Poster: DataTags, Data Handling Policy Spaces and the Tags Language

Michael Bar-Sinai
Computer Science Dept.
Ben-Gurion University of the Negev
Be'er-Sheva, Israel
mbarsinai@iq.harvard.edu

Latanya Sweeney
Data Privacy Lab
Harvard University
Cambridge, MA
latanya@fas.harvard.edu

Mercè Crosas
Institute for Quantitative Social Science
Harvard University
Cambridge, MA
mcrosas@iq.harvard.edu

We propose an extensible, formal, machine-actionable model for describing and reasoning about dataset handling policies. Additionally, we present the Tags programming language and toolset, created for working with the proposed model. We use Tags to create the sample set of datatags proposed in a recent paper, and apply it to HIPAA. We also present some of Tags' tools, such as visualizers, development environment, and code analyzers.

Dataset handling policies set various aspects of dataset handling, such as access requirements, and required security features for transmission and storage. Modeling data handling policies as points in a multidimensional, ordinal space (called Data Handling Policy space, or DHP space, for short), allows for precise reasoning about their properties. We define binary operators for policy composition and for comparing policy strictness, allowing formal phrasing of propositions such as "this policy is stricter than that policy", "these two policies have different lenient aspects", and "dataset X can be stored in data repository R".

A set of datatags, as proposed in recent paper by the authors, can be understood as a list of carefully selected points in a DHP space, fully ordered by strictness. A datatags-compliant data repository is represented in a DHP space by the datatags it implements. Datasets are represented in DHP space by the policy required to handle them. Once both dataset and data repository are present in the same DHP space, it is easy to decide whether the dataset can be deposited in said data repository, and if so, under which tag.

But how can a researcher with no legal or data security expertise decide on the proper policy to handle a dataset?

To this end, Tags allows the creation of interactive, user-friendly questionnaires, modeled after the familiar metaphor of "an interview with an expert". In addition to questions, the questionnaire contains instructions for composing a data handling policy based a given DHP space. An execution environment for the questionnaires allows dataset depositors to go through a friendly, interactive interview where they describe the provenance of the dataset in question, and then receive a data handling policy fit for it. The interview is also available as a service on the web, so it can be integrated into the dataset depositing process of general-purpose data repositories, such as Dataverse. In case the data stewards use a file management system, they can go through the interview, receive the generated policy and implement it on their own, as the generated policy is human-readable. Moreover, the interview can be used before the data is collected, in the research design phase, to predict the required data handling policy should a given design be selected.

We present part of the growing Tags toolset, including visualizers for the questionnaire and the DHP space, refactorings, and query engine, capable of finding all answer sequences that will result in a given set of policies. The latter allows for automatic questionnaire validation, such as "this questionnaire never gets to a state where a dataset contains private and confidential data, and the resultant policy allows for non-encrypted data transmission".

Finally, we present a Tags questionnaire covering HIPAA, vetted by a legal expert. We conclude by proposing future directions for research, and possible applications for Tags and the DHP space model, e.g. for comparative legal studies.