When deposited in datatags-compliant data governed by certain law combinations without a repository, data sharing [15]. To share data while respecting legislation, it is important to consider the strictness of the policy. For instance, propositions like “this policy is stricter than that policy” must be evaluated at the repository level, with all necessary information available on demand. A systematic approach to managing privacy involves the development of repositories that can support data sharing and compliance with legal requirements.

Methodologies for questionnaire development. Data for this paper was gathered using datatags, a set of tools that make it easy to create a user-friendly web-based interview system. We present some of the datatags’ tools, such as interview systems, visualizers, development inspection tools.

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 collection methods, and legal and technological expertise — matching requires familiarity with the dataset’s history and the datatag most appropriate for it has to be created. Such prediction of the legal and technological requirements, and the strictness of the DUA terms and execution.

Datatags-based systems can predict the legal and technological dimensions, allowing an efficient and unambiguous description of each dataset properly, from a legal and contractual standpoint. By making decisions about data sharing policies applicable to a given repository, we can ensure that data is shared in a manner that is consistent with legal requirements.

Abstract

Technologically, there is an ever-growing set of solutions disseminated under binding contracts, data use agreements, data sharing and credible. Funding agencies and publications increasingly require data sharing. 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 collection methods, and legal and technological expertise — matching requires familiarity with the dataset’s history and the datatag most appropriate for it has to be created. Such prediction of the legal and technological requirements, and the strictness of the DUA terms and execution.

Datatags-based systems can predict the legal and technological requirements, and the strictness of the DUA terms and execution.

Managing Privacy in Research Data Repositories Workshop

July 13th, 2016

@michbarsinai
Based in part on:

Sweeney L, Crosas M, Bar-Sinai M. **Sharing Sensitive Data with Confidence: The Datatags System.** *Technology Science [Internet].* 2015. Technology Science

We present a framework for formally describing, reasoning about, and arriving at data-handling policies.
We present a framework for formally describing, reasoning about, and arriving at data-handling policies. Making it *Easier to store and share* scientific datasets.
Why Share Data?

- **Good Science**
  - Transparency
  - Collaboration
  - Research acceleration
  - Reproducibility
  - Data citation
- **Compliance** with requirements from sponsors and publishers
Why Share Data?

- Good Science
- Transparency
- Collaboration
- Research acceleration
- Reproducibility
- Data citation

**Compliance** with requirements from sponsors and publishers

*image: © Sesame Workshop*
Sharing Data is Nontrivial

- **Sharing may harm the data subjects**
- **Law** is complex
  - 2187 privacy laws in the US alone, at federal, state and local level, usually context-specific [Sweeney, 2013]
- **Technology** is complex
  - E.g. encryption standards change constantly, as new vulnerabilities are found
- **Specific dataset provenance** (may be) complex
Dataset handling policies play the critical role of balancing privacy risks and scientific value of sharing datasets.
Here are some Data Handling Policies.
Here are some new Data Handling Policies
Formal CS DHPs

W3C’s Privacy Preference Project (P3P)
- Focuses on web data collection

Open Digital Rights Language (ODRL)
- Models DRM, supports privacy and rule-based assertions

PrimeLife Policy Language (PPL)
- Focuses on downstream usage, using rules

Data-Purpose Algebra
- Models restriction transformation along data processing path

Robot Lawyers
- See next session
## DataTags

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

**DataTags and their respective policies**

Data-handling policies consist of independent aspects. Encryption at rest, transfer type, access credentials, etc.
Data-handling policies consist of independent *aspects*. Encryption at rest, transfer type, access credentials, etc.

Each aspect has multiple *possible requirements*, and can be defined such that these requirements are ordered.
DHPs: From Text to Space

Data-handling policies consist of independent aspects. Encryption at rest, transfer type, access credentials, etc.

Each aspect has multiple possible requirements, and can be defined such that these requirements are ordered.

Construct a data-handling policy space by viewing aspects as axes, where each aspect’s possible requirements serves as its coordinates.
## Going from this...

<table>
<thead>
<tr>
<th>Tag Type</th>
<th>Description</th>
<th>Security Features</th>
<th>Access Credentials</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue</td>
<td>Public</td>
<td>Clear storage, Clear transmit</td>
<td>Open</td>
</tr>
<tr>
<td>Green</td>
<td>Controlled public</td>
<td>Clear storage, Clear transmit</td>
<td>Email- or OAuth Verified Registration</td>
</tr>
<tr>
<td>Yellow</td>
<td>Accountable</td>
<td>Clear storage, Encrypted transmit</td>
<td>Password, Registered, Approval, Click-through DUA</td>
</tr>
<tr>
<td>Orange</td>
<td>More accountable</td>
<td>Encrypted storage, Encrypted transmit</td>
<td>Password, Registered, Approval, Signed DUA</td>
</tr>
<tr>
<td>Red</td>
<td>Fully accountable</td>
<td>Encrypted storage, Encrypted transmit</td>
<td>Two-factor authentication, Approval, Signed DUA</td>
</tr>
<tr>
<td>Crimson</td>
<td>Maximally restricted</td>
<td>Multi-encrypted storage, Encrypted transmit</td>
<td>Two-factor authentication, Approval, Signed DUA</td>
</tr>
</tbody>
</table>
...to This*

* Shown here is a 2-D projection over the DUA Agreement Method and Authentication axes.
...to This*

Yellow = \langle \text{click-through, password, clear, encrypt, registered, approval} \rangle

* Shown here is a 2-D projection over the DUA Agreement Method and Authentication axes.
<table>
<thead>
<tr>
<th>DUA agreement Method</th>
<th>Authentication</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Email/OAuth</td>
<td>Email/OAuth</td>
</tr>
<tr>
<td>Password</td>
<td>Password</td>
</tr>
<tr>
<td>Two Factor</td>
<td>Two Factor</td>
</tr>
</tbody>
</table>

**Strictness**

- **Sign**: None
- **Click-Through**: Email/OAuth
- **Implied**: Password

**DUA agreement Method**

- **None**
- **Email/OAuth**
- **Password**
- **Two Factor**
Strictness

DUA agreement Method
- Sign
- Click-Through
- Implied

Authentication
- None
- Email/OAuth
- Password
- Two Factor

compliance(P)
compliance(P)

Strictness

DUA agreement Method

Sign
Click-Through
Implied

Authentication
None
OAuth
Password
Two Factor

All policies that do not breach P
Lenience

<table>
<thead>
<tr>
<th>DUA agreement Method</th>
<th>None</th>
<th>Email/OAuth</th>
<th>Password</th>
<th>Two Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sign</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Click-Through</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Implied</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

P
Lenience

DUA agreement Method

- Sign
- Click-Through
- Implied

Authentication

- None
- Email/OAuth
- Password
- Two Factor

support(P)
Lenience

DUA agreement Method

- Sign
- Click-Through
- Implied

Support (P)

Authentication
- None
- Email/OAuth
- Password
- Two Factor

All policies that P does not breach
A Dataset and a Repository walk into a DHP space…

<table>
<thead>
<tr>
<th>DUA agreement Method</th>
<th>Sign</th>
<th>Click-Through</th>
<th>Implied</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>Blue</td>
<td>Green</td>
<td>Orange</td>
</tr>
<tr>
<td>Email/OAuth</td>
<td></td>
<td></td>
<td>Red</td>
</tr>
<tr>
<td>Password</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two Factor</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A Dataset and a Repository walk into a DHP space…

<table>
<thead>
<tr>
<th>DUA agreement Method</th>
<th>None</th>
<th>Email/OAuth</th>
<th>Password</th>
<th>Two Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sign</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Click-Through</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Implied</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Dataset**
- Blue
- Green

**Authentication**
- Orange
- Red
A Dataset and a Repository walk into a DHP space...

[Diagram showing a matrix with DUA agreement methods (Sign, Click-Through, Implied) on the y-axis and authentication methods (None, Email/OAuth, Password, Two Factor) on the x-axis. The matrix includes cells for compliance levels (Red, Orange, Blue, Green).]
A Dataset and a Repository walk into a DHP space…

![Diagram showing the relationship between Dataset, Repository, and DUA agreement methods.](image-url)
A Dataset and a Repository walk into a DHP space…

DUA agreement Method

Sign

Click-Through

Implied

Authentication

None

Email/OAuth

Password

Two Factor

Dataset

Blue

Green

Orange

Red
DataTags Tools

Open source on GitHub
A tag space is a hierarchical structure that defines a DHP space, with some assertion dimensions added.
Tag Space

BlueToCrimson.ts

/**
 * This is the tag space for the DataTags set proposed at:
 * Latanya Sweeney, Mercè Crosas, and Michael Bar-Sinai. Sharing sensitive data with confidence.
 */

DataTags: consists of Security, AccessCredentials

Security: consists of Storage, Transmit

AccessCredentials: consists of Authentication, Registration, Approval, DUAAcceptance.

Storage[How are data stored on disk]: one of clear [No encryption used], encrypt [Data are stored encrypted on disk], multiEncrypt [Data are encrypted on disk, in a way that the server cannot unencrypt the data]

Transmit[How are data travelling through networks]: one of clear, encrypt.

Authentication: one of none, password, twoFactor.
Tag-Space Visualized

Visualization using CliRunner (on a later slide) and Graphviz (www.graphviz.org).
Arriving at a DHP

Question: Please select one answer

Do the data contain information from a covered entity or business associate of a covered entity?

Terms

**Business associate**
A business associate is any person or organization, including a subcontractor, that acts on behalf of, or provides services to, a covered entity involving the use or disclosure of protected health information. This includes, but is not limited to, legal, actuarial, accounting, consulting, claim processing, data analysis, administration, utilization review, quality assurance, billing, benefit management, practice management, and re-pricing activities.

**Covered entity**
A covered entity is a health plan, health care clearinghouse, or health care provider that transmits any health information in electronic form.

Yes  Not Sure  No
Arriving at a DHP
Tags Questionnaire

❖ “Interview with an expert” metaphor

❖ Consists of a tag space and a decision graph
Decision Graph - Visualized

### simple.dg

**eduCompliance**

- **ask** Was written consent obtained?
  - **no** REJECT Cannot handle educational records without written consent.
  - **yes**
    - Set Assertions={educationalRecords} Handling=[Transit:encrypt Storage:encrypt]

- **ask** Do the data concern humans?
  - **yes** Set Assertions={humanData}
  - **no**
    - Set Handling=[Transit:clear Storage:clear]

- **ask** Does the data contain educational records?
  - **yes**
    - Set Handling=[Transit:encrypt]
  - **no**
    - todo Handle IP issues here

---

**[1]**

- **ask** Can handle educational records without written consent.
  - **yes**
    - Set Assertions={educationalRecords} Handling=[Transit:encrypt Storage:encrypt]
  - **no**
    - todo Handle IP issues here
Does the data visually adhere to the HIPAA Safe Harbor provision?

Did the business associate agreement have any additional restrictions on sharing?

Has an expert certified the data as being of minimal risk?

Did you acquire the data under a HIPAA Business Associate agreement?

Are you an entity that is directly or indirectly covered by HIPAA?

Did the consent have any restrictions on sharing?

Did you acquire the data under a HIPAA limited data use agreement?

Is there any reason why we cannot store the data?

For how long should we keep the data?

Is a qualified person prohibited from re-identify but not contact people whose information is in the data?

Is a qualified recipient prohibited from identifying and contacting people or organizations in the data?

Is a qualified recipient allowed to use the data results based on the data:

How should a qualified user accept the data use agreement?

How may a qualified recipient use the data?

How may the data be shared?

Is a qualified person allowed to sign digitally, with ID?

Is a qualified person prohibited from sign, with ID?

Does the data contain personal health information?

Does the data visually adhere to the HIPAA Safe Harbor provision?
Decision Graph - Visualized

### HIPAA Compliance - Decision Graph

- **DataType**
  - [Basis: HIPAAStatistician]
  - Harm: noRisk
  - Effort: deIdentified

- **Handling**
  - Storage: clear
  - Transit: clear
  - Authentication: none

- **DataTags**
  - no

- **3.1.3.1**
  - ask dua

- **3.1.1.1**
  - ask

- **3.6**
  - ask

- **3.5**
  - ask

- **3.3**
  - ask

- **DUA**
  - [Sharing: none]

- **Acceptance**
  - [signed]

- **Publication**
  - [noRestriction]

- **Use**
  - [noRestriction]
  - research
  - IRB
  - noProduct

- **Reidentify**
  - [noMatching]
  - [noPeople]
  - [noEntites]

- **Sharing**
  - [none]
  - 5years
  - _2years
  - _1year

- **Storage**
  - [clear]
  - encrypt
  - multiEncrypt

- **Transit**
  - [clear]
  - encrypt

- **Authentication**
  - [clear]
  - encrypt

- **TimeLimit**
  - _1year
  - _5years
  - none

- **Handling**
  - Storage: encrypt
  - Transit: encrypt
  - Authentication: twoFactor

- **DataTags**
  - no

- **3.2**
  - ask

- **Is a qualified recipient prohibited from identifying and contacting people whose information is in the data?**
  - yes

- **Is a qualified person allowed to contact organizations in the data?**
  - yes

- **Is a qualified person allowed to contact people whose information is in the data?**
  - yes

- **Does your data include personal medical records information?**
  - yes

- **Are you an entity that is directly or indirectly covered by HIPAA?**
  - yes

- **Was the data received from a HIPAA Business Associate agreement?**
  - no

- **Was the data received from a HIPAA Covered Entity agreement?**
  - no

- **Did the limited data use agreement have any additional restrictions on sharing?**
  - no

- **Has an expert certified the data as being of minimal risk?**
  - yes

- **Does the data visually adhere to the HIPAA Safe Harbor provision?**
  - yes

- **Publication**
  - [noRestriction]

- **Acceptance**
  - [signed]

- **Sharing**
  - [anyone]

- **Handling**
  - Storage: encrypt
  - Transit: encrypt
  - Authentication: twoFactor

- **DataTags**
  - no

- **Is a qualified recipient prohibited from re-identifying and matching the data to other data?**
  - yes

- **Is a qualified person prohibited from re-identifying but not contacting people whose information is in the data?**
  - yes

- **Consent Details**
  - explicitConsent

- **Publication**
  - [noRestriction]

- **Acceptance**
  - [signed]

- **Sharing**
  - [anyone]

- **Handling**
  - Storage: encrypt
  - Transit: encrypt
  - Authentication: twoFactor
CliRunner

- Questionnaire Development Console
- Run, debug, visualize

Query:
What answer sequences result in encryption=clear, harm=severe?
Collaborating

- Can use existing social coding tools

 Future Work: collaborative development environment

Screenshot: GitHub line comment
Dataverse Integration - Upload

Mockup!
Dataverse Integration - Upload

MOCKUP!

"Help me tag"
Dataverse Integration

Behind the scenes

1. Help Me Tag
2. Request Interview( questionnaire:Q, id:1 )
3. [link 1]
4. Redirect To [link 1]
5. [link 1]

Interview Process

6. Question
7. Answer

8. Interview( id:1 ) done + tags
9. Update deposit session with tags
10. [link 2]
11. Redirect to [link 2]
12. [link 2]

deposit continues
Dataverse Integration

Behind the scenes

User

Dataverse

1. Help Me Tag
2. Request Interview (questionnaire:Q, id:1)
3. [link 1]
4. Redirect To [link 1]
5. [link 1]
6. Question
7. Answer
8. Interview(id:1) done + tags
9. Update deposit session with tags
10. [link 2]
11. Redirect to [link 2]
12. [link 2]

deposit continues
Dataverse Integration - View

Description
This dataset contains files referred to in the paper. They were not included in the paper, as it could also be understood without them. Thus, this dataset could be viewed as an appendix for those who would like to deepen their understanding of the paper.

Subject
Law; Computer and Information Science

Files
14 Files
- algorithm-trace.html
- algorithm-trace.md
- BlueToCrimson-ts.pdf
- BlueToCrimson.dg
Dataverse Integration - View

Description
This dataset contains additional datasets referred to in the paper. They were not included in the paper, as it could also be understood without them. Thus, this dataset could be viewed as an appendix for those who would like to deepen their understanding of the paper.

Subject
Law; Computer and Information Science

Files
14 Files

- algorithm-trace.html
  - text/html - 4.5 KB - Feb 7, 2016 - 0 Downloads
  - MD5: 61c867f8541b4b72034033654f94e4a1
  - Viewable trace of the slot unfolding algorithm.

- algorithm-trace.md
  - text/markdown - 1.4 KB - Feb 7, 2016 - 0 Downloads
  - MD5: 23200ec0863f51c3ab8b6a497c4a386
  - Trace of the slot unfolding algorithm (editable markdown form).

- BlueToCrimson-ts.pdf
  - Adobe PDF - 16.6 KB - Feb 8, 2016 - 0 Downloads
  - MD5: d5836778795cd2f3268f6c537f3d7e
  - Visualization of the tagspace for the datatags set defined in http://techscience.org/a/2015101601/

- BlueToCrimson.dg
  - Unknown - 1.7 KB - Feb 8, 2016 - 0 Downloads
  - MD5: d32d25bcf80c4c7debe2023a5cc9345
  - Unknown file type.

Search this dataset...
DataTags

Create an automated interview for composing data policies, such that a tagging process has to yield a sub-graph in which the root node (shown as [PrivacyTagSet]) is the only possible result of a tagging decision.

A tagging process has to yield a sub-graph in which the root node (shown as [PrivacyTagSet]) is the only possible result of a tagging decision. The programming language for Tag Space and Harmonized decision-graph is a formal language, designed to allow legal experts with little or no cognitive effort to reason about tagging statements, suggested wording for questions, sub-questions, and parts of data policy.

The programming language for Tag Space and Harmonized decision-graph is a formal language, designed to allow legal experts with little or no cognitive effort to reason about tagging statements, suggested wording for questions, sub-questions, and parts of data policy. The tagging process is executed and reasoned about. HTML and Graphviz describe the data policy space. The programming language for Tag Space and Harmonized decision-graph is a formal language, designed to allow legal experts with little or no cognitive effort to reason about tagging statements, suggested wording for questions, sub-questions, and parts of data policy.

The programming language for Tag Space and Harmonized decision-graph is a formal language, designed to allow legal experts with little or no cognitive effort to reason about tagging statements, suggested wording for questions, sub-questions, and parts of data policy. The programming language for Tag Space and Harmonized decision-graph is a formal language, designed to allow legal experts with little or no cognitive effort to reason about tagging statements, suggested wording for questions, sub-questions, and parts of data policy.

The programming language for Tag Space and Harmonized decision-graph is a formal language, designed to allow legal experts with little or no cognitive effort to reason about tagging statements, suggested wording for questions, sub-questions, and parts of data policy. The programming language for Tag Space and Harmonized decision-graph is a formal language, designed to allow legal experts with little or no cognitive effort to reason about tagging statements, suggested wording for questions, sub-questions, and parts of data policy.

Usability is a major challenge for Usability is a major challenge for Usability is a major challenge for Usability is a major challenge for Usability is a major challenge.