Sweeney L, Crosas M, Bar-Sinai M. Sharing Sensitive Data with Confidence: The DataTags System. *Technology Science*. 2015101601. October 16, 2015. http://techscience.org/a/2015101601

Sharing Sensitive Data with Confidence: The Datatags System

Latanya Sweeney, Mercè Crosas, and Michael Bar-Sinai

Highlights

- We introduce datatags as a means of specifying security and access requirements for sensitive data.
- The datatags approach reduces the complexity of thousands of data-sharing regulations to a small number of tags.
- We show implementation details for medical and educational data and for research and corporate repositories.

Tag Type	Description	Security Features	Access Credentials
Blue	Public	Clear storage, Clear transmit	Open
Green	Controlled public	Clear storage, Clear transmit	Email- or OAuth Verified Registration
Yellow	Accountable	Clear storage, Encrypted transmit	Password, Registered, Approval, Click-through DUA
Orange	More accountable	Encrypted storage, Encrypted transmit	Password, Registered, Approval, Signed DUA
Red	Fully accountable	Encrypted storage, Encrypted transmit	Two-factor authentication, Approval, Signed DUA
Crimson	Maximally restricted	Multi-encrypted storage, Encrypted transmit	Two-factor authentication, Approval, Signed DUA

Definitions for each of six ordered Blue to Crimson sample datatags.

Abstract

Society generates data on a scale previously unimagined. Wide sharing of these data promises to improve personal health, lower healthcare costs, and provide a better quality of life. There is a tendency to want to share data freely. However, these same data often include sensitive information about people that could cause serious harms if shared widely. A

multitude of regulations, laws and best practices protect data that contain sensitive personal information. Government agencies, research labs, and corporations that share data, as well as review boards and privacy officers making data sharing decisions, are vigilant but uncertain. This uncertainty creates a tendency not to share data at all. Some data are more harmful than other data; sharing should not be an all-or-nothing choice. How do we share data in ways that ensure access is commensurate with risks of harm?

Results summary: We introduce the notion of datatags as a means of identifying handling and access requirements for a file. Handling includes security features, such as the use of encryption in the storage and transmission of files. Access requirements for those receiving files include providing credentials and agreeing to terms of use. A datatags repository shares data having varying levels of sensitivity by assigning tags that encode varying levels of handling and sharing restrictions. Although there are thousands of data sharing laws and regulations, and numerous ways to specify security for any given file, the datatags approach reduces this complexity to a few well-defined choices. A datatags-compliant repository provably complies with the policies associated with the designated tag to make sure promised and legally necessary handling requirements are met. There are many possible ways to construct a datatags repository, and which one is best depends on use. We introduce a model set of six tags to support options from data having no risk to data requiring maximum protection. We use the set of model datatags to present exemplar architectures for research labs, research repositories, government repositories, multinational corporations, and institutional review boards. We show implementation details for medical data, provide an interview system for tagging medical and educational data, and demonstrate how to construct a global research repository. Finally, decision makers and scholars can use a datatags repository, even without access to data, to study, compare, and analyze data sharing regimes.

Introduction

Researchers, companies, universities, and governments often store and share sensitive data and need to do so in a way that respects a multitude of legal commitments, best practices, local policies, and ethical promises. Here are some examples.

The United States federal government wants to spur scientific breakthroughs and economic advances by making research results more available to innovators. Under a White House directive, each federal agency with a research and development budget in excess of $100 million must develop its own public access policies to ensure that the public can read, download, and analyze in digital form published works and data arising from federally funded research [1]. Some government funders are developing repositories for specific disciplinary research, such as genomics and environmental health. Other funders and academic publishers push this responsibility onto the researchers themselves.

- Adam leads a medical research team at a large university hospital complex. His team receives millions of dollars on a multitude of research projects based on data drawn from the hospital's medical information system. Research results from the team have led to groundbreaking improvements in clinical care. Adam wants to share versions of his datasets for research validity and reproducibility in keeping with the goals established by his funders and publishers, but he wants to do so in such a way that he has some knowledge of who receives copies of the data. Additionally, Adam's research team works with almost 100 researchers at other institutions on collaborative projects, and there is no centrally accessible archive for data and interim research files. In his lab, research files tend to be stored on the individual computers of the researchers doing the work, with no centralized archive. Last year, one of his researchers left abruptly and Adam remains unable to locate the files that formed the basis of published results. Adam wonders whether there is a way to technologically manage his team's data sharing needs.

A huge quantity of digital research data exists solely as files on the computers of individual researchers, even in cases where the impact of the research based on the data has been dramatic. The scientific community started preserving data so that it would not be lost forever. Data needs to be identified, located, assessed, acquired, processed, preserved, and shared to extend useful life and protect against computer failure [2]. Over the past 7 years, different Web archives have emerged for sharing data studies. Most are discipline-specific, with variations in the quality and nature of management (e.g., geophysics [3], biosciences [4], earth sciences [5], clinical trials [6], and medicine [7]). None currently offers the full spectrum of sharing that Adam needs.

- Betty is a social science researcher who often works alone. She received federal funding to collect and analyze sensitive data from social media about sexual and other preferences that she acquired with the consent of participants. Her goal is to learn more about how friendships form and change over time. In accepting the grant, she agreed to share a version of the information with other researchers for research validity and to reduce data collection costs for future studies.

Leading research archives are capable of publishing, citing and sharing research data, and many tend to be powered by the same software. For example, the Dataverse [8] [9] [10] [11], an open source software package founded at Harvard in 2006, has been installed and used in a multitude of universities around the world as a research data sharing platform for those institutions (e.g., a consortium of universities in the Netherlands, Fudan University in China, and the British Columbia Research Data Services in Canada). Harvard's Dataverse hosts more than 59,000 data sets from many disciplines, including the world's largest collection of social science research data [12]. The Harvard Dataverse currently stores non-sensitive, publicly available data only. In general, anyone can access any file. Betty's data are sensitive and cannot be stored there. However, she could place a file there to notify others to contact her

for the data. Betty wonders whether Dataverse might evolve to store and share sensitive data directly, and whether her university would allow data sharing through the Dataverse.

In the United States, institutional review boards (IRBs) act as independent third parties and decide whether researchers can access personal data and, if so, what forms of the data may be shared [13]. Notwithstanding the fruitful results that may be realized from all areas of research, disasters in research have occurred. Promulgated in 1981, Title 45, Code of Federal Regulations, Part 46, Protection of Human Subjects (45 CFR 46), requires IRB approval for all federally funded research involving human participants. Research institutions receiving substantial federal research dollars must provide an IRB panel to review and approve all research conducted at the institution involving human subjects.

- Charles leads the IRB at his university. He worries that his committee may not always use the best measures when it decides about data sharing. As members rotate on and off the committee, research standards change, new kinds of protocols appear, and knowledge of new risks emerges. Thus, committee discussions evolve, sometimes in a circular fashion, and judgments may oscillate. A committee that learns and changes is good, but Charles wonders whether change could be made more productively. Because of committee decisions, some researchers may be burdened with unnecessary security concerns while others are not taking sufficient care. Can technology help the committee systematically organize its decisions and impose consistent requirements?

- Diane is the chief privacy officer of a gigantic multinational corporation that prides itself on the many ways it uses data to drive innovation and productivity within its various project groups. Today's technology allows the company to share information globally among its employees so that the best of its analysts can process the most relevant information available to the company, no matter where in the world the analysts and data reside. The seamless flow fosters a belief within the company that it can freely share data held internally, but Diane realizes it is not that simple. Different data have different legal, regulatory, and contractual restrictions. Data collected under one agreement may need to be treated differently than data collected under another. Data collected in one country may not be able to be seamlessly transmitted to analysts in another country, even though the recipient is an employee at the same company. Luckily, the company has not had any problems, but Diane attributes that to the fact that most data remains within the smaller units in which it originates. However, increasingly data are being shared widely, beyond the context of the smaller units, and she is concerned that serious problems will emerge because employees sharing the data with other employees may not realize all the legal and corporate ramifications. She and her company want to honor their obligations while promoting data sharing within the company. Can technology help?

All these people need to store and share data in a manner that respects legal commitments and ethical promises. They need a way to associate requirements with data and a way to make sure all stores and transmissions of the data and all access requests made on the data provably respect those requirements. This sounds simple, but doing so may not be easy.

Background

The need for computers to store and retrieve sensitive information is not new. Databases already use passwords and other credentials to store and retrieve sensitive values such as a name or birthdate. Of course, in the examples described previously, people share files and not data values. Not all files lend themselves to efficient database storage, especially as the size of the files increases. Even if a file could be loaded into a database, there would be substantial overhead if retrieval required database exporting beforehand. Database software alone is not an answer.

Multiuser file systems already manage file reading and writing permissions for groups of users on the system. This is often called mandatory access control. In some of our examples, it may be possible to use a mandatory access file system, as a partial solution. Our hypothetical medical research team leader, Adam, might use a Unix file system to help with internal file access. Data files would be stored in directories and access controlled by group permissions set by the system administrator. A practice might be to include a text file named README in each directory. The contents of the README file would describe notice and handling requirements. An honor system asks users to read the README files before copying or viewing the data to be sure proper care and access occurs. Of course, maintaining hundreds of user accounts takes ongoing resources to set up accounts, reset passwords, and set and change permissions. Additionally, some files are highly sensitive and should not be stored on the file system unencrypted. Encryption key management across hundreds of non-local users can be onerous.

Role-based access systems [14] might be useful within organizations, at least in part. For example, our hypothetical chief privacy officer, Diane, could use the security credentials her organization already maintains for employees. Within her organization, employees have different roles, and those roles already have associated access permissions that allow employees access to certain physical spaces and computer systems. These could also apply to categories of computer files in the globally networked cloud storage. By virtue of their assigned roles, employees could have access only to specific groups of files. As described earlier, a README file within each file directory could describe notice and handling requirements for the files in the directory. A concern is that different and unexpected employees increasingly request access to different data files, requiring overhead to alter permissions. This approach works best if the number of roles is few or hierarchical. In addition, as stated previously, sensitive files will be encrypted, and key management adds significantly to management overhead.

A critical limitation in using mandatory and role-based access with our examples is the diversity, number, and ever-changing nature of data requesters. Computer scientists use policy specification languages for environments in which resources and users are not pre-determined, the environment is constantly changing, and the number of requesters may be very large [15] [16]. This approach, adapted to our cases, employs declarative policies that allow access based on attributes of data requesters and of files. These policies are in a machine-understandable form. The starting point is an ontology of requesters and terms used to assess permissions. A policy is a general statement or rule about data sharing. Below are some examples of policy statements related to sharing medical data.

1. If all the participants whose data appears in the file consented to sharing, allow any requester a copy of the file.

2. If there is no personal information in the file, allow any requester a copy of the file.

3. If the data includes AIDS or HIV information about participants, only allow requesters having medical information clearance a copy of the file.

4. Do not share any file if prohibited by a law or regulation.

5. If the rules conflict, disallow the request.

An attractive feature of this approach is that it allows high level reasoning over policies and permissions. For example, our hypothetical leader of an IRB, Charles, could codify the IRB's principles and then later inquire whether a given use is consistent with those principles. Consider the example policies above. Rule 1 respects human autonomy in decision-making. Rule 2 limits decision making to files that actually contain personal information. Rule 3 protects against widespread sharing of sensitive medical information. Rule 4 ensures legal compliance. Finally, Rule 5 resolves conflicts by conservatively disallowing the request.

A challenge to using this approach involves unforeseen permissions that may inadvertently be allowed (or not) when policies combine and when terms appearing in policy statements are nuanced. For example, suppose there is a data file containing counts of the number of AIDS cases by state in the United States. An arbitrary public user requests a copy of the file. Rule 2 allows the sharing. Rule 3 disallows the sharing. These rules conflict, in part, because of fine distinctions between personal information and information about participants. Rule 5 resolves the conflict by not allowing the file to be shared, even though there is no good reason for the file not to be shared.

Suppose we swap the default in Rule 5. Instead of disallowing a request when a conflict occurs, the rule allows the request when a conflict occurs. Having the inverse rule now yields the desired result in the previous example. However, other problems may still exist. Consider a data file that contains geo-locations from a mobile app used to help patients with complex care needs maintain schedules for costly medical appointments. Only the geo-locations from

app users who granted permission to share their geo-locations appear in the data. The data also includes the type of medical specialist visited. Some of the people visited AIDS clinics. Rule 1 would allow the file to be shared widely because of its blanket consideration of consent. Rule 3 would not. Rule 4 recognizes that it is against the law to share the information widely. However, Rule 5, now reversed, would allow the sharing anyway. These rules conflict, in part, because of a lack of precedence as to whether personal consent should override a law or whether a law should override personal consent. Together, these examples show how high level rules can combine to yield unforeseen and undesirable consequences.

If one could develop correct and well-formed rules, then technology could enforce them. A middleware technology named iRODS supports policy based control over access to and use of data [17]. Its features include: data and metadata search, and management, control, and tracking of all data access and manipulation, workflows, annotations, and subscriptions. It has a rule engine that applies user-defined policies and rules to data to automatically enforce governance and management policies, enabling large-scale collections with complex requirements. The critical requirement to make iRODS operational in our use cases is defining needed rules and policies.

Methods

What is needed is a repository capable of handling sensitive and non-sensitive files in accordance with different security requirements. Computer security concerns that usually arise when storing and sharing files in a repository include the visibility of file contents during transit and in storage, and the credentials necessary to access each file.

For example, files may transmit over a network in clear text, in which case an eavesdropper on the communication line can view the contents. Alternatively, the contents or transmission may be encrypted, thwarting the efforts of the casual eavesdropper.

Similarly, files stored on disk can be in a native application format, in which case someone who has access to the file system can inspect a file's contents at will. We say the file is stored in the clear. Alternatively, a file may be stored on an encrypted drive or in an encrypted form, thereby limiting access to its contents to those having access to the encryption key.

Of course, there are other security features to consider. We do not mean to imply that the only security features to consider are whether file storage and transmission occur in the clear or are encrypted. We present them in this writing as a good set of features to consider.

There are numerous ways to authenticate a person's credentials before providing a copy of a file to the person. In this context, authentication means identifying a means of communicating with the person, and it may include the ability to show that the means of communication works. Here are some examples: an email address verified by a passcode sent in an email message; social media accounts verified by social media login credentials;

mobile phone numbers verified by a passcode sent in a text message; a photocopy of a driver's license; and challenge questions verified from public record information.

We introduce the notion of a *datatags repository* as one that stores and shares data files in accordance with different levels of security and access requirements. We focus on security requirements because they identify necessary physical machine and network requirements. Our definition appears below.

A *datatags repository* is a repository of files held for data sharing that satisfies the following conditions:

1. A datatag is a set of security features and access requirements for file handling. A datatags repository has a finite, partially ordered set of datatags, where the strictness and strength of datatags' security features and access requirements dictate the ordering. A repository must have more than one datatag.

2. All files in the repository must have a datatag, and each file in the repository has one and only one datatag. A file may optionally have additional handling requirements, such as an audit trail log or an expiration date. A file may optionally require additional terms for a data use agreement or additional terms of access by a recipient of the file from the repository. A file may have attributes that further describe it for reporting purposes. None of the optional requirements may weaken or replace the security requirements for the file's assigned datatag, and none may adjust a datatag's security requirements to be the same as another datatag or stronger than a more restrictive datatag.

3. A recipient who receives a file from the repository must satisfy the file's associated access requirements, produce sufficient credentials as requested, and agree to any terms of use required to acquire a copy of the file.

4. Technological guarantees exist that the requirements in 1 and 2 are satisfied for all files in the repository and for all accesses to those files from the repository. This imposes auditing obligations on transactions in the repository.

Security features and access credentials are independent components of a datatag and at least one must be ordered to satisfy the first condition. For example, if A and B are security features in that order and X and Y are unordered access credentials, then possible datatags are: AX, AY, BX, and BY. It is not true that AX < AY, nor is it true that AX > AY. Instead, the datatags would be: {AX, AY} < {BX, BY}. However, if X and Y are ordered, the ordered tags become AX < AY < BX < BY.

Consider a data-sharing repository that holds only non-sensitive, publicly available data. Anyone can visit the website, click on a link, and receive the data file immediately. Transmissions over the HTTP protocol are in the clear. Behind the scenes, file storage is also

in the clear. As described, this is not a datags repository because there are no options for handling files in any other way. Similarly, consider a highly secure silo that encrypts all transmissions, encrypts all files, and limits access to employees who have identification cards with the appropriate machine-readable embedded code. This too is not a datatags repository because there is no variation in security requirements.

Suppose the public file repository described above handled sensitive data by placing a file online that describes data. The file does not actually contain the data. Instead, the file includes contact information and explains the procedure potential recipients must follow to acquire the data. This variant provides two different access requirements, one for publicly available data and one for sensitive data. Therefore, this would qualify as a datatags repository having just two levels, sensitive or not. Notice that access to sensitive data in this repository is not online. Automation is preferred but is not a requirement.

Of course, we envision a datatags repository with several levels. Table 1 describes a model datatags repository having six color-coded levels that we introduce for demonstration purposes. As the levels increase, so do the transmission, storage, and access requirements. For reference purposes, we call these six levels the "model datatags." The datatags differ in security features by whether encryption is required during storage or transmission. Encryption during transmission seeks to thwart communication eavesdroppers. Encrypted storage seeks to make sure that of those who may gain internal access to files, only those who also have access to the encryption key can view file contents. The strongest is multiparty encryption, which means that even the system administrator may not know the contents of a file because the administrator may hold only one of the multiple encryption keys necessary to gain access to content. For efficiency, we refer to this as multi-encryption or multi-encrypted storage in this writing.

Table 1 also shows a partial ordering of access credentials associated with the model set of datatags. As the levels increase, so does the confidence with which a recipient of the file can be identified and contacted. At the lowest level, the Blue datatag requires no access credentials. The Green datatag requires that a requester's email address be verified, presumably by sending a link in an email message to which the requester must respond. Alternatively, the Green datatag may use credentials verified from a social media account. From the Yellow datatag onwards, the requester must submit an application and receives access only after approval. As the levels increase, so do the required steps by which the requester accepts the terms of a data use agreement. The Blue and Green datatags may have a minimal or no data use agreement. The Yellow datatag has a data use agreement that the requester accepts using a click-through agreement. From the Orange datatag onwards, the data use agreement must be signed. The Red and Crimson datatags requires two-factor authorization, which could require verifying both the requester's email address and mobile phone number. As stated earlier, the definitions of these model datatags are for demonstration and discussion purposes. They are not the only definitions datatags may assume.

Tag Type	Description	Security Features	Access Credentials
Blue	Public	Clear storage, Clear transmit	Open
Green	Controlled public	Clear storage, Clear transmit	Email- or OAuth Verified Registration
Yellow	Accountable	Clear storage, Encrypted transmit	Password, Registered, Approval, Click-through DUA
Orange	More accountable	Encrypted storage, Encrypted transmit	Password, Registered, Approval, Signed DUA
Red	Fully accountable	Encrypted storage, Encrypted transmit	Two-factor authentication, Approval, Signed DUA
Crimson	Maximally restricted	Multi-encrypted storage, Encrypted transmit	Two-factor authentication, Approval, Signed DUA

Table 1. Definitions for each of six ordered Blue to Crimson model datatags that are introduced for demonstration purposes. A data file is tagged with one level based on its handling needs. Its assigned level dictates security features, which in these models concern the use of encryption for storage and transmission. An assigned level also specifies access mechanics. In the examples shown, the strength of the access credentials increases as the security level increases. The lowest level, 1, is colored blue (on top) and has the least restrictions to access the data. The levels proceed with more stringent storage, transmission, and access requirements as the level increases from level 2 green, level 3 yellow, level 4 orange, and level 5 red, to level 6 crimson, the highest and most restrictive level. Multi-encrypted storage refers to a use of multi-party encryption.

Support for this approach comes from a real-world example. Years earlier, Harvard University's Provost's Office designed and adopted an Information Security Policy to standardize the handling of research data across the University. Their security policy has five security grades that prescribe what kind of data fits into each grade and, based on its grade, how those data are stored and transmitted [18]. Here is a summary of the Harvard Security grades and how we relate them to the model datatags.

Harvard Grade 1 data consist of non-confidential and non-personal information that can be stored and shared freely [18]. Harvard's examples include non-personal research data that has been openly published, openly available statistics, public use files, publicly available de-identified personal information, and research data that has been de-identified in accordance with applicable rules. Consistent with Harvard's security standards, the handling of Grade 1 data maps to Blue, for non-personal data. We additionally map Grade 1 data to Green if it contains personal data. We make this distinction because a re-identification vulnerability

may become learned later (e.g., [19]), in which case having verified email addresses of recipients of the files and a data use agreement allows some minimal recourse upon discovery of a re-identification vulnerability.

Harvard Grade 2 data is information that Harvard considers "confidential but the disclosure of which would not cause material harm" [18]. Harvard's examples include unpublished research work and intellectual property not otherwise classified at a higher level. Consistent with Harvard's security standards, the handling of Grade 2 data maps to Yellow.

Harvard Grade 3 data is information that Harvard believes could pose "risk of material harm" to individuals if disclosed [18]. Harvard's examples include information protected by the Family Educational Rights and Privacy Act (FERPA) that is not Grade 1, such as non-directory student information and directory information about students who have requested a FERPA block. Other examples include school ID numbers associated with names that could identify students, identified personnel records, and individuals' identified personal financial records. Consistent with Harvard's security standards, the handling of Grade 3 data maps to Orange.

Harvard Grade 4 data is information that Harvard believes would likely "cause serious harm to individuals if disclosed," including "risk of serious social, psychological, reputational, financial, legal or other harm to an individual or group" [18]. Harvard's examples include information that associates an individual's name together with any of the following data about that individual: Social Security number, bank or other financial account numbers, credit or debit card numbers, driver's license number, passport number, other government issued identification numbers, biometric data, or health and medical information. Consistent with Harvard's security standards, the handling of Grade 4 data maps to Red.

Harvard Grade 5 data is information that Harvard believes would "cause severe harm to individuals" [18] if accessed by non-authorized individuals. Harvard's examples include information covered by a regulation or agreement that requires that data be stored or processed in a secure manner but that can still be shared using proper encryption protections on data networks, including individually identifiable medical records and identifiable genetic information. Consistent with Harvard's security standards, the handling of Grade 5 data able to be stored on a networked computer maps to Crimson.

The model set of datatags are not the only possible levels or ways of defining datatags. Different uses may warrant different tag definitions and different ways of specifying security levels. The number of levels may also vary. We use the model datatags as a way of thinking about and discussing datatags, but we do not mean to imply the tag definitions shown are the only possible tags. Similarly, Harvard's way of assigning data to levels is not appropriate for all kinds of data and settings. We do not promote Harvard's assignments of data to these levels. We use them only to show datatags assignments to a real-world example of delineations.

Consider a datatags repository with thousands of security levels, where almost each file has a unique datatag. In this case, it may become difficult to reason about the performance of the system and its consistency in treatment of similar files without grouping tags into fewer categories. These groups might be learned through analysis or automated clustering. Arguably, these fewer categories might better serve as the repository's datatags, and the handling differences among files within the same category might become options within those datatags. Having fewer datatags usually allows opportunities for more robust feedback on the nature and consistency of data sharing decisions made by a repository and more easily allows technological guarantees and enhancements. Therefore, unless otherwise stated, we assume a datatags repository has a small number of datatags.

Security levels or datatags describe the minimal, uniform security protections available for a file. Because not all files tagged at the same level may have exactly the same requirements for access and handling, datatags discretize options.

To compare two different datatag repositories, both need to be compliant to the same security levels. We use the term "datatags-compliant repository" to denote a datatags repository that is compliant to a specific set of datatags. For convenience in this writing, when we say a datatags-compliant repository we mean one that is compliant to the model datatags, unless otherwise stated.

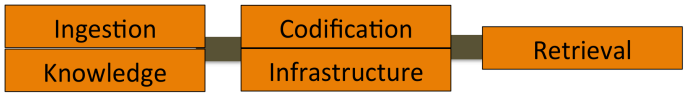
A datatags repository can be used to study data sharing regimes, such as privacy laws, or decisions made by a research review committee. A datatags repository can help evolve datatag definitions by learning from prior data sharing decisions.

When architecting a datatag repository for real-world use, we first specify the security levels or tags for a set of security features, such as transmission and storage standards, and delineate access requirements. Then, it is often convenient to think about the design of overall operation in terms of ingestion and decision making knowledge, codification and infrastructure (collectively called handling), and retrieval and credentialing. Figure 1a shows these three design groups.

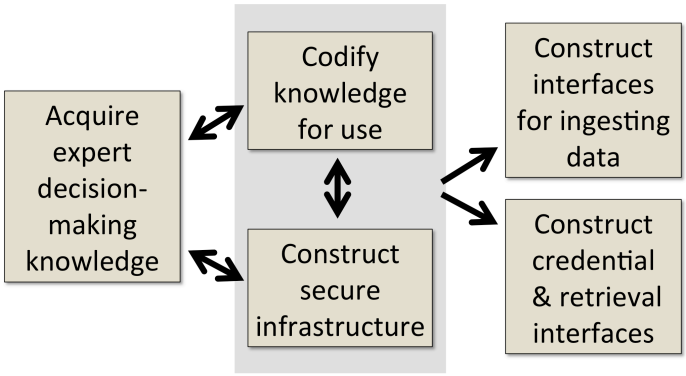
Ingestion and decision making knowledge refers to accepting data into the repository, with appropriate tag and any supplemental requirements. Credentials and retrieval refers to authenticating the recipient and acquiring acceptance of terms in accordance with the requirements associated with the requested file. Codification and infrastructure is the machinery between ingestion and retrieval; it executes the security levels of the repository. Sometimes codification may include ingestion. This occurs when tagging a file is simple or automated. Codification may also include retrieval. This usually occurs when there are no additional access requirements, only those specified by the security level.

When constructing a datatags repository, however, the components are more distinct and follow a flow. The critical components are: acquiring knowledge for decision making; codifying that knowledge into a machine readable form; constructing the secure

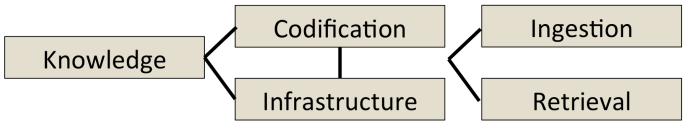
infrastructure; constructing interfaces for ingesting data; and constructing interfaces for retrieving data. Figure 1b shows how these relate. A feedback loop exists between knowledge acquisition, codification of that knowledge, and the secure infrastructure. The feedback is likely to establish or refine datatag definitions, as different security and access features may become necessary and others unnecessary. Codification of knowledge may reveal inconsistencies and inefficiencies. Interfaces for ingesting and receiving data may be human facing or machine interacting; these appear on the right. While it is convenient to think about ingestion early in the design stage (Figure 1a), during construction, ingestion refers to the interface for data acquisition, so in Figure 1b, ingestion appears on the right.



(a)



(b)



(c)

Figure 1. Design (a), constructed parts of a datatags repository (b), and their derived operational part names (c). Design has three groups: ingestion and decision making knowledge; codification and infrastructure; and retrieval and credentialing. Ingestion from (a) moves to the right in (b) and (c). The feedback loop between knowledge acquisition, knowledge codification, and the secure technical infrastructure establishes terms of agreements and properties to track, and may refine tag definitions. Any particular implementation may combine one or more of these components.

Each component in Figure 1b has its own notion of optimal efficiency and correctness. Knowledge acquisition wants to efficiently and correctly acquire and represent needed

knowledge and make human expert review efficient. Codification seeks to be accurate in its representation and use of the acquired knowledge. The secure infrastructure should technologically make assurances about the security features on which decision-making relies and include oversight and integrity checks. Finally, the interfaces seek to be unambiguous and time efficient.

Although we describe these components as if they are different sets of programs or tools, it is not necessary that they be separate. We merely use these distinctions to identify tasks and goals. There are many possible ways to construct a datatags repository. But no matter how constructed, a datags repository should operationally achieve the components described in Figure 1.

Automated Datatagging

An ambitious vision for ingestion includes an automated interview system that engages in a question-and-answer session with an arbitrary data depositor, and/or optionally inspects data files being ingested. There are more than 2000 applicable data sharing laws at the state and federal levels in the United States [20]. Additionally, some data sets are subject to binding contracts, data use agreements, data sharing restrictions, etc. An interview with a panel of humans, such as an IRB, or with a human expert, such as a privacy officer, is feasible because these parties already make security level (or datatags) determinations. Can an automated system do the same with arbitrary data?

In computer science, an expert system is a computer program that emulates human decision-making ability [21]. Computer scientists heavily pursued the development of autonomous, standalone expert systems in the 1980s with limited success [22]. However, human-assisted approaches are popular today. For example, TurboTax is a program that interactively works with taxpayers in the United States to help them complete their personal tax forms [23]. American tax law is horribly complicated, yet the system has been used successfully by millions of Americans. The TurboTax approach explicitly embeds knowledge into a carefully structured series of yes-no (or multiple-choice) questions, and then walks the user stepwise through the questions.

This approach suggests three areas of consideration, from knowledge acquisition, through a machine-readable medium, to the user interface that would actually interact with data depositors (see the topmost path in Figure 1). Tools for knowledge acquisition extract data sharing decision-making knowledge from human experts or documents and allow human experts to subsequently edit and review the acquired knowledge. Codification of the knowledge involves machine interpretation of the knowledge sufficient to generate executable interferences from inputs provided by a data depositor. Finally, ingesting data involves the actual user interface through which data depositors engage. Tools in this component may include automated data inspection programs as well as question-and-answer displays.

Formal Description

More formally, a datatags repository combines a classification problem with an access-control problem.

Let D be the set of datatag levels for a datatags repository. The datatags are partially ordered and at least two distinct datatags must be strictly ordered.

$$D = \{d_1, \cdots, d_n\}.\ \forall d_i \in D: \{d_1, \cdots, d_i\} < \cdots < \{d_j \cdots\ d_n\}\ and\ d_1 \neq d_n$$

Let F be the files assigned to the repository. Then, there is a function G that maps each file in F to a datatag such that if the information in two files requires handling with similar security features and access credentials, they have the same datatag.

$$G: F \rightarrow D.\ Let\ f_i, f_j \in F\ and\ d_k \in D.\ Then\ G(f_i) = d_k\ and\ G(f_j) = d_k\ iff\ f_i \approx f_j$$

This is a classification problem in machine learning, suggesting that cluster-analysis algorithms may be effective. We envision using cluster analysis to learn optimal sets of datatag definitions from previously tagged files, as well as to identify which datatag cluster a new file most closely matches.

Determining whether a requester should have access to a file is a classic access control problem. If requesters are given explicit, hard-coded permissions to a file, a datatag, or groups of datatags, then the approach is the same as mandatory access control. If requesters obtain permission by their role assignments, then the approach is the same as role-based access. More generally, if requesters assert attributes about themselves (RA) and datatags are expressed with attributes (DA), then the approach is one of matching the attributes of requesters at request time to tagged files.

The files available to any single requester are a subset of the many-to-many assignment relation:

$$RA \times DA$$

Clearly, there are many ways to construct datatag repositories.

Results

In this section, we use the definitions and notions we introduced about datatags to architect examples of datatag repositories for real-world use.

HIPAA Research Archive

The Health Information Portability and Accountability Act (HIPAA) is a 1996 federal statute that authorized the U.S. Department of Health and Human Services to establish privacy rules

governing individually identifiable health information, rules that specify with whom and how physicians, hospitals, and insurers may share a patient's medical information.

The Safe Harbor provision of the HIPAA Privacy Rule prescribes a way to share medical data publicly [24]. Dates may only include the year. HIPAA requires that postal codes ("ZIPs" in the United States) contain only the first three digits if the population in those ZIP codes is greater than 20,000. ZIP codes for populations less than 20,000 report a null ZIP of 00000. No explicit identifiers, such as name, Social Security numbers, or addresses, can appear. HIPAA allows researchers access to more detailed data, including full dates and postal codes, with approval of an IRB [25] [13].

Adam, the hypothetical medical researcher mentioned earlier, wants to use a datatags repository to store both HIPAA Safe Harbor and IRB-approved versions of his team's various research data sets. His repository will use the Green, Red, and Crimson datatags described in Table 1.

Because HIPAA covers the medical data, the data files are tagged Red, with the exception of any file containing clinical notes, psychiatric records, or HIV-AIDS information. Those have greater disclosure restrictions and are assigned Crimson.

Because all the data files originate from the medical-information system, Adam writes a simple computer script to use the names of the data fields to automatically generate a HIPAA Safe Harbor version of the files. The Safe Harbor version is tagged Green. He selects Green and not Blue for Safe Harbor files because he is concerned that even though Safe Harbor data are considered fully anonymous by regulation, a risk of re-identification remains. Therefore, he requires the email address of anyone downloading the data, and he drafts an accompanying data use agreement requiring recipients to notify him if they find a vulnerability. The agreement also requires recipients to cease use and reapply for any data when notified by Adam that there is a vulnerability. (The data use agreement does not prohibit data linkage because doing so limits worthy uses of the data and does not allow Adam to learn if real risks exist [19].) A design summary appears in Table 2 and an operational depiction in Figure 2.

Ingestion and Decision-making Knowledge	A HIPAA-consistent Safe Harbor script redacts data files to produce a version for sharing under the Green tag. It assigns a Crimson tag to any file if finds that contains clinical notes, psychiatric notes, or HIV-AIDS information. It assigns a Red tag to all other data files and to the original non-redacted files that are not Crimson.
Codification and Infrastructure	See Table 1 for Green, Red, and Crimson tags.
Credentials and Retrieval	Data use agreements. Red and Crimson are limited to those who qualify based on IRB review. Data use agreements describe handling requirements beyond the repository for downloaded files.

Table 2. Design of a datatags-compliant repository for sharing research data under the HIPAA Privacy Rule.

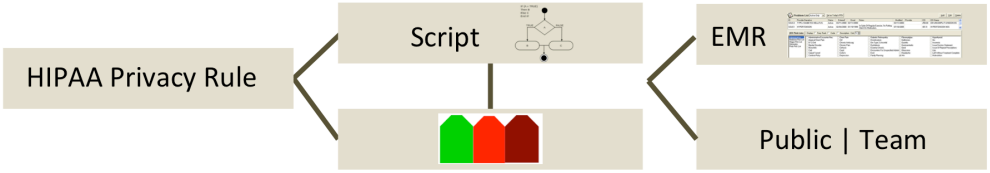


Figure 2. Operational components for a datatags-compliant repository for a medical-research team to share data consistent with the HIPAA Privacy Rule. Decisions that rely on HIPAA are codified in a program that makes appropriate Green, Red, and Crimson datatags from patient medical information drawn from the University hospital's EMR (electronic medical records) system. Compliant members of the public have access to files tagged Green. Only appropriate research team members have access to files tagged Red or Crimson.

Multinational Corporation Archive

Multinational corporations acquire data from many diverse sources, under a multitude of agreements, and are subject to many laws, regulations, and business practices. Some data pose more liability for the company than do other data, and some data have geographical restrictions that prohibit sharing with all parts of the corporation. Diane, the chief privacy officer for the hypothetical corporation mentioned earlier, wants to use a datatags repository to keep track of different restrictions and requirements on files and to enforce sharing obligations throughout the corporation.

Diane's repository will use all six of the datatags described in Table 1. She replaces the notion of a data use agreement described in Table 1 with a document or click-through notice that describes important use and handling instructions for employees. Assignments of datatags will not necessarily be based on the sensitivity of personal information. Instead, in Diane's repository, datatags measure the likelihood of corporate liability and concerns about compliance and adverse publicity.

Some of the data include personal information, such as registration and real-time use information from buyers of company products, members of its focus groups, and visitors to its websites, along with individually identifiable market information purchased from data brokers. Other data include business analytics, such as sales forecasts, customer profiles, strategic market information, and purchased market analytics from other companies. Different agencies and regulatory bodies regulate some categories of data.

Data comes into the company through divisions and project groups, each of which has a privacy officer responsible for making sure the company complies with its obligations. These local privacy officers work with project leaders to make decisions about data handling. Diane

informs them that they will be responsible for codifying their data sharing decisions by tagging data files as the files come into the repository. The tagging includes assigning a datatag, specifying any additional handling restrictions like geography or the kind of employee that can access the file, and notifying any employee who downloads the file of any special handling requirements.

Access to data files relies on employee identification credentials and roles assigned to employees. For example, company offices in some parts of the world may have access to some files and not others. Employees on one project may have access to project files but not other files. Diane relies on the company's existing role-based access system [26] for access credentials in the datatags repository. She also asserts that each employee must receive a copy of data directly from the datatags repository. Employees cannot share files, even with other employees. A design summary appears in Table 3 and an operational depiction in Figure 3.

Once the system is operational, Diane monitors tagging decisions and access logs for consistency and spot checks for appropriateness. She can now answer questions about the company's internal data sharing policy for any specific data set or any kind of data.

Ingestion and Decision-making Knowledge	Local privacy officers and project leaders determine which datatags apply to which data sets and specify any additional restrictions or notices that apply.
Codification and Infrastructure	See Table 1 for all six tags: Blue, Green, Yellow, Orange, Red, and Crimson, with access based in part on the company's role-based access system.
Credentials and Retrieval	Employees having appropriate credentials in the company's role-based access system may access a file in the datags repository after acknowledging receipt of any notices about special handling required for the file. Employees may not share the files, even with other employees.

Table 3. Design of a datatags-compliant repository to help a multinational corporation's internal data sharing comply with a multitude of obligations.

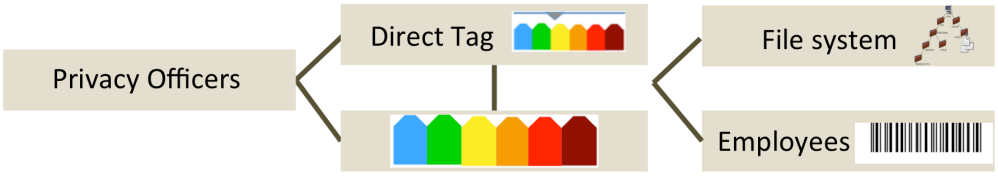


Figure 3. Operational components for a datatags-compliant repository for a multinational corporation. Privacy officers local to the projects that source the data directly tag files from the corporate file system. Employees with the appropriate credentials (or roles) may access corresponding files.

An Institutional Review Board Archive

All federally funded research studies in the United States involving human subjects must (a) ensure that research procedures fulfill the principles of voluntary participation and informed consent, (b) maintain the confidentiality of information obtained from or about human participants, and (c) adequately address possible risks to participants, including data sharing risks. A research initiative might be objectionable or problematic if subjects experience physical, social, or economic harm, or if subjects are wronged in the sense that the research violates a dignity or intrinsic interest, even if subjects do not experience a setback or damage as a result. An IRB applies these general principles when reviewing a specific research protocol to make determinations, including decisions about data collection and data sharing. Charles, the hypothetical leader of an IRB, constructs a datatags repository to track decision-making by his IRB. His university does not give him the funds to actually implement a system capable of storing data, but he does have the resources to construct a datatags repository that documents data sharing permissions granted by the IRB.

Charles' repository uses all six of the datatags described in Table 1. Rather than holding actual data files, the repository contains descriptions of the data and the accompanying research protocol. All IRB committee members can access any file in the repository, but there is no electronic implementation of access restrictions. The access requirements associated with a datatag merely document the classification decision made by the committee.

When the IRB reviews a research study, it stores any decisions it makes about data handling in the repository. The decision includes which datatag applies and any special handling requirements imposed on the researcher. The researcher is independently responsible for adhering to the provisions on her computer or in a third-party repository. Files related to the IRB decision, such as a description of the data, a summary of the research study, and notes from the committee, are stored in the repository under the same datatag and marked with the same handling requirements as the IRB applies to the data and the researcher. The purpose of storage in the repository is not to enforce data access but to document decisions made about data access and handling. Therefore, any member of the IRB can retrieve any files by password, regardless of the access requirements associated with the datatag.

Additionally, Charles uses programs that produce summary information from the repository. One program displays the number and nature of files stored with each datatag, along with handling variations that appear within the same datatag, thereby summarizing decisions for each datatag. Another program reports on the treatment of similar kinds of data across datatags, thereby possibly revealing inconsistencies. A third program accepts a description of a research study, searches the repository for similar cases, and reports tag and handling decisions made for those cases, thereby providing recommendations to the IRB based on its historical decisions. A design summary appears in Table 4 and an operational depiction in Figure 4.

Ingestion and Decision-making Knowledge	The IRB determines which datatags apply to which data sets and specifies any additional restrictions that apply. A copy of IRB documents appears as files in the repository, and not the data themselves.
Codification and Infrastructure	See Table 1 for all six tags: Blue, Green, Yellow, Orange, Red, and Crimson. However, the access requirements associated with the tags are not used to access the IRB files. IRB members have password access to any file in the repository.
Credentials and Retrieval	IRB committee members can retrieve documents describing the data, as well as summary reports about the nature of data archived at each level.

Table 4. Design of a datatags-compliant repository to document IRB data sharing decisions.

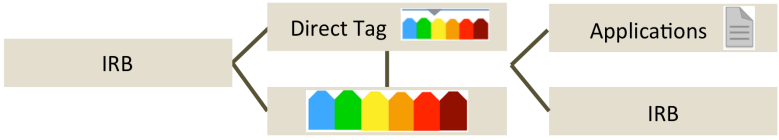


Figure 4. Operational components for a datatags-compliant repository for an IRB to track its decisions about applications.

Global Research Repository

Betty, the hypothetical sole researcher, learns that Jane, a hypothetical developer, maintains a Global Research Repository that supports a full range of sensitive data from any researcher in the world and has all the archival and support features described of Dataverse. The Global Research Repository uses all six of the datatags described in Table 1.

When a researcher submits data to the Global Research Repository, the researcher may not know which datatag applies or what additional terms are necessary. For example, a researcher submitting video data captured in the United States may be unaware that American jurisprudence has different standards for photographs than for sound. One can generally photograph anyone in public, but recording conversations may require explicit consent from all parties, depending on the state in which the recording is made. Violators can be subject to felony offenses under wiretap laws and it is impermissible to share the data through a public repository.

Jane envisions two options at ingestion. The researcher could have an IRB determine the appropriate datatag and specify any additional handling requirements. Alternatively, the researcher could use an interview system that offers expert advice on an appropriate datatag and handling requirements. Jane wants to avoid having researchers select their own datatags because she worries that researchers may not know all the legal requirements that

may apply or may have a tendency to over- or under protect data. Table 5 has a design summary, and Figure 5 shows an operational depiction.

Ingestion and Decision-making Knowledge	IRB determination or an interview system.
Codification and Infrastructure	See Table 1 for all six tags: Blue, Green, Yellow, Orange, Red, Crimson.
Credentials and Retrieval	Beyond what appears in Table 1, different files may require specific terms of use based on legal or regulatory requirements or adopted best practices.

Table 5. Design of a datatags-compliant repository used as a general-purpose research repository.

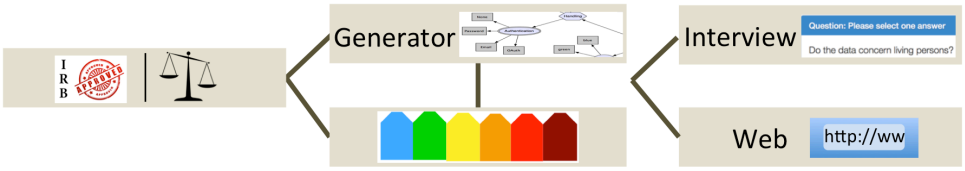


Figure 5. Operational components for a datatags-compliant repository for general-purpose research data sharing. Researchers deposit data through an online question-and-answer interview that tags the data in accordance with a codification of IRB decisions and legal jurisprudence. Other researchers and the public can obtain data through a website after providing appropriate credentials.

Jane's vision of an automated interview system that engages in a question-and-answer session with an arbitrary researcher, and/or optionally inspects research data at ingestion, is ambitious. To show feasibility, we constructed demonstration versions of the three critical components — knowledge acquisition, codification, and the interface for ingestion — described in Figure 1.

As a demonstration of a knowledge-acquisition tool, we constructed a decision tree that embeds a sequence of yes-or-no questions for assigning relevant datatags pursuant to consent and HIPAA standards. We had human experts review the tree for accuracy; see Table 6. We use decision trees for demonstration purposes. There are many possible ways to elicit and represent expert knowledge. More details about Table 6 appear shortly.

We also surveyed 50 different data use agreements currently associated with health data shared through government agencies, hospital associations, and others (see the Data citation in this paper for a copy and [27] for originating sources). We harvested operational terms found in these data use agreements because data acquired under some of these data use agreements and then deposited into a datatags repository may have to apply some of the

terms to data recipients. Table 7 lists the terms and their options. More details about Table 7 appear below.

Start						
1	Do your data include personally identifiable information?					
2		No	[BLUE, basis=not personal info, identity=not person-specific]			
3	Did each person whose information appears in the data give explicit permission to share the data?					
4		Yes	Did the consent have any restrictions on sharing?			
5			No	[GREEN, basis=Consent, identity=___]		
6			Yes	[GREEN, basis=Consent, identity=___] Add special terms (Table 7).		
7	Do the data contain personal health information?					
8		Yes	Were the data received from a HIPAA-covered entity or a business associate of one?			
9			Yes	Do the data visually adhere to the HIPAA Safe Harbor Provision (e.g., dates in years and first two digits of ZIPs)?		
10				No	[GREEN, basis=HIPAA Safe Harbor, identity=de-identified]	
11			Yes	Has an expert certified the data as being of minimal risk?		
12				Yes	[GREEN, basis=HIPAA Statistician, identity=de-identified]	
13			Yes	Did you acquire the data under a HIPAA limited-data use agreement?		
14				Yes	Did the limited-data use agreement have any restrictions on sharing?	
15					No	[ORANGE, basis=HIPAA Limited Data Set, identity=identifiable]
16					Yes	ORANGE, basis=HIPAA Limited Data Set, identity=identifiable] Add special terms (Table 7).
17			Yes	Did you acquire the data under a HIPAA Business Associate agreement?		
18				Yes	Did the business associate agreement have any restrictions on sharing?	
19					No	[RED, basis=HIPAA Business Associate, identity =identifiable]
20					Yes	[RED, basis=HIPAA Business Associate, identity =identifiable] Add special terms (Table 7).
21			Yes	Are you an entity that is directly or indirectly covered by HIPAA?		
22				Yes	[RED, basis=HIPAA Covered Entity, identity=identifiable]	
23		No	Did the data have any restrictions on sharing (e.g., stated in an agreement or policy statement)?			
24			No	[GREEN, basis=Agreement, identity=___]		
25			Yes	[GREEN, basis=Agreement, identity=___] Add special terms (Table 7).		
26	Unable to tag. This version processes consent and medical data only.					

Table 6. Decision-tree flowchart designed for human-expert review to assign datatags in an interview process. It covers consent and medical data in the United States. Questions appear in blue; terminating assignments appear in orange. The value of an assignment is a tuple [datatag, basis, identity], where datatag is one of those identified in Table 1, basis identifies the governing standard, and identity describes the kind of personal information expected to be stored. In lines 6, 16, 20, and 25, processing continues to Table 7 to establish additional terms for the file's data use agreement.

In the decision tree in Table 6, the order of questions (colored blue) starts at the top, and execution flows downwards until encountering a terminating node (colored orange). Different choices of yes and no lead to different paths. A terminating node makes a datatag assignment and optionally continues execution to the second decision tree to add terms to a data use agreement. A tag appears in the decision tree as a tuple [datatag, basis, identity] in this example. *Datatag* is one of the six security levels listed in Table 1 and identified by the name of its assigned color, Blue, Green, Yellow, Orange, Red, or Crimson. *Basis* stores a reference to the relevant standard used for making the determination. *Identity* is a reference as to whether the personal information stored in the file appears with explicit identifiers, such as name or address (identified) or without these kinds of identifiers (de-identified). We added basis and identity as examples of attributes that can be additionally tagged with a file for overall reporting about repository contents.

The first question, on line 1, asks whether the data include any personal information. If the file does not contain personal information, a determination results on line 2 that the data are tagged Blue and concludes the path through the decision tree. On the other hand, if the file does contain personally identifiable information, the question on line 3 asks whether the subjects of the information gave explicit permission for sharing [28]. If so, the question on line 4 is next. Regardless of the answer, the file is tagged Green because of the consent, and if there are restrictions imposed on the consent (line 6), execution continues with Table 7, even though the identifiability of the data is unknown (specified by a blank assignment to identity).

Suppose instead that the data came from a HIPAA covered entity, which is a healthcare provider or insurance company. Then there are five possible questions to ask, shown on lines 9, 11, 13, 17, and 21, which are asked in turn if no terminating node is encountered.

Some of the lines make a tag assignment and then continue to Table 7 to add terms to a data use agreement; see lines 6, 16, 20, and 25. Some repositories may allow any data depositor to set these terms, or may provide defaults. This is an exemplary, not an exhaustive list.

Unlike in Table 6, each of the leftmost groups in Table 7 specifies exemplary attributes that must be set and that dictate additional terms for the data use agreement. These terms concern time limit, sharing limitations, linking, publication, uses, and approvals. Each must be set in turn. A qualified recipient of the file may or may not be required to delete all copies of the data based on a time limit, allowed to share the file with others, allowed to link or

match the file with others or contact research participants, allowed to publish freely, allowed to use the data for any purpose, or required to obtain the depositor's review for each data request.

Examine the terms in the agreement under which the data originated and set the following terms accordingly. Provide a selection for each of the six terms.

Select one of the following to specify a time limit.

- May a qualified person use the data indefinitely?
  - Yes timelimit=none
- May a qualified person use the data for one year?
  - Yes timelimit=1yr
- May a qualified person use the data for two years?
  - Yes timelimit=2yrs
- May a qualified person use the data for five years?
  - Yes timelimit=5yrs
- Repeat until timelimit is set.

Select one of the following to specify further data sharing.

- May a qualified person share the data with others freely?
  - Yes sharing=anyone
- May a qualified person share the data with others provided the data is not made publicly available online?
  - Yes sharing=not online
- May a qualified person share the data with others in his organization?
  - Yes sharing=organization
- May a qualified person share the data with others in his immediate work group?
  - Yes sharing=group
- Is a qualified recipient prohibited from sharing the data with anyone?
  - Yes sharing=no one
- Repeat until sharing is set.

Select one of the following to specify whether the recipient can link the file to other data.

- Is a qualified person prohibited from matching data to other data?
  - Yes linking=none
- Is a qualified recipient prohibited from identifying and contacting people or organizations in the data?
  - Yes linking=no entities
- Is a qualified recipient prohibited from identifying and contacting people whose information is in the data?
  - Yes linking=no people
- Is a qualified person allowed to re-identify and contact people whose information is in the data?
  - Yes linking=no prohibition

Is a qualified person allowed to re-identify but not contact people whose information is in the data?
- Yes linking=re-identify

Is a qualified person allowed to contact people whose information is in the data?
- Yes linking=contact

Repeat until linking is set.

Select one of the following to specify any publication restrictions.

May a qualified user freely publish papers using the data?
- Yes publication=no restriction

Must a qualified recipient notify you of any publications using the data?
- Yes publication=notify

Must a qualified recipient get your pre-approval before publishing a paper that uses the data?
- Yes publication=pre-approve

Are there to be no publications using your data?
- Yes publication=prohibited

Repeat until publication is set.

Select one of the following to specify any restrictions on use.

May a qualified user freely use the data for any purpose?
- Yes use=no restriction

Must a qualified recipient use the data for research only?
- Yes use=research

Must a qualified recipient use the data for IRB-approved research only?
- Yes use=IRB

Is a qualified user prohibited from incorporating the data into products or other data?
- Yes use=no product

Repeat until use is set.

Select one of the following to specify whether you want to personally approve each data request.

May a qualified user use the data with no further approval from you?
- Yes approval=none

Do you want to approve each application for use by email?
- Yes approval=email

Do you want to approve each application by submitting your digitally signed approval?
- Yes approval=signed

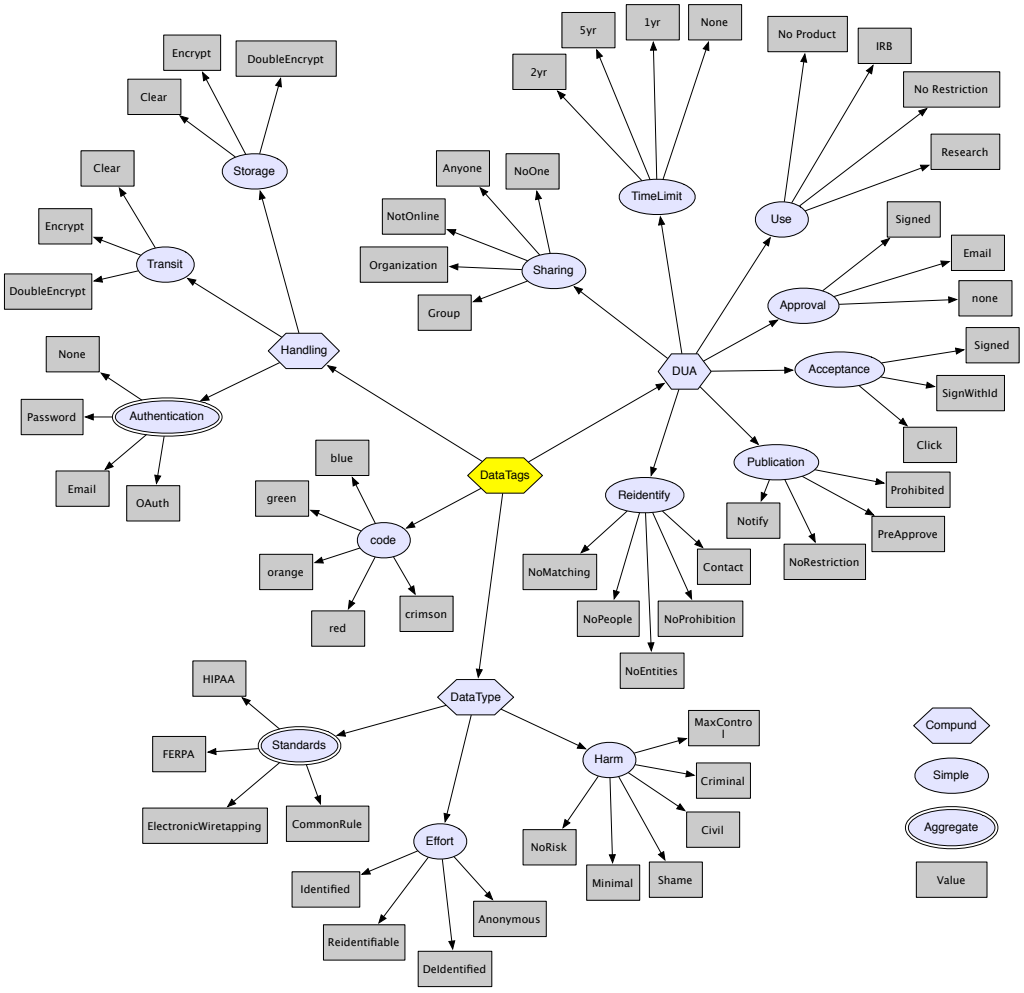
Repeat until approval is set.

Table 7. Additional terms for data use agreements. Each group specifies a setting for one term. Possible terms concern time limits, data sharing, linking, publication, uses, and approvals.

Human experts do not need decision trees. Instead, experts help build decision trees as a way to represent their knowledge for non-experts to use. Therefore, a critical question is whether decision trees are an efficient means for human expert review. Decision trees offer several advantages. First, the representation is succinct, making it easier for human experts to focus on decision-making at a high level. The tree localizes decisions, making it easier to understand the exact criteria used. As discussed earlier, policy language specification can yield unforeseen decisions because all applicable rules apply in combination, and some combinations may not be anticipated. In comparison, the decision tree approach, like the decision tree in Table 6, has a strict ordering that embeds question precedence based on the order in which questions appear in the tree. This makes unexpected decisions less likely. We are not necessarily advocating that decision trees are the only or best way to work with human experts. We mention these advantages in comparison to policy language specification in order to underscore the kinds of issues to consider in any implementation.

Of course, while decision trees may be efficient for human expert review, they are not necessarily appropriate or efficient for data requesters to use. For example, a multiple-choice option may be more efficient for data depositors to use at times than a series of questions. Terms used in the decision tree assume expert knowledge, and data depositors may not be able to apply the terms appropriately. Further, the tree format may be easier for experts to review because of its succinctness, but it is not necessarily a good format for computational processing to guarantee that all paths end with a well-formed assignment and that there is no redundancy or ambiguity in questions. To address these concerns, we introduce a formal language and language interpreter for representing and using the knowledge and incorporating user-friendly text as an example of codification.

Representing the tag space as a graph allows us to reason about it using graph theory. Under these terms, assigning a datatag and specifying any additional data use agreement terms translates to selecting a sub-graph from the tag space graph. A single node n is fully specified in sub-graph S, if S contains an edge from n to one of its leaves. A compound node c is 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. Figure 6 shows the tag space for HIPAA generated by our language interpreter (see the data citation of this paper for a copy).



(a)

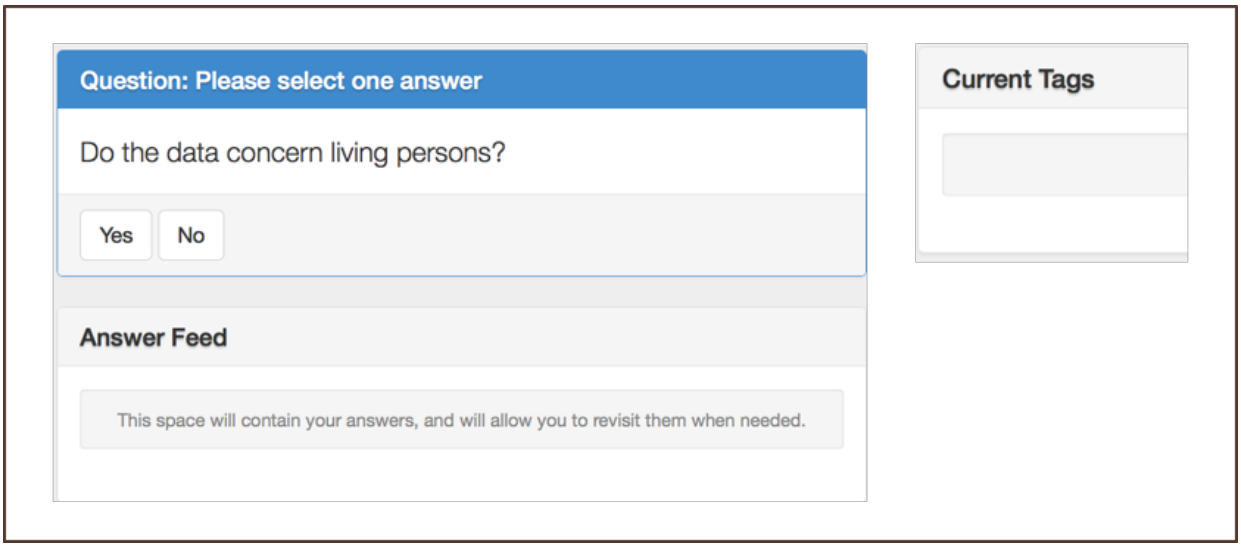
```
DataType: Standards, Effort, Harm.

Standards: some of HIPAA, FERPA,
                   ElectronicWiretapping,
                   CommonRule.

Effort: one of Identified, Identifiable,
               DeIdentified, Anonymous.

Harm: one of NoRisk, Minimal, Shame, Civil,
             Criminal, MaxControl.
```

(b)

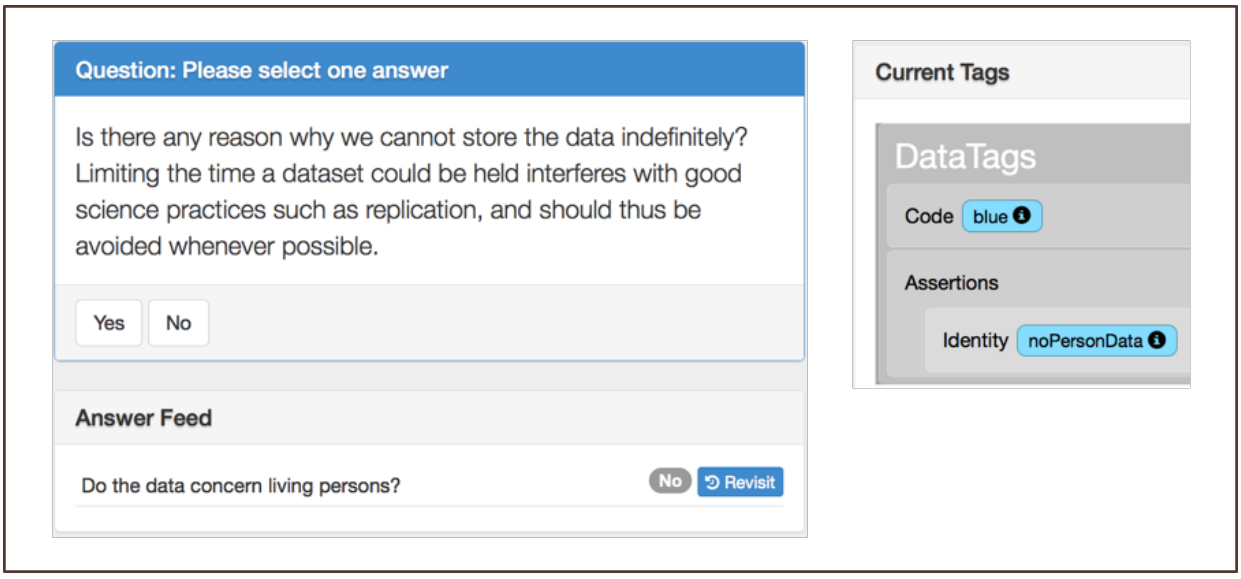
Figure 6. The tag space graph needed for HIPAA compliance (a), and part of the code used to describe it (b). Our language interpreter created the base graph for the diagram.

Finally, we constructed a user interface and improved it through usability testing. Usability is critical to convincing researchers to invest the time and energy necessary to determine the correct datatag on ingestion of files into the repository. 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. Iterative improvements resulting from usability testing included being able to convey the progress made so far in a gratifying way; keeping the user engaged in the process; and making sure that whenever a technical or a legal term appears, an explanation is readily available. These improvements help make the tagging process approachable and non-intimidating.

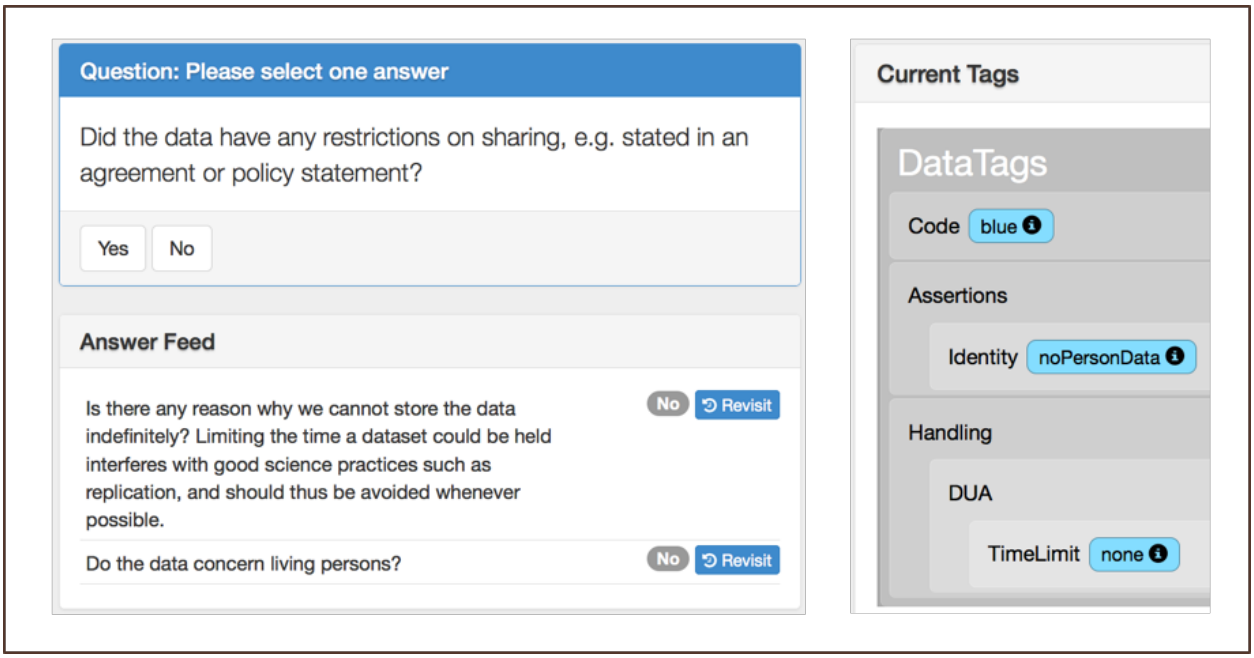
The demonstration remains live at datatags.org and tags datasets for which consent, HIPAA, education, or government records laws apply [29]. Figure 7 contains a sequence of screenshots. The contents of the user interface come directly from the language specification to ensure accuracy of representation. Figure 7 shows the execution of lines 1 and 2 from Table 6 and the start of Table 7. (In the usable version, execution continues to Table 7 regardless of whether a predicating agreement exists.) Figure 8 shows a trace of the engine for the steps shown in Figure 7.



(a)



(b)



(c)

Figure 7. Execution of the user-level demo at datatags.org for (a) lines 1 of Table 6, (b) line 2 of Table 6, and (c) start of Table 7. Questions are on the left and results on the right.

Engine Trace

Useful for debugging the interview

Id	Type	Info
duaAdditional	ask	Did the data have any restrictions on sharing, e.g. stated in an agreement or policy statement?
$275	set	[TagValue name:null type:[CompoundType name:DataTags]]
duaTimeLimit	ask	Is there any reason why we cannot store the data indefinitely? Limiting the time a dataset could be held interferes with good science practices such as replication, and should thus be avoided whenever possible.
dua	todo	Data use agreements
$11	call	dua
$10	set	[TagValue name:null type:[CompoundType name:DataTags]]
$0	ask	Do the data concern living persons?

Figure 8. Engine trace for the steps itemized in Figure 7 of the user-level system.

By providing these exemplars of decision trees, tagging language and interpreter, and a usable system, we provide evidence of an automated interview system.

Discussion

We define a datatag repository as capable of storing and sharing files that have different security-level requirements. Benefits include reducing handling complexity to a few well-formed choices and providing guarantees about appropriate handling, which may include auditable accounting. Datatag repositories seem useful in a variety of government, research and corporate applications, especially when the data recipients are numerous or diverse. We provided architectures of datatag repositories for several real-world uses. Some of these architectures are directly implementable.

The most ambitious undertaking is a global research repository that ingests and widely shares virtually any kind of data. Many large-scale repositories limit by discipline or kind the nature of ingested data, thereby limiting the type of tagging decisions required. However, in the most general case, the ingested data could be from any real-world domain, subject to many possible laws, regulations, and best practices. Developing an interview system or a tagging strategy is a large effort requiring a community of experts to review and compose tagging decisions. We envision a Wikipedia-like effort in which lawyers, researchers, government agencies, advocacy groups, scholars, and the public can view, edit, and compose decision-making knowledge using decision trees or some other representation. Alternatively, we envision hiring a team of experts to develop the necessary decision-making knowledge

over time. We developed the decision tree that is in use at the demo application running at datatags.org with this method. It relies on the legal and technical expertise of members of Harvard's Data Privacy Lab and Berkman Center for Internet and Society.

A datatags repository is not only for sharing files; it also documents data sharing decisions. In this way, a datatags repository can be useful in characterizing, quantifying and analyzing the nature and consistency of data sharing regimes, including those defined by IRBs, policies, best practices, or laws and regulations. In these cases, one develops the datatags repository for the study of how different kinds of data can be shared, even though the repository itself does not share actual data.

There are many possible studies for future work and no shortage of open questions remaining. These include, but are not limited to, studies on datatag definitions, data classifications, flow analyses, use cases, access control models, optimization, data sharing policies, and so on.

References

1. Holdren J. Memorandum for the Heads of Executive Departments and Agencies. Office of Science and Technology Policy. Executive Office of the President of the United States. February 22. 2013. http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

2. Altman M, Adams M, Crabtree J, et al. Digital preservation through archival collaboration: the data preservation alliance for the social sciences. The American Archivist. 72 (Spring/Summer 2009). pp170-184. http://archivists.metapress.com/content/eu7252lhnrp7h188/fulltext.pdf

3. Roeder J. Considerations in the development of a National Geophysical Data Policy. EOS, Transactions American Geophysical Union. June 3, 2011.

4. Vision T. Open Data and the Social Contract of Scientific Publishing, BioScience. 2010. 60(5):330-330. doi:10.1525/bio.2010.60.5.2

5. Grobe H, Diepenbroek M, Dittert N, Reinke M, Sieger R. Archiving and distributing earth-science data with the PANGAEA information system. Antarctica: contributions to global earth sciences; Proceedings of the IX International Symposium of Antarctic Earth Sciences Potsdam, 2003 / Hrsg. Dieter Futterer; Detlef Damaske; Georg Kleinschmidt, Hubert Miller, Franz Tessensohn; Springer, Berlin, 2006. pp 403-406.

6. Hrynaszkiewcz I and Altman D. Towards agreement on best practice for publishing raw clinical trials data. Trials 2009,10:17. http://www.biomedcentral.com/content/pdf/1745-6215-10-17.pdf

7. Schwartz A, Pappas C, and Sandlow L. Data repositories for medical education research: issues and recommendations. Academic Medicine. 85(5). May 2010. pp. 837-843

8. The Dataverse Project. http://dataverse.org

9. Crosas M. A Data Sharing Story. Journal of eScience Librarianship. 2013. 1 (3):17379.

10. Crosas M. The Dataverse Network: An Open-Source Application for Sharing, Discovering and Preserving Data. D-Lib Magazine. 2011. 17 (12).

11. King G. An Introduction to the Dataverse Network as an Infrastructure for Data Sharing. Sociological Methods and Research. 2007. 36 (2): 17399.

12. The Harvard Dataverse. http://dataverse.harvard.edu

13. Common Rule and Institutional Review Board. 45 CFR Part 46 Protection of Human Subjects. Revised 2009.

14. Ferraiolo D, Sandhu R, Gavrila S, Kuhn D, Chandramouli R. Proposed NIST standard for role-based access control. ACM Transactions on Information and System Security (TISSEC) 4(3) (2001) 224–274

15. Kagal L, Finin T, Joshi A. A policy language for pervasive systems. Fourth IEEE International Workshop on Policies for Distributed Systems and Networks. 2003.

16. Tonti G, Bradshaw J., Jeffers R, Montanar R, Suri1 N, Uszok1 A. Semantic web languages for policy representation and reasoning: A comparison of kaos, rei, and ponder. In: Proceedings of the 2nd International Semantic Web Conference (ISWC2003), Springer-Verlag. 2003.

17. iRODS. University of North Carolina. http://irods.org

18. Harvard Security Levels. (This writing uses “grade” to refer to “level” on the web page.) Harvard University. Web September 9, 2015. http://policy.security.harvard.edu/view-data-security-level

19. Sweeney L. Only You, Your Doctor, and Many Others May Know. Technology Science. 2015092903. September 29, 2015. http://techscience.org/a/2015092903

20. Smith R. The Compilation of State and Federal Privacy Law. Privacy Journal. 2014.

21. Jackson P. Introduction To Expert Systems (3 ed.). Addison Wesley. 1998

22. Russell S and Norvig P. Artificial Intelligence: A Modern Approach. Simon & Schuster. 1995. pp. 22–23.

23. Chipman M. TurboTax.1981. People.forbes.com. 2012-04-18. Current versions available at https://turbotax.intuit.com/.

24. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. U.S. Health and Human Services. Web September 9, 2015. http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html

25. Health Information Portability and Accountability Act (HIPAA) Privacy Rule. 45 CFR Part 160 and Subparts A and E of Part 164. http://www.hhs.gov/ocr/privacy/hipaa/administrative/privacyrule/

26. Ferraiolo D, Sandhu R, Gavrila S, Kuhn D, Chandramouli R. Proposed NIST standard for role-based access control. ACM Transactions on Information and System Security (TISSEC) 4 (3). 2001. pp 224–274

27. Hooley S and Sweeney L. Survey of Publicly Available State Health Databases. Harvard University. Data Privacy Lab. 1064-1. June 2013. http://thedatamap.org/1075-1.pdf

28. There are many forms of consent, including the limitation of sharing only with a specific individual, for a specific purpose, or until a specific date. Restrictions can come from the data subject, data source, or researcher depositing the data. Depending on the use, it may become necessary to identify the source of a restriction or just the nature of the restriction.

29. The DataTags Interview Demonstration. http://datatags.org

---

Authors

Latanya Sweeney is Professor of Government and Technology in Residence at Harvard University, Director of the Data Privacy Lab at Harvard, Editor-in-Chief of Technology Science, and was formerly the Chief Technology Officer of the U.S. Federal Trade Commission. She earned her PhD in computer science from the Massachusetts Institute of Technology and her undergraduate degree from Harvard. More information about Dr. Sweeney is available at her website at latanyasweeney.org. As Editor-In-Chief of Technology Science, Professor Sweeeney was recused from the review of this paper.

Mercè Crosas is the Director of Data Science at Harvard's Institute for Quantitative Social Science (IQSS). Her team combines expertise in software engineering, statistical innovation and data curation and management to develop software applications for data sharing and analysis. The team's applications include Dataverse for publishing research data, Zelig and TwoRavens for statistical analysis, DataTags for sharing and handling sensitive data, and Consilience for clustering analysis and annotation of text. Before joining IQSS, Crosas led the software development efforts in educational and biotech industries. Prior to that, she was a pre- and post-doctoral fellow and a researcher at the Harvard-Smithsonian Center for Astrophysics. She holds Ph.D. in Astrophysics from Rice University and a B.S. in Physics from the Universitat de Barcelona, Spain.

Michael Bar-Sinai is a PhD candidate in Computer Science at the Ben-Gurion University of the Negev, Israel, and a fellow at the Institute for Quantitative Social Science at Harvard University. His research interests include programming language, software engineering, and issues laying at the intersection of society and software systems, such as privacy. Prior to resuming his academic studies, Michael worked as a software consultant in the UK, the US, and Israel.

The authors thank Sean Hooley, Alexandra Wood, David O'Brien, Stephen Chong, Salil Vadhan, Gary King, and the other members of the Privacy Tools Project at Harvard. This work was funded by grant CNS-1237235 from the National Science Foundation.

Referring Editor: Urs Gasser, Daniel Weitzner

Citation

Sweeney L, Crosas M, Bar-Sinai M. Sharing Sensitive Data with Confidence: The DataTags System. *Technology Science*. 2015101601. October 16, 2015. http://techscience.org/a/2015101601

Data

Under review for data sharing classification. See also http://datatags.org