Position Statement: Empowering Researchers in Times of Crisis

Sophie Stalla-Bourdillon, Kenesa Ahmad, Elena Elkina,
Alexsis Wintour, Claire O'Hanlon, Steven Davenport, Daniel Wein

As the COVID-19 crisis deepens, so does the need for robust, data-driven research. Leading modelling methods, such as machine learning algorithms, are built to learn from the past: historical data is systematically analyzed, labelled and clustered to discover new correlations and patterns. Although such data use increases the likelihood of valuable inferences when applied to representative samples, these models also increase the likelihood of data misuse and harm. We believe that current research efforts require accelerated access to data and strong data governance, particularly when involving access to individual-level data.

Building a scalable and replicable research solution to achieve those goals requires an institutional approach; i.e., a set of rules and personas created to promote good data governance within data science environments.[1] We suggest that establishing trusted intermediaries in the form of integrated research platforms can help solve data discovery, interoperability and governance challenges. We define "integrated research platform" as an environment that embeds organizational and technical controls to ensure data protection and oversight of the data users' activities.

---

1. Trusted intermediaries help solve data discovery and interoperability challenges

Researchers often operate within environments that are technologically precarious and driven by long and duplicative processes. Generally, researchers initiating a new research project must start from scratch, with a pre-investigation phase to identify relevant data sources. Data discovery and access are rarely straightforward, particularly when the research requires combining data sources owned by different data providers. First, data curation standards can vary across organizations. Second, researchers must follow several compliance and ethical pathways to satisfy data providers' requirements, as well as requirements set by their originating institutions. Multi-party data sharing, especially involving health and behavior data, usually involves complex extract, transform and load processes (i.e., processes by which data is copied from one or more sources and shipped to a destination system), which can take several months, if not years.

Position Statement: Empowering Researchers in Times of Crisis

Sophie Stalla-Bourdillon, Kenesa Ahmad, Elena Elkina,
Alexsis Wintour, Claire O'Hanlon, Steven Davenport, Daniel Wein

We believe that it is possible to address these challenges by building integrated research platforms that are trusted by both data providers and data users. The COVID Alliance Research Platform ("Platform") is the product of a non-profit initiative aimed at providing researchers with standardized access to geolocation, public health and medical datasets in a cloud-computing environment. The Platform leverages privacy-preserving technologies and supports collaboration between data users and data governance teams, while standardizing discovery of and access to valuable data pipelines. Using privacy-preserving technologies while also designing for a high-quality user experience is a delicate exercise and requires a series of testing and iterations with different solution providers. The COVID Alliance recently opened the Platform to external research partners as well as investigative journalists, who need access to quality data to assess COVID-19 responses and the evolution of the pandemic. Both individuals and institutions are welcome to apply to the Platform's online portal.

2. Trusted intermediaries help reduce tensions between privacy and the need for data access

Standardizing data access through technology is a double-edged sword. If not combined with strict data governance controls guided by legal and ethical standards, the likelihood of misuse and harm increases. However, implementing data governance controls can slow access to data and drastically reduce data utility. Privacy protection is thus in constant tension with data access.

Typically, in research projects, assessments of compliance, privacy, ethics and domain expertise are conducted in silos and reconciliation of approaches is difficult. Appropriately setting the privacy versus utility trade-off[2] is not a given. We propose a two-pronged strategy for trusted intermediaries to reduce tensions between privacy and the need for data access: 1) grow a multiskilled and proactive data governance team through direct dialogue with privacy and ethics experts to streamline workflows and facilitate assessment integration; and 2) organize data science

Position Statement: Empowering Researchers in Times of Crisis

Sophie Stalla-Bourdillon, Kenesa Ahmad, Elena Elkina,
Alexsis Wintour, Claire O'Hanlon, Steven Davenport, Daniel Wein

activities by projects, establishing common denominators across projects while refining privacy-preserving controls by project.

One key control to implement within data science environments is data de-identification or anonymization. This is true even if, in a machine learning age, privacy attacks go well beyond re-identification of individuals whose data has been used during the training phase.[3] Yet, many universities or research institutions are poorly equipped to implement such a control. A common misunderstanding is that if data has been de-identified once, stringent ethical review is not needed again.

Trusted intermediaries, such as integrated research platforms, offer the opportunity to build de-identification expertise. The trusted intermediary keeps de-identified or anonymized data within the remit of data governance teams and regularly monitors the data environment, which is key as re-identification risks evolve over time. The Platform combines a variety of de-identification and anonymization techniques to increase the range of options available to the COVID Alliance's data governance team. These techniques are supplemented by other supporting organizational and technical data governance controls. This risk-based approach[4] has been applied to geolocation data to develop tools to support clear understanding of and rapid response to risky mass gathering events in the United States, such as the Sturgis Motorcycle Rally and Lake of the Ozarks BikeFest.[5] The COVID Alliance's work has started to inform public health response, such as measures taken by the St. Louis City Department of Health.[6]

Conclusion

Empowering researchers in times of crises requires both accelerating data discovery and access and establishing controlled environments for data operations. We suggest that trusted intermediaries, such as the COVID Alliance Research Platform, can help solve data discovery, interoperability and governance challenges that research organizations currently face.

Position Statement:
Empowering Researchers in Times of Crisis

Sophie Stalla-Bourdillon, Kenesa Ahmad, Elena Elkina,
Alexsis Wintour, Claire O'Hanlon, Steven Davenport, Daniel Wein

Citations

[1] See the Open Data Institute, Workstream on data institutions, https://theodi.org/project/rd-data-institutions/

[2] Privacy is not black and white, and data protection principles are usually conceived as goals. Finding a middle ground is key as long as minimum requirements are met. This is commonly referred to as the "privacy vs. utility trade-off."

[3] See e.g, S. Stalla-Bourdillon et al., Warning Signs - The Future of Privacy and Security in the Age of Machine Learning, FPF and Immuta Whitepaper, https://www.immuta.com/warning-signs-the-future-of-privacy-and-security-in-the-age-of-machine-learning/ (2019).

[4] There is growing consensus that a risk-based approach to de-identification & anonymization makes sense & is compatible with privacy & data protection frameworks. See e.g., Sophie Stalla-Bourdillon and Alfred Rossi, 'Aggregation, Synthesisation and Anonymisation: A Call for a Risk-Based Assessment of Anonymisation Approaches,' Computer Privacy and Data Protection 2020 Conference Book (Hart Publishing), *forthcoming*.

[5] Steven Davenport et al., 'Geolocation data suggest Lake of the Ozarks' BikeFest may be a regional coronavirus 'superspreader.' These 23 counties should be on alert' (2020).

[6] KMOV.com, 'Lake of the Ozarks motorcycle rally attendees encouraged to get tested for Covid-19' (2020).