DataTags

Framework & Development

Mercè Crosas
Institute for Quantitative Social Science

David O'Brien
Berkman Center for Internet & Society

Privacy Tools for Sharing Research Data
Presentation to the National Science Foundation
December 1, 2014

Overview

- Concept
- Research and Development
- Demo
- Plans for the Future

Concept

Benefits of Sharing Data

- Transparency
- Collaboration
- Research acceleration
- Reproducibility
- Data citation
- Compliance with requirements from sponsors and publishers

Difficulty of Sharing Sensitive Data

- **Complexity of law**
  - Thousands of privacy laws in the US alone, at federal, state and local level, usually context-specific
  - e.g., HIPAA applies only to covered entities (health plans, health care clearinghouses, and health care providers) and their business associates (legal, consulting, data aggregation, etc.)
- **Options available to researchers**
  - **Restrict access** under terms of a user agreement → *very few people can access, and use restrictions may be unclear or overly constraining*
  - **Share a deidentified dataset** → *many people can access, but often utility is reduced and/or privacy protection is ineffective*
  - **Not share at all** → *nobody can access*

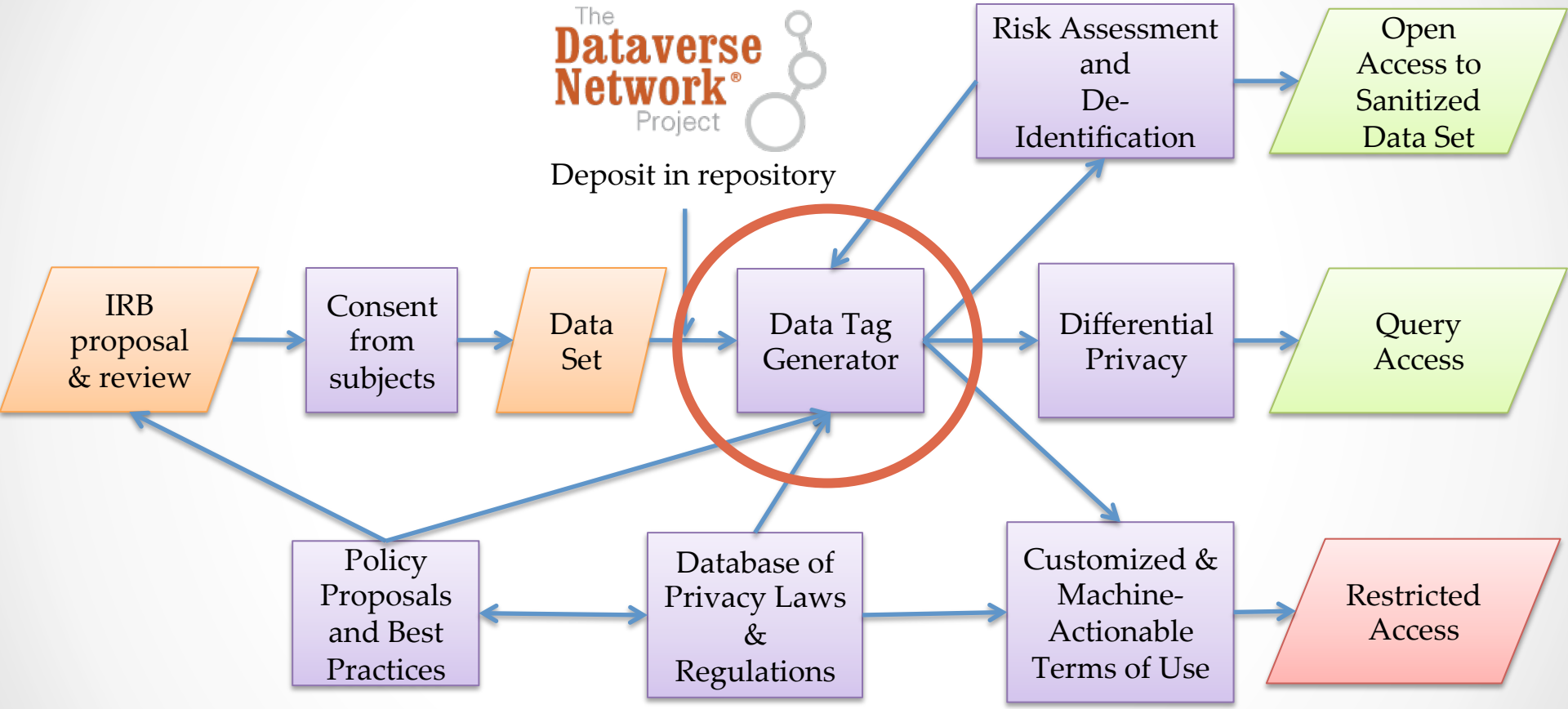
How about providing a tool that generates

a policy for your sensitive data

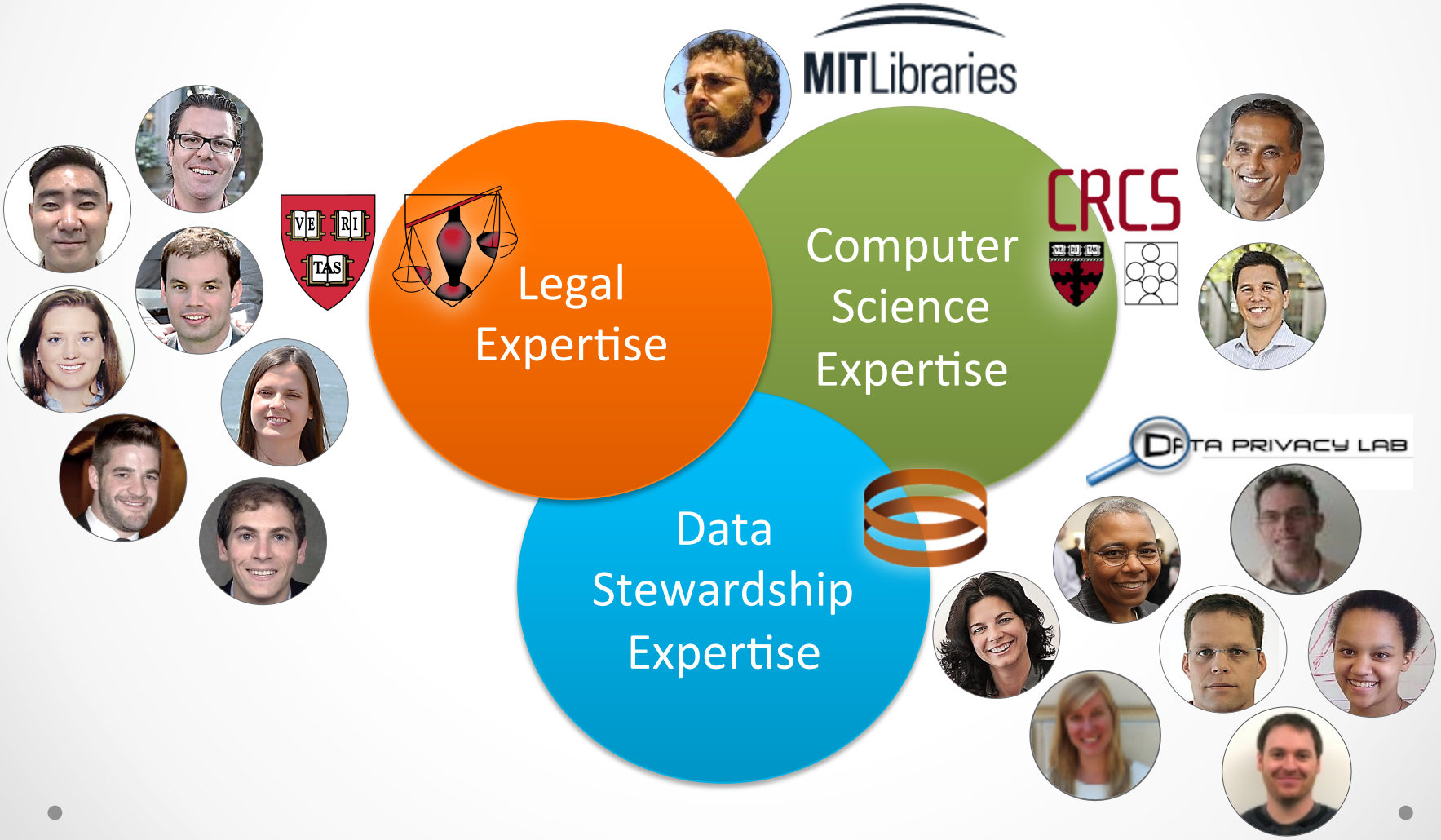
that defines how to transfer, store,
access and use those data?

. . . and DataTags was born

Integration of Tools



Building DataTags: A necessary collaboration



We benefited from …

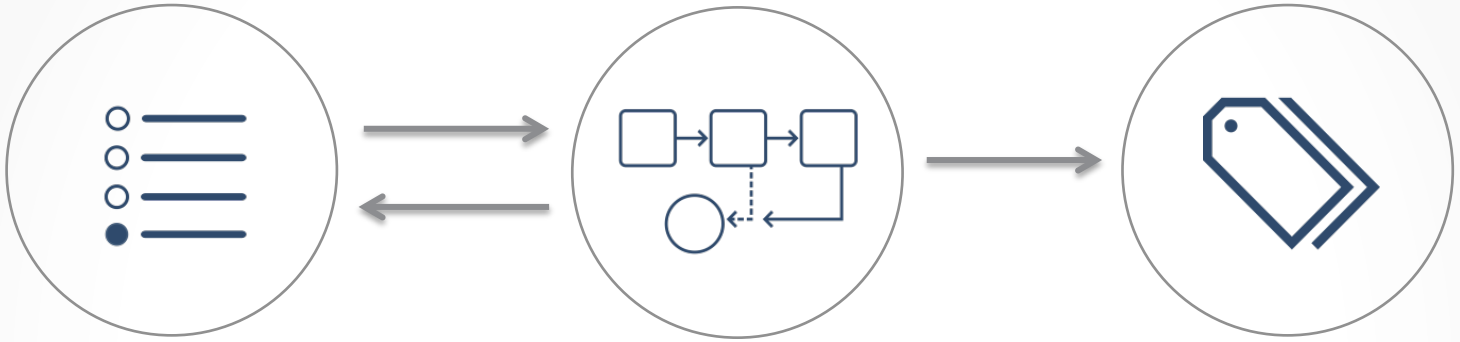
- **Initial work** from Latanya Sweeney to group the vast array of federal and state privacy laws into ~ 30 general classifications
- **Security levels** set by Harvard University for confidential information:

Level 1	**Non-confidential research information** *Public information*
Level 2	**Benign information to be held confidentially** *Information the disclosure of which would not cause material harm, but which the University has chosen to keep confidential*
Level 3	**Sensitive or confidential information** *Information that could cause risk of material harm to individuals or the University if disclosed*
Level 4	**Very sensitive information** *Information that would likely cause serious harm to individuals or the University if disclosed*
Level 5	**Extremely sensitive information** *Information that would cause severe harm to individuals or the University if disclosed*

DataTags Design Goals

1. **Define simple, iconic labels (tags)** to set requirements for transmitting, storing and using the dataset.
2. **Construct a decision graph** to map legal restrictions to dataset properties
3. **Automate a user-friendly interview** to elicit properties of the data and the collector from the user
4. **Generate a final report** with machine-actionable tags and related documentation
5. **Interoperate with data repositories** (e.g., Dataverse)

How DataTags Works



Step 1
Dynamic Questionnaire

DataTags elicits key properties of a given dataset by asking the user a sequence of questions that adjusts based on the user's answers

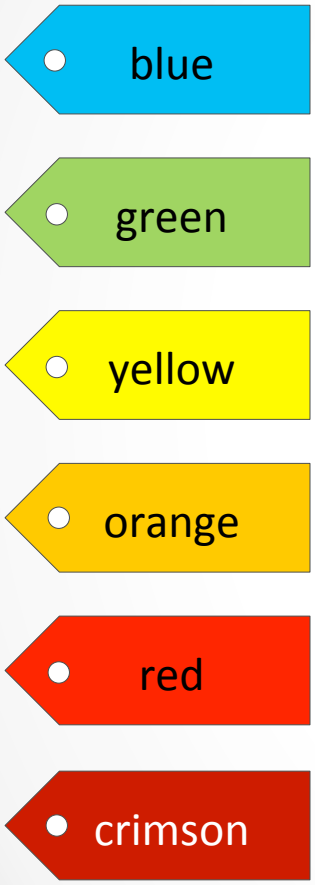
Step 2
Inference

Based on the user's responses to the questionnaire, DataTags applies inference rules to determine which legal restrictions are applicable and therefore which data handling tags should be assigned

Step 3
Assignment

DataTags generates simple, iconic tags that indicate how the dataset can be stored, transmitted, or used based on the dataset's properties and applicable legal restrictions

Draft Tag Levels



Non-confidential information that can be stored and shared freely

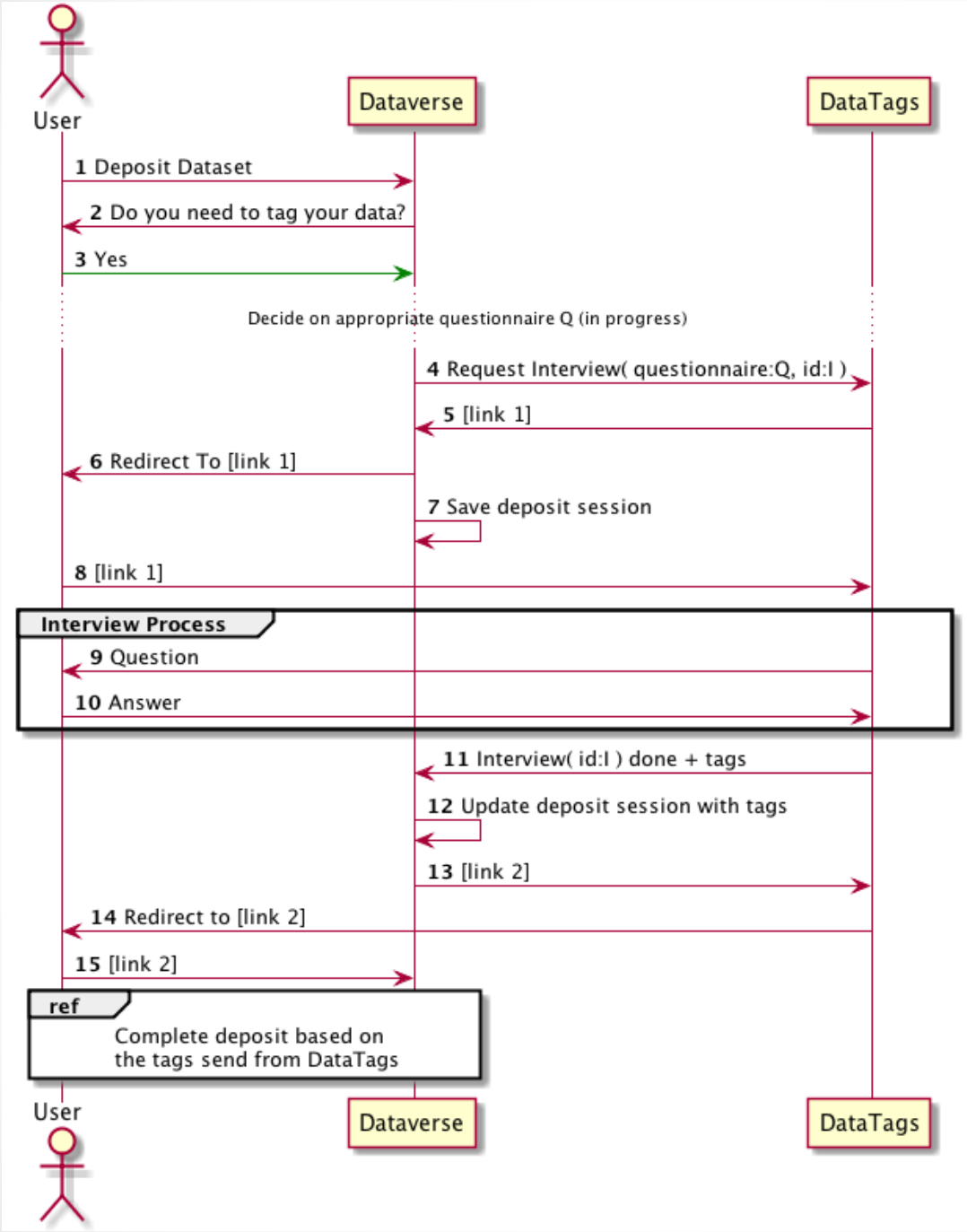
Potentially identifiable but not harmful personal information, shared with some access control

Potentially harmful personal information, shared with loosely verified and/or approved recipients

May include sensitive, identifiable personal information, shared with verified and/or approved recipients under agreement

Very sensitive identifiable personal information, shared with strong verification of approved recipients under signed agreement

Requires explicit permission for each transaction, using strong verification of approved recipients under signed agreement

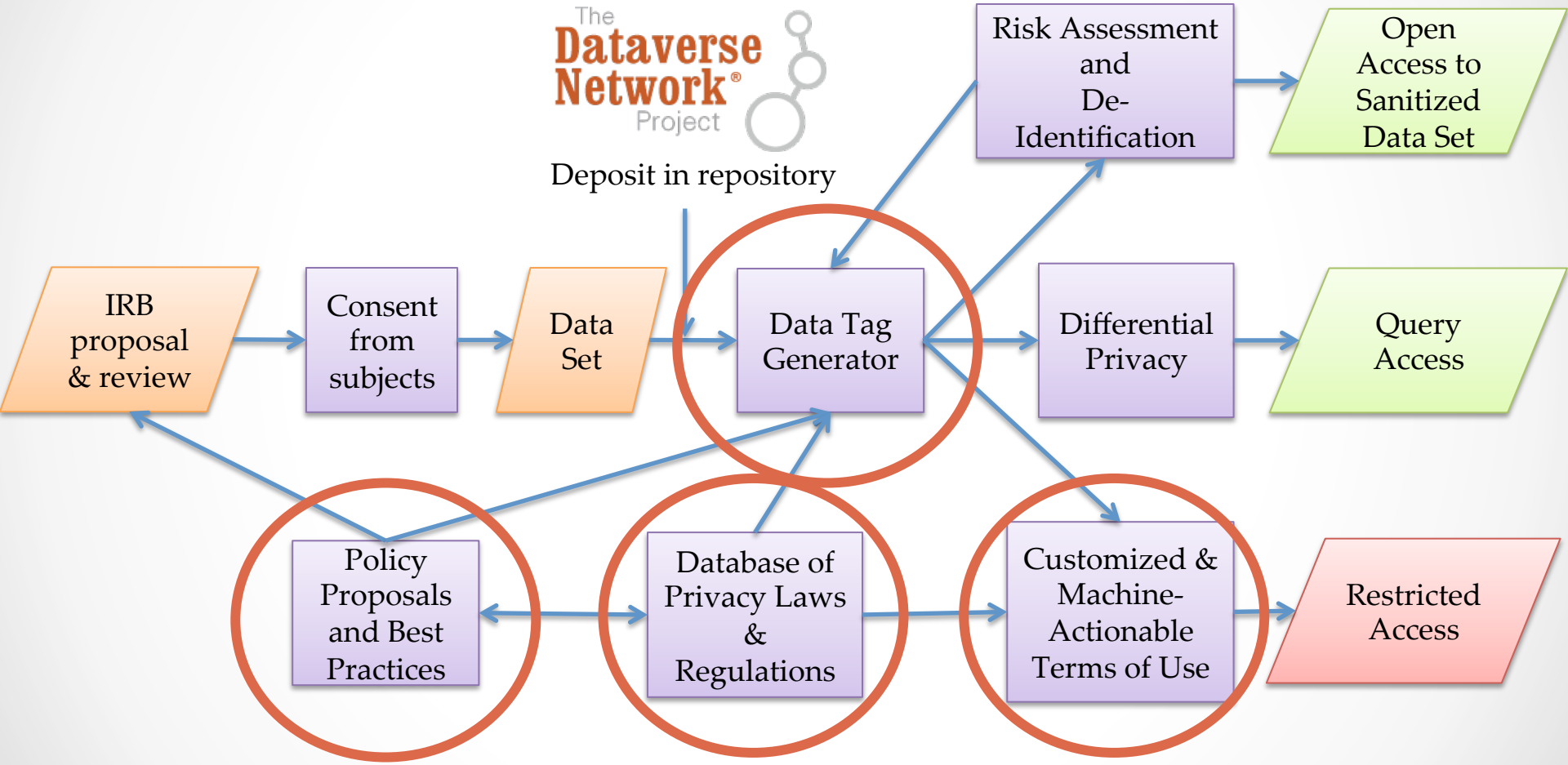


DataTags API

REU Project

Research and Development

Integration of Tools



Scope

Medical records

- HIPAA Privacy Rule
- Substance abuse confidentiality regulations

Student records

- FERPA
- Protection of Pupil Rights Amendment
- Education Sciences Reform Act

Government records

- Privacy Act of 1974
- Confidential Information Protection and Statistical Efficiency Act
- Title 13 (Census Bureau)
- Driver's Privacy Protection Act

Legal Research Objectives

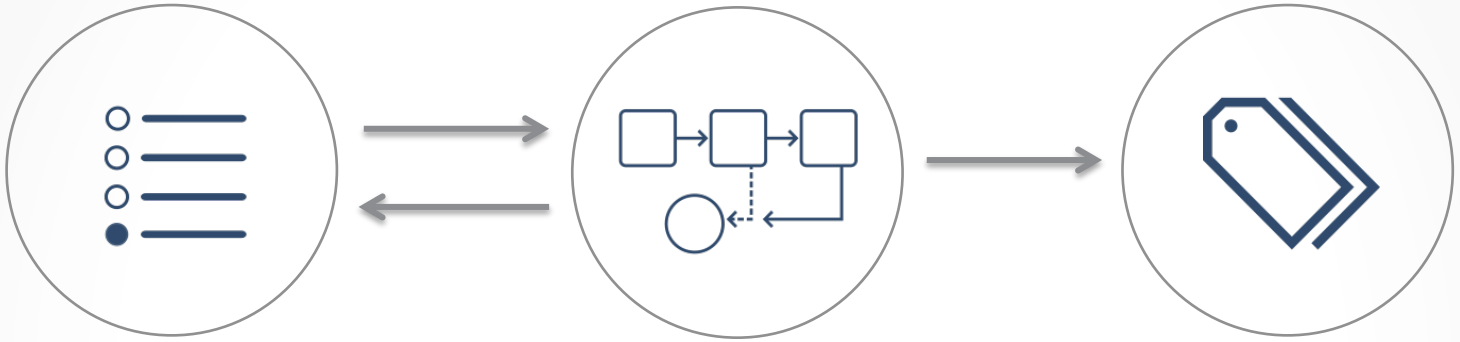
1. To better understand

- data sharing challenges,
- flows of information,
- the regulatory framework, and
- common institutional and contractual approaches

2. To identify typologies of

- regulatory and contractual requirements, and
- common practices from different disciplines

How DataTags Works



Step 1
Dynamic Questionnaire

DataTags elicits key properties of a given dataset by asking the user a sequence of questions that adjusts based on the user's answers

Step 2
Inference

Based on the user's responses to the questionnaire, DataTags applies inference rules to determine which legal restrictions are applicable and therefore which data handling tags should be assigned

Step 3
Assignment

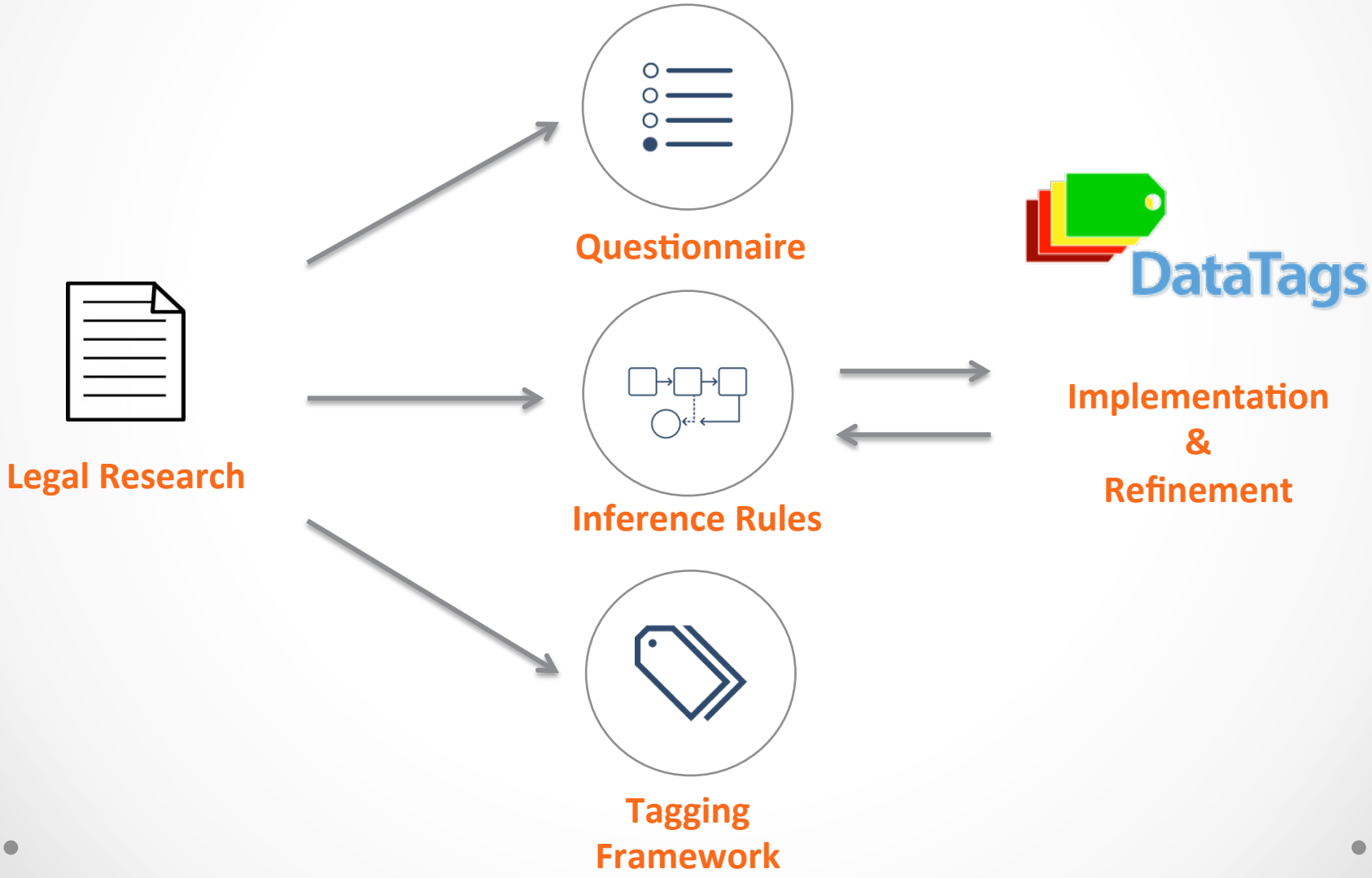
DataTags generates simple, iconic tags that indicate how the dataset can be stored, transmitted, or used based on the dataset's properties and applicable legal restrictions

Legal Research Activities

Legal memoranda and whitepapers analyzing

- Privacy statutes, regulations, case law, agency and institutional interpretations of laws
- Definitions of personally identifiable information and examples of sensitive information from different laws and regulations
- IRB and university data classification policies
- Common contractual approaches to data sharing

Research Integration



Questionnaire Development

1. **Identifying** regulatory and contractual requirements most relevant to research data sharing
2. **Clustering** legal requirements and approaches into general categories
3. **Drafting** annotated DataTags questionnaires
4. **Supporting** implementation process

Annotated Questionnaires

- **Question** (written in lay terms)
- **Defined terms**
- **Examples** of information for each definition
- **Citations** to and **excerpts** of legal authority
- **Interpretations** from case law, agency policies, and the policies of a variety of institutions
- **Tags** to be assigned based on answers given (and how they should be interpreted)
- **Rationale** (why the question is included, written the way it is, and how it translates and reconciles different legal interpretations)

Example Question: "education records"

Do the data contain any information derived from **institutional records directly related to a student**? This refers to records maintained by an educational agency or institution, or by a person on behalf of the agency or institution, for each student in the normal course of business. For the purposes of this questionnaire, information may satisfy this definition even if it has been **deidentified**.

Examples of records directly related to a student include:

- Directory information such as name, address, telephone listing, e-mail address, date and place of birth, dates of attendance, number of course units in which enrolled, class level, major field of student, last school attended, degrees and honors received, participation in official student activities, and student athletes' weight and height
- Demographic information such as gender, race, ethnicity, nationality, citizenship
- Identification photographs
- Current academic status, class schedule, courses taken, academic specialization and activities, units attempted and completed, instructors, past academic status, official communications regarding academic status
- Academic evaluations, including student examination papers, transcripts, grade point average, grades in courses, test scores, recorded communications that are part of the academic process, and other academic records
- …

Examples of records <u>not</u> directly related to a student include:

- Personal memory aids held in the sole possession of their creator (such as a teacher's notes)
- Records maintained by the law enforcement division of an educational agency or institution
- Employment records for an educational agency or institution
- Records produced by a physician, psychiatrist, or other professional for treatment purposes
- Grades on peer-graded papers before a teacher has recorded them
- Records directly related to other individuals, such as records of teacher misconduct or complaints against school employees, that only tangentially refer to students
- Information obtained from observation and not from an education record

Usability Testing

- **Feedback** received from 6 groups of potential DataTags users, including researchers, repositories, and academic journals
- **Report** compiling and analyzing recommendations from users:
    - More detailed documentation to explain the concept
    - Increased transparency into the tagging process
    - Enhanced definitions and examples to minimize ambiguity
- **Revisions** based on feedback incorporated in DataTags v. 1.0

Demo

Hypothetical Dataset

The dataset **contains information collected in a 10 year longitudinal health and education study**. The research is conducted by academic researchers and funded by their university and a private foundation (however, no government funding is involved)

- **Subjects:** public school students in grades 6-12. Parents received notice and provided written consent.
- **Dataset contains:**
  - grades collected pursuant to transcript requests
  - responses to questionnaires distributed at school that cover sensitive topics, such as drug and alcohol use
  - biomarkers like blood pressure and heart rate
  - direct identifiers, such as names and addresses have been removed, but indirect identifiers such as gender, race, height, weight, and ZIP code remain

Plans for the Future

Plans for the Future

- Conduct additional **usability testing and outreach**
- Add support for **data use agreements, consent terms, and IRB policies**
- Improve assessment of **identifiability, sensitivity, and the effectiveness of deidentification**
- Improve **harmonization** of approaches across different areas of law
- Develop a **modular license generator**
- **Integrate with repository** (Dataverse) and differential privacy tools
- Add formal **verification**, **validation**, and **optimization** theory and tools