[19] Benjamin E. Ujcich, Adam Bates, and William H. Sanders. 2018. A Provenance Model for the European Union General Data Protection Regulation. In *Provenance and Annotation of Data and Processes (Lecture Notes in Computer Science)*, Khalid Belhajjame, Ashish Gehani, and Pinar Alper (Eds.). Springer International Publishing, 45–57. https://doi.org/10.1007/978-3-319-98379-0_4

[20] Yang Xiao, Ning Zhang, Jin Li, Wenjing Lou, and Y. Thomas Hou. 2020. PrivacyGuard: Enforcing Private Data Usage Control with Blockchain and Attested Off-Chain Contract Execution. In *Computer Security – ESORICS 2020*. Springer, Cham, 610–629. https://doi.org/10.1007/978-3-030-59013-0_30

[21] Rui Zhao and Malcolm Atkinson. 2019. Towards a Computer-Interpretable Actionable Formal Model to Encode Data Governance Rules. In *2019 15th International Conference on eScience (eScience)*. 594–603. https://doi.org/10.1109/eScience.2019.00082