DataTags

Michael Bar-Sinai, Latanya Sweeney, Mercè Crosas

Institute of Quantitative Social Science at Harvard University

Privacy Tools for Sharing Research Data

A National Science Foundation Secure and Trustworthy Cyberspace Project

INTRODUCTION

Datasets used in social science research are often subject to legal and human subjects protections. Not only do laws and regulations require such protection, but also, without promises of protection, people may not share data with researchers. On the other hand, "good science" practices encourage researchers to share data to assure their results are reproducible and credible. Funding agencies and publications increasingly require data sharing too. Sharing data while maintaining protections is usually left to the social science researcher to do with little or no guidance or assistance.

It is no easy feat. There are about 2187 privacy laws at the state and federal levels in the United States [1]. Additionally, some data sets are collected or disseminated under binding contracts, data use agreements, data sharing restrictions etc. Technologically, there is an ever-growing set of solutions to protect data – but people outside of the data security community may not know about them and their applicability to any legal setting is not clear.

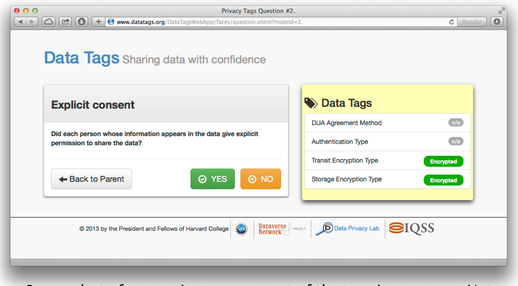
The DataTags project aims to help social scientists share their data widely with necessary protections. This is done by means of interactive computation, where the researcher and the system traverse a decision graph, creating a machine-actionable data handling policy as they go. The system then makes guarantees that releases of the data adhere to the associated policy.

OBJECTIVES

Create and maintain a user-friendly system that allows researchers to share data with confidence, knowing they comply with the laws and regulations governing shared datasets.

We plan to achieve the above by the following efforts:

1. Describe the space of possible data policies using orthogonal dimensions, allowing an efficient and unambiguous description of each policy.
2. Harmonize American jurisprudence into a single decision-graph for making decisions about data sharing policies applicable to a given dataset.
3. Create an automated interview for composing data policies, such that the resulting policy complies with the harmonized laws and regulations (initially assuming the researcher's answers correctly described the dataset).
4. Create a set of "DataTags" – fully specified data policies (defined in Describing a Tag Space), that are the only possible results of a tagging process.
5. Create a formal language for describing the data policies space and the harmonized decision-graph, complete with a runtime engine and inspection tools.
6. Create an inviting, user-friendly web-based automated interview system to allow researchers to tag their data sets, as part of the Dataverse system.



Screenshot of a question screen, part of the tagging process. Note the current data tags on the right, allowing the user to see what was achieved so far in the tagging process.

DataTags and Harm Levels

Harvard Research Data Security Policy[2] describes a 5-level scale for researchers to handle research data. We extend this to a 6-level scale for specifying data policies regarding security and privacy of data. The scale is based on the level of harm malicious use of the data may cause. The columns represent some of the dimensions of the data policy space.

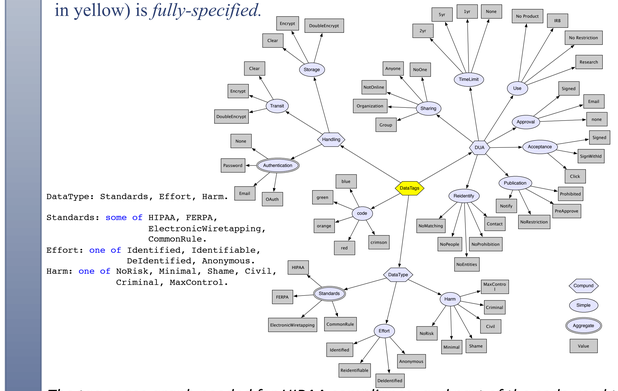
Harm Level	DUA Agreement Method	Authentication	Transit	Storage
No Risk	Implicit	None	Clear	Clear
Data is non-confidential information that can be stored and shared freely				
Minimal	Implicit	Email/OAuth	Clear	Clear
May have individually identifiable information but disclosure would not cause material harm				
Shame	Click Through	Password	Encrypted	Encrypted
May have individually identifiable information that if disclosed could be expected to damage a person's reputation or cause embarrassment				
Civil Penalties	Signed	Password	Encrypted	Encrypted
May have individually identifiable information that includes Social Security numbers, financial information, medical records, and other individually identifiable information				
Criminal Penalties	Signed	Two-Factor	Encrypted	Encrypted
May have individually identifiable information that could cause significant harm to an individual if exposed, including serious risk of criminal liability, psychological harm, loss of insurability or employability, or significant social harm				
Maximum Control	Signed	Two-Factor	Double Encrypted	Double Encrypted
Defined as such, or may be life-threatening (e.g. interviews with identifiable gang members).				

Describing a Tag Space

In order to define the tags and their possible values, we are developing a formal language, designed to allow legal experts with little or no programming experience to write interviews. This will enable frequent updates to the system, a fundamental requirement since laws governing research data may change. Below is the full tag space needed for HIPAA compliance, and part of the code used to create it.

Representing the tag space as a graph allows us to reason about it using Graph Theory. Under these terms, creating DataTags to represent a data policy translates to selecting a sub-graph from the tag space graph. A *single* node *n* is said to be *fully-specified in sub-graph S*, if S contains an edge from *n* to one of its leafs. A *Compound* node *c* is said to be *fully-specified in sub-graph S* if all its single and compound child nodes are fully specified in sub-graph S.

A tagging process has to yield a sub-graph in which the root node (shown in yellow) is *fully-specified.*



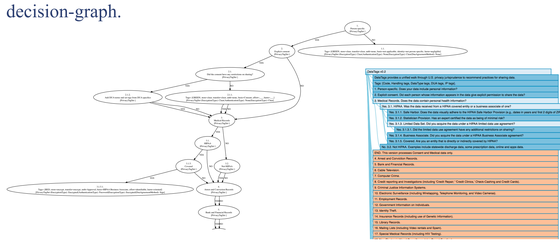
The tag space graph needed for HIPAA compliance, and part of the code used to describe it. Base graph for the diagram was created by our language interpreter.

Harmonized Decision-Graph

Harmonized decision-graphs are the programs interactively executed by the runtime and the researcher. The language we develop to create them will support tagging statements, suggested wording for questions, sub-routines and more. As we realize harmonized decision-graphs take a long time to create and verify legally, we plan to support a special TODO type, such that partially implemented harmonized decision-graphs can be executed and reasoned about.

Part of the tooling effort is creating useful views of the harmonized decision-graph and its sub-parts. Below are two views of a harmonized decision-graph – one interactive (based on HTML5) and another static (based on Graphviz). The latter was automatically generated by our interpreter. Nodes show technical information as well as basic wording (actual wording presented to the researcher may be different).

We have already harmonized regulations related to IRBs, consent and HIPAA and made a summary flow chart of questions for an interview of a researcher. We have also had legal experts review our approach and all agreed it was sufficient, proper and prudent with respect to data sharing under HIPAA. The views below show parts of the HIPAA harmonized decision-graph.

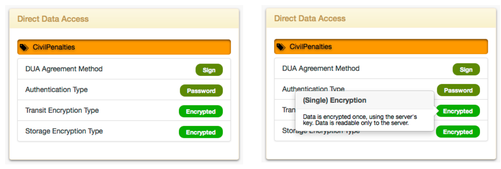


Two views of the same harmonized decision graph, computing HIPAA compliance

User Interface

Usability is a major challenge for DataTags to be successful. From the data publisher point of view, a data tagging process may be experienced as a daunting chore containing many unfamiliar terms, and carrying dire legal consequences if not done correctly. Thus, the interview process and its user interface will be designed to be inviting, non-intimidating and user-friendly. For example, whenever legal or technical terms are used, a layman explanation will be readily available.

As the length of the interview process depends on the answers, existing best practices for advancement display (such as progress bars or a check list) cannot be used. Being able to convey the progress made so far in a gratifying way, keeping the user engaged in the process is an open research question which we intend to study.



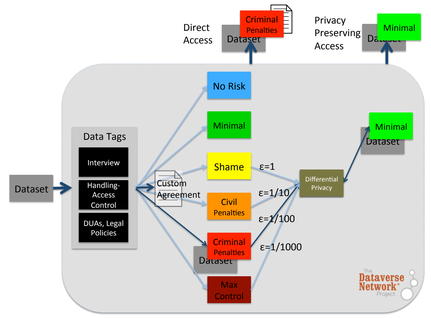
In order to make the tagging process approachable and non-intimidating, whenever a technical or a legal term is used, an explanation is readily available. Shown here is part of the final tagging page, and an explained technical term.

CONCLUSIONS

The DataTags project will allow researchers to publish their data, without breaching laws or regulations. Using a simple interview process, the system and researcher will generate a machine actionable data policy appropriate for a dataset – its "DataTags". This policy will later by used by systems like Dataverse to decide how the data should be made available, and to whom. The system will also be able to generate a customized DUA based on these tags – a task that is currently done manually, consuming a lot of time and resources.

The programming language for Tag Space and Harmonized decision-graph description, and the tools related to it, will be able to describe general harmonized decision-graphs, not just in the legal field. While easy to learn, the language relies on Graph Theory, a robust foundation that will allow various tools, including model checking and program/harmonized decision-graph validations.

We believe DataTags will dramatically improve the rate of data sharing among researchers, while maintaining legal compliance and at no cost to the researcher or her institution. As a result, we expect more data to be available for researchers, with fewer barriers of access.



Overview of a dataset ingest workflow in Dataverse, showing the role of the DataTags project in the process.

REFERENCES

[1] Sweeney L. Operationalizing American Jurisprudence for Data Sharing. White Paper. 2013

[2] http://www.security.harvard.edu/research-data-security-policy

ACKNOWLEDGEMENTS

Bob Gellman – validating the current harmonized decision-graph we have is HIPAA compliant.

CRCS Center for Research on Computation and Society at Harvard School of Engineering and Applied Sciences

BERKMAN CENTER FOR INTERNET & SOCIETY AT HARVARD UNIVERSITY

The Institute for Quantitative Social Science at Harvard University

DATA PRIVACY LAB